A gLite FTS based solution for managing user output in CMS

Mattia Cinquilli[1], Hassen Riahi[2], Daniele Spiga[1], Claudio Grandi[3], Valentina Mancinelli[1], Marco Mascheroni[1], Francesco Pepe[3], Eric Vaandering[4], *on behalf of CMS collaboration.*

1) CERN 2) INFN Perugia 3) INFN Bologna 4) FNAL

Analysis use case

CMS[1] (LHC[2]) data is generated at experiment, processed and distributed worldwide over more than 100 sites connected through the Grid.

- Analysis jobs are sent with the data location driven model.
- Users output files can have a significant size (more than 1GB/job).
- Users need outputs in a "friendly" storage element.
- Need to store the output on a defined storage element for further access.

The production CMS Remote Analysis Builder (CRAB) [3] implements direct remote stage-out: jobs running in the worker node and copying each output file to a user pre-defined remote location at the end of job execution.

Synchronous stage-out experience

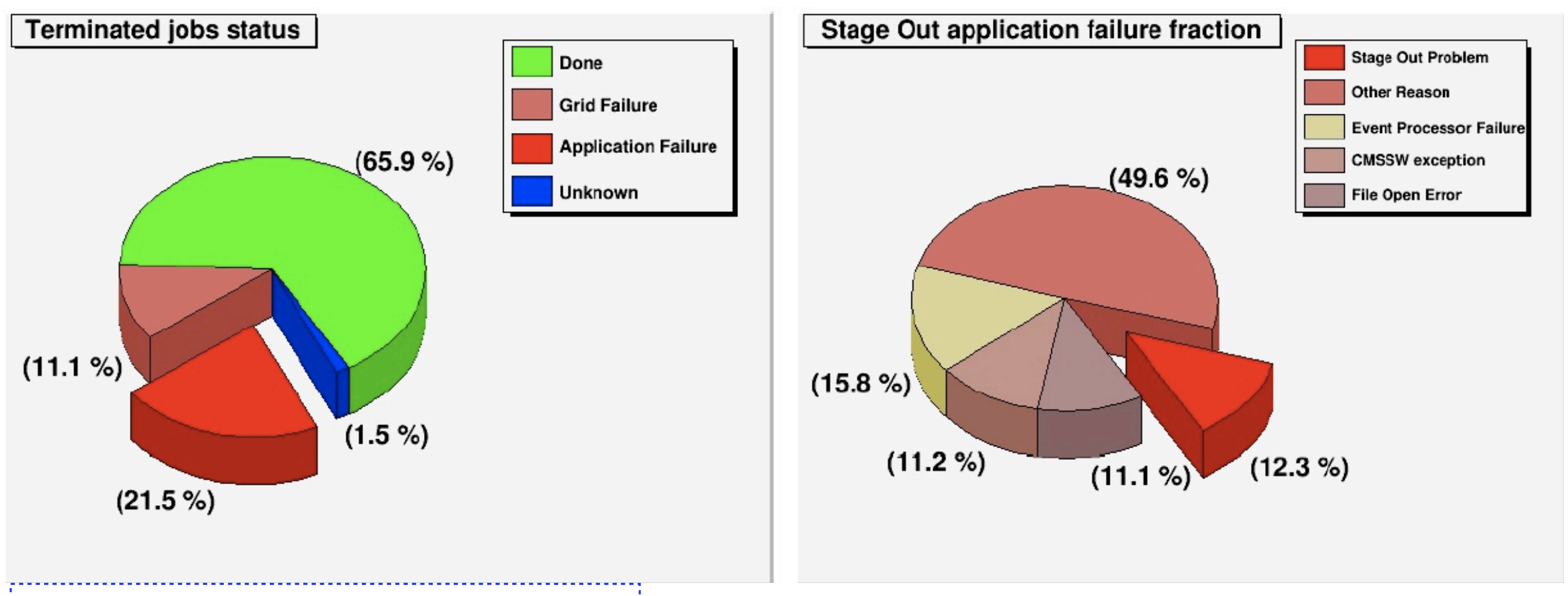
Synchronous remote stage-out has been demonstrated to work but has also highlighted various issues...

