A comparison of different database technologies for the CMS AsyncStageOut transfer database

Eric Vaandering
for
Marco Mascheroni, Diego Ciangottini, Justas Balcas

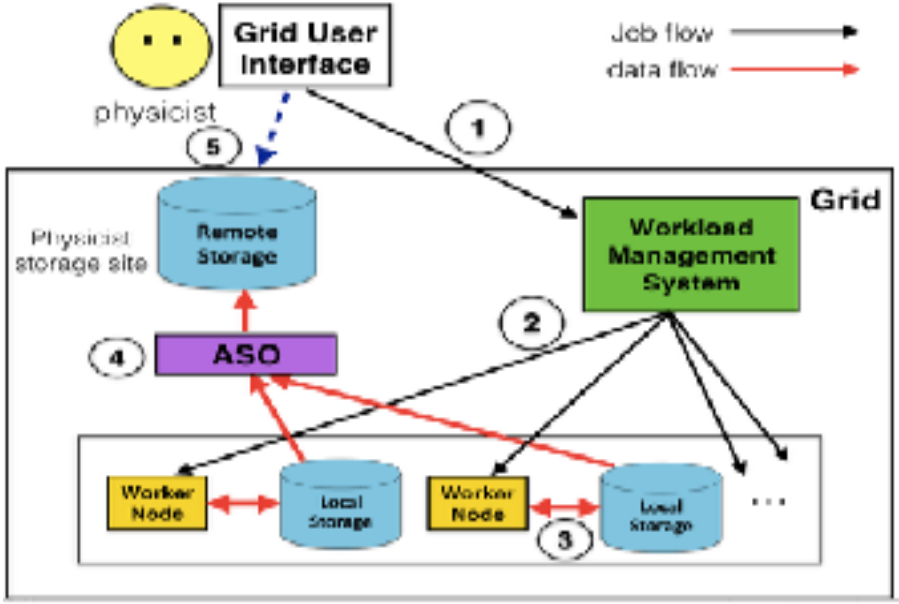
What is AsyncStageout (ASO)

ASO takes care of transfer management for CMS **users**

A physicist runs jobs through the workload management system

Job's output is stored in a temporary area of the storage element of the site where the job ran (**local** storage)

All outputs are