wasting precious resources

infrastructure issues

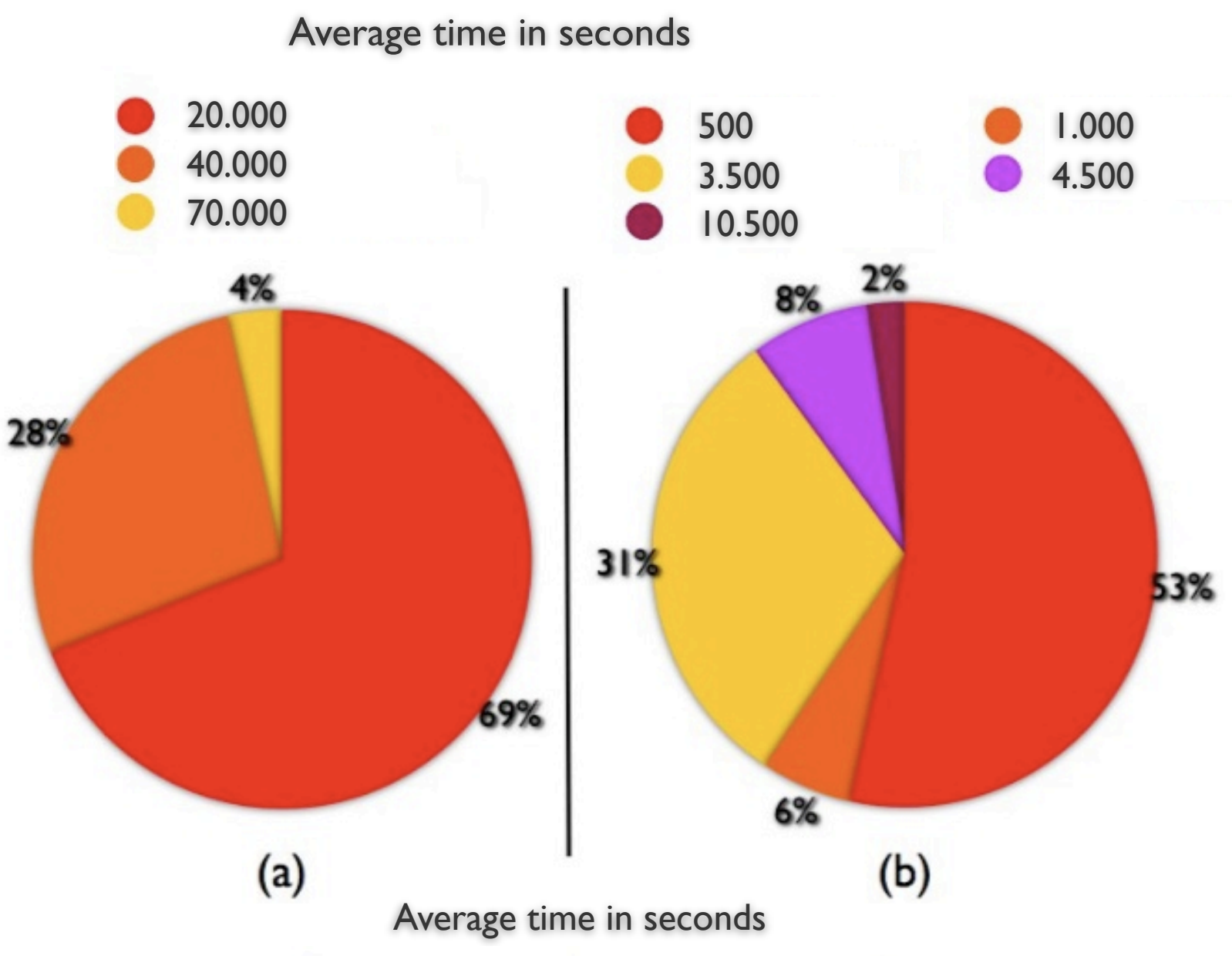
delaying user analysis

decreasing job efficiency

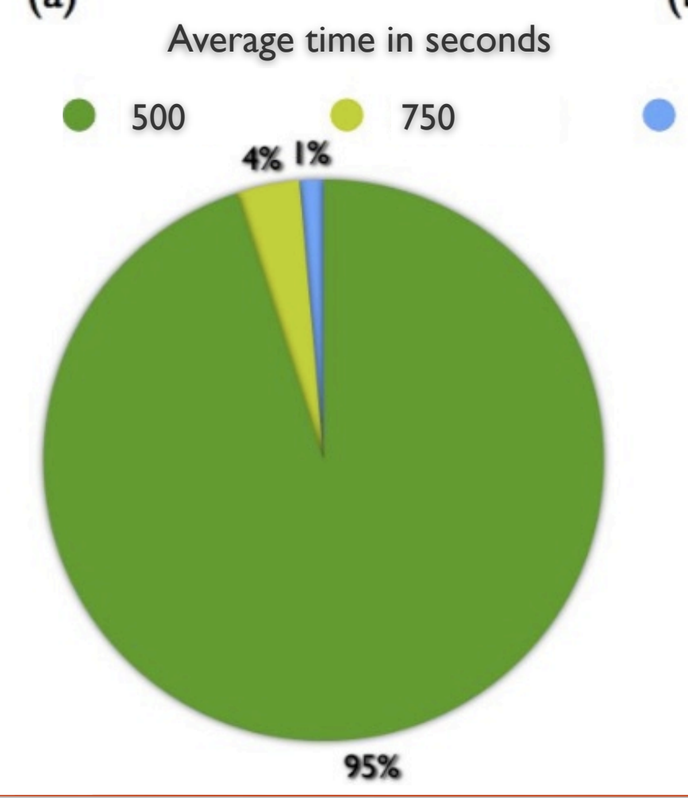


fractions of CMS analysis jobs by terminated status (a) and grouped by error types (b) during 2010

distribution of the number of jobs which failed remote stage-out over average CPU wall-clock time spent (left) and for the remote stage-out (right)



distribution of number of CMS analysis jobs by average CPU wall-clock time spent doing remote stage-out from worker nodes

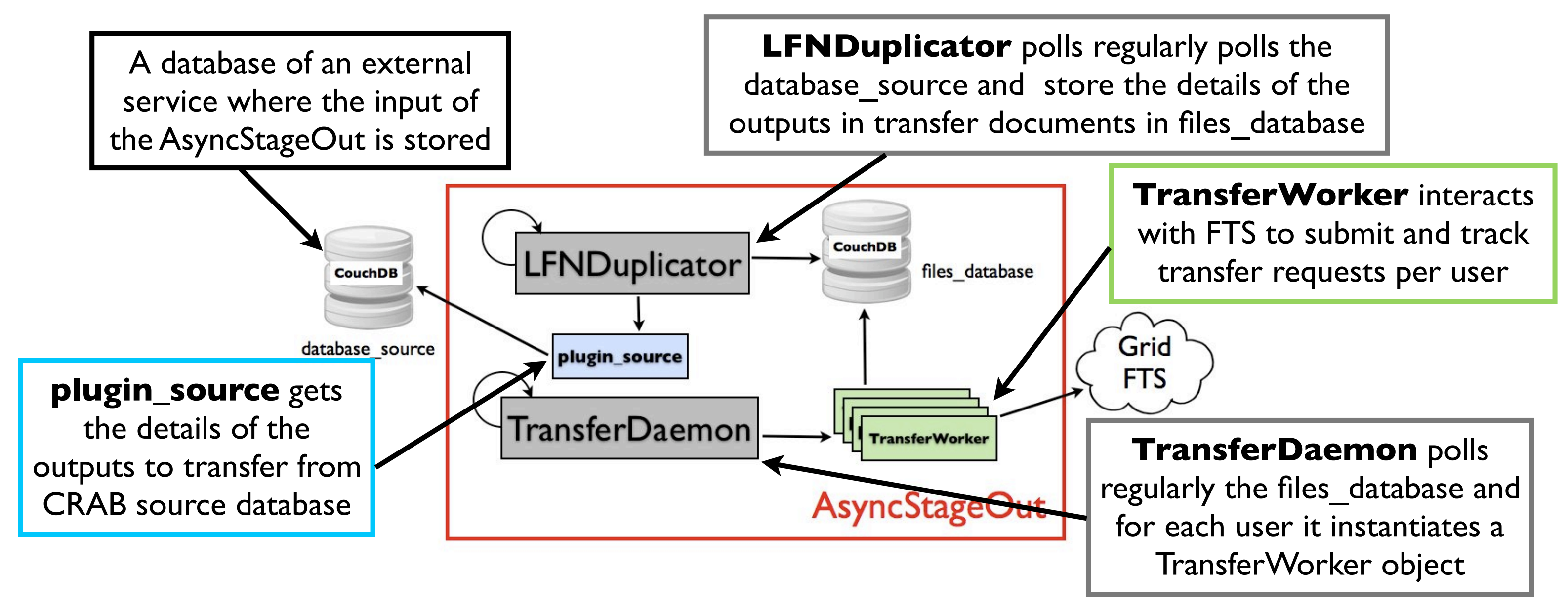


During July 2011 CMS had wasted about 24500 days of CPU wall-clock time due to remote stage-out from worker node.

Asynchronous stage-out

1) Local site storage at job runtime used as cache for output files.
2) AysncStageOut[4] tool creates, submits and manages gLite File Transfer Service (FTS)[5] jobs to transfer these outputs to the final destination site.

Architecture



A database of an external service where the input of the AsyncStageOut is stored

LFNDuplicator polls regularly polls the database_source and store the details of the outputs in transfer documents in files_database

TransferWorker interacts with FTS to submit and track transfer requests per user

plugin_source gets the details of the outputs to transfer from CRAB source database

TransferDaemon polls regularly the files_database and for each user it instantiates a TransferWorker object

Workflow

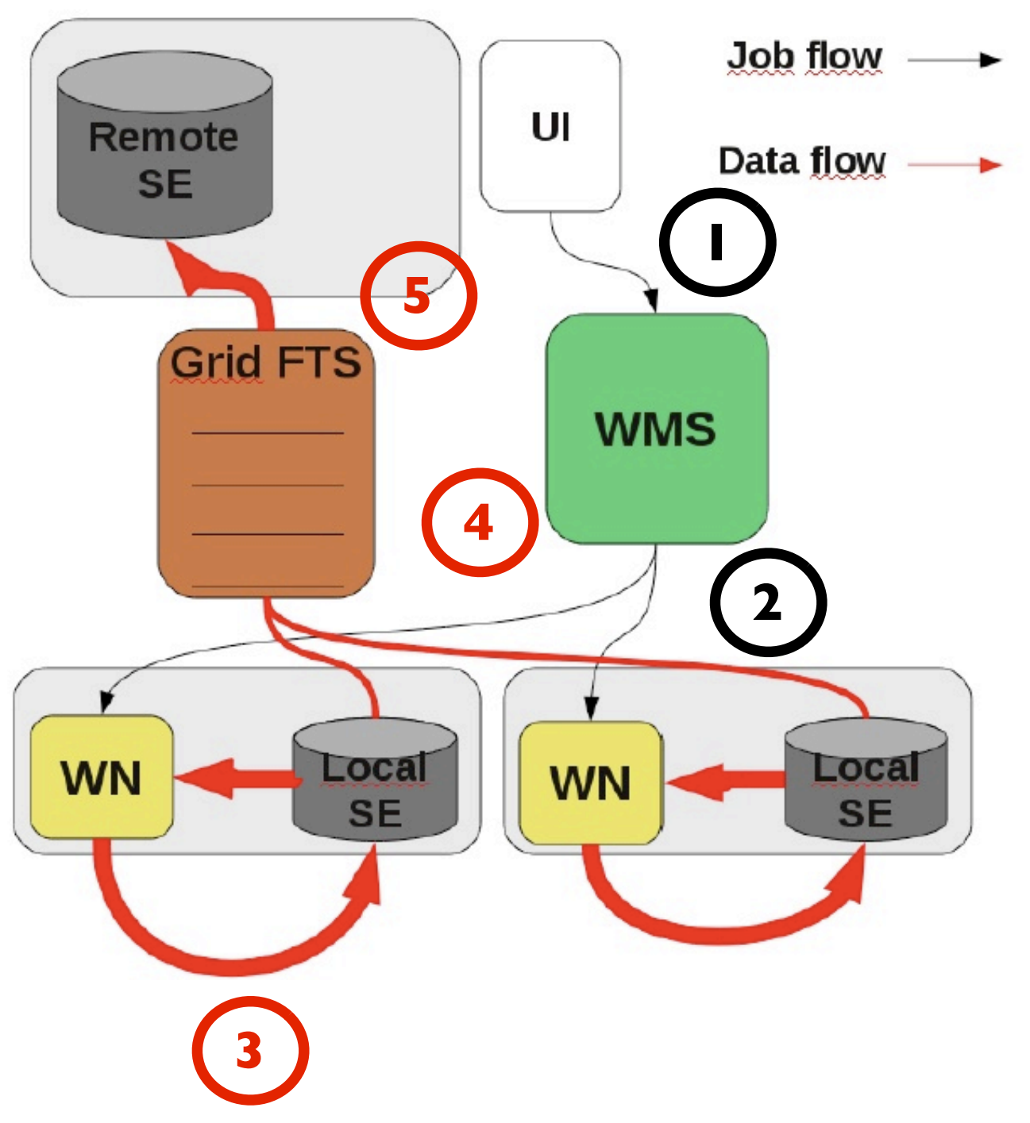
1. User submits his analysis workflow from his UI,

2. based on the configuration and the location of the data, the jobs are scheduled by the WMS to run in matched Tier-2s,

3. once the execution of the analysis code is done, the output is copied in the local SE of the site (**Local stage-out**) in **/store/temp/user**,

4. if the local copy of the output succeeds, a request is automatically submitted to FTS to copy the output to the remote SE (**Remote stage-out**) in **/store/user**,

5. the transfer request is then tracked and resubmitted if required.



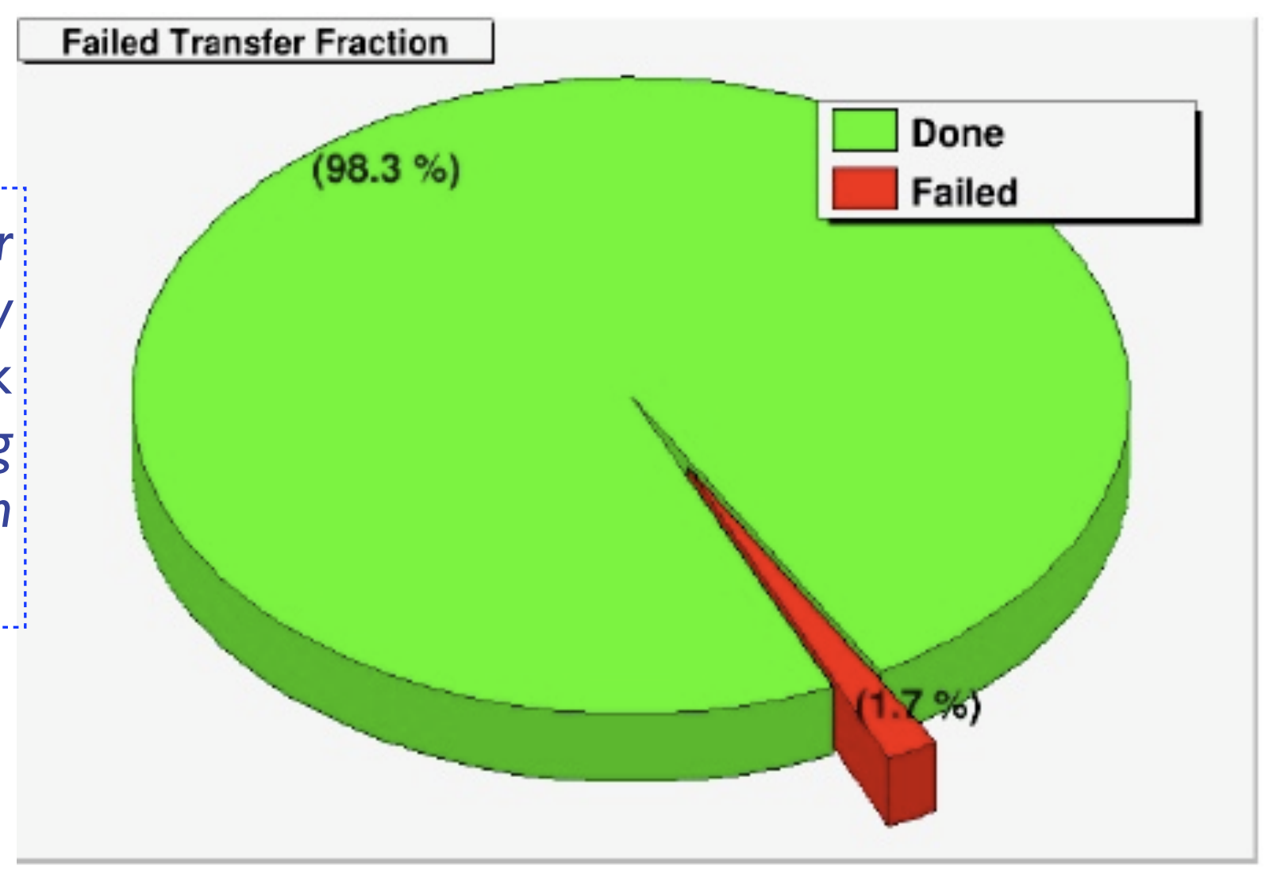
Tests and results

First functional tests on real CMS analysis scenario with 1700 successful jobs.

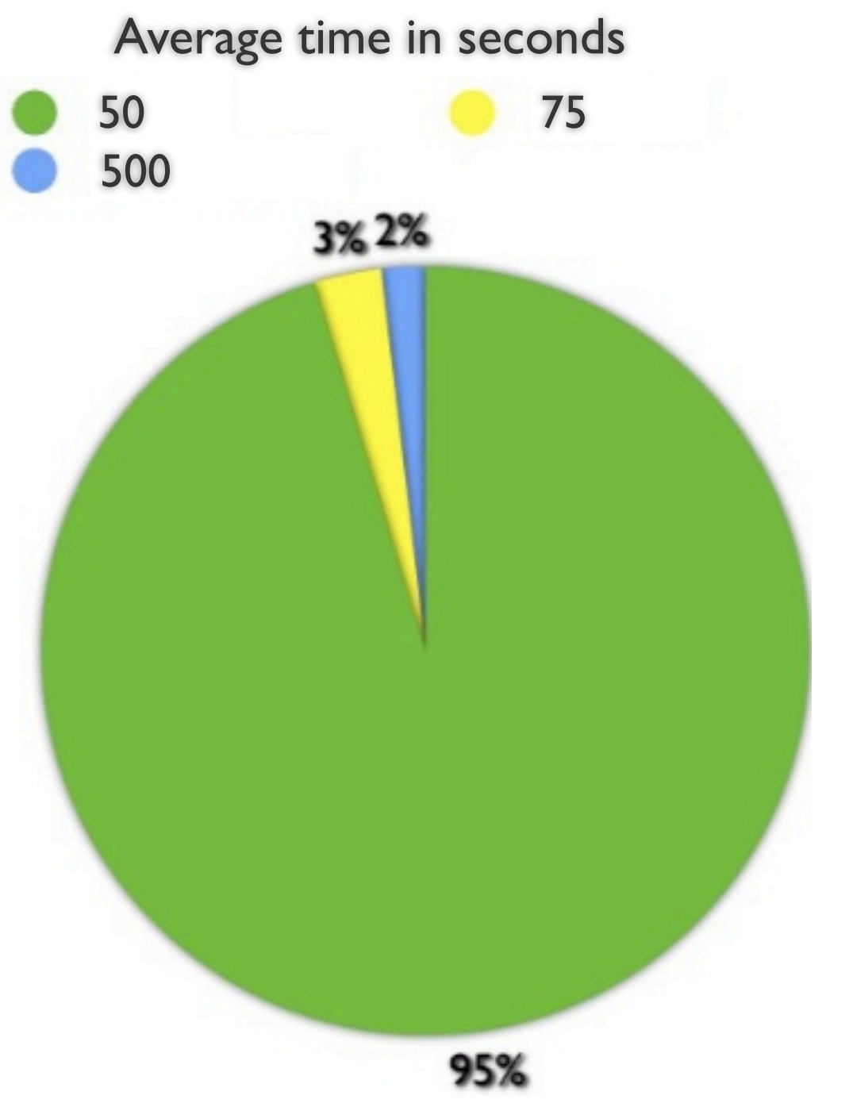
More scalability tests with ~24.000 successful jobs.

- Global asynchronous stage-out efficiency at 98%.
- Transfer retry improves the success rate and decreases the wast of resources.
- Possible and easy to spot site issues.
- Local stage out takes less than 75 sec in more than 95% of the jobs.

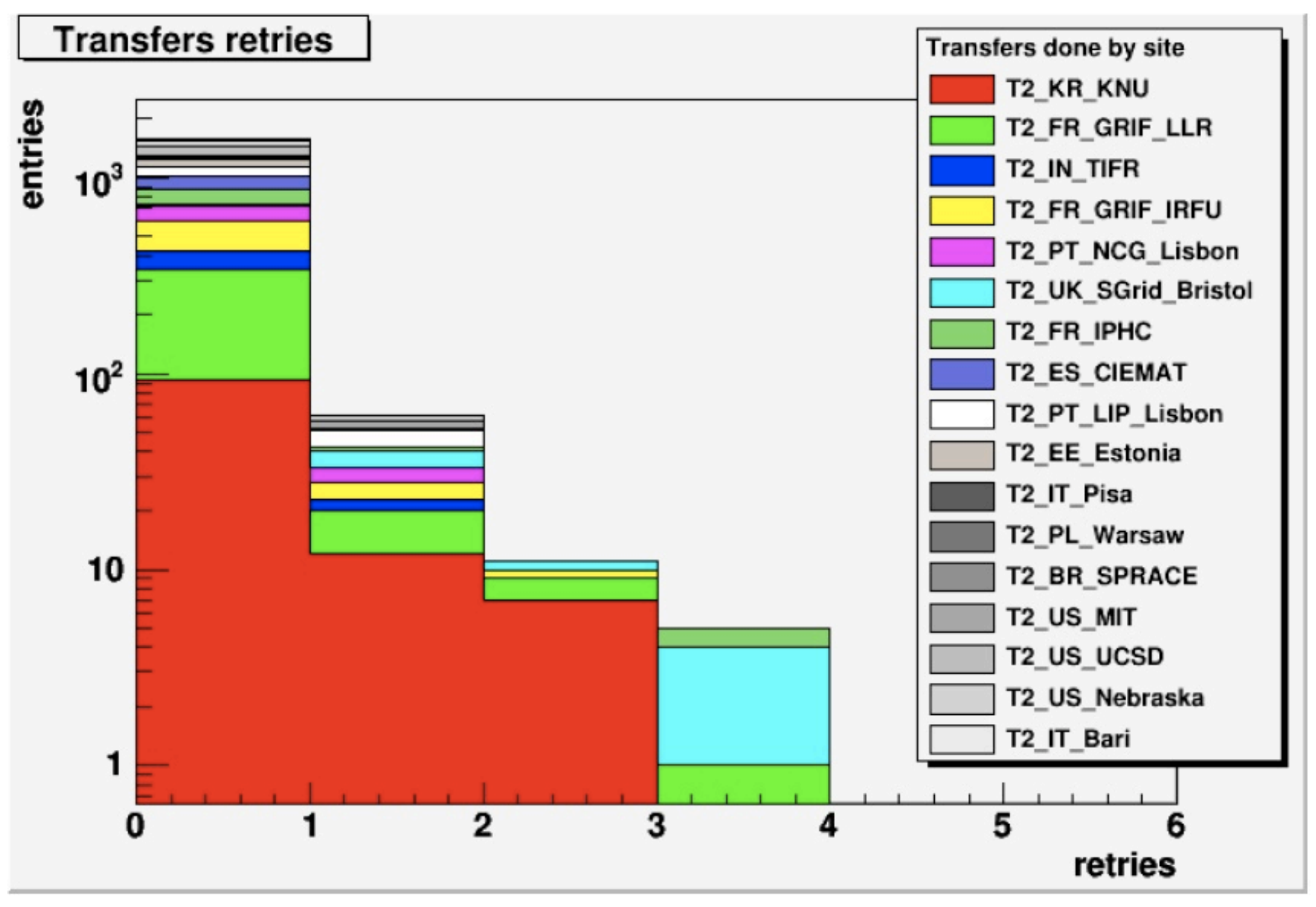
distribution of number of CMS analysis jobs by average CPU wall-clock time spent doing remote stage-out on worker nodes



distribution of number of CMS analysis jobs by average CPU wall-clock time spent doing local stage-out from worker nodes



number of transfer attempts by site



Conclusions

Asynchronous stage-out improves the management of CMS user analysis workflows.

Compared to the synchronous remote stage-out:

- ✓ Avoids wasting CPU wall-clock time:
  - reducing failures on worker nodes,
  - avoiding resubmission of full analysis job in case of stage-out failure,
  - local stage-out on worker node takes 10 times less than remote.
- ✓ Reduces latency in executing analysis workflows
  - avoiding manual resubmission of stage-out failed jobs,
  - easier life for the analysis users.
- ✓ Improves the usage of the underlying infrastructure
  - using dedicated services for Tier-2s sites transfers,
  - avoiding to overload networks and storage systems.

Future work:

- to be included by default in CMS analysis workflows once CRAB3 goes in production,
- evaluate if while running at scale introduces delays to FTS transfers of other CMS activities (e.g. PhEDEx[6] data transfers).

CERN IT
Experiment Support

1. The Compact Muon Solenoid experiment http://cms.web.cern.ch
2. The Large Hadron Collider http://lhc.web.cern.ch
3. Cinquilli, M. et al. "CRAB3: Establishing a new generation of services for distributed analysis at CMS" - Poster #206 of this conference.
4. H. Riahi "Design optimization of the Grid data analysis workflow in CMS" - Ph.D. Thesis, University of Perugia, Italy, 2012.
5. A, Frohner et al. "Data management in EGEE" - Journal of Physics Conference Series 219 062012, 2009.
6. J. Rehn, et al. "PhEDEx high-throughput data transfer management system" - Proceedings of Computing High Energy Physics, 2006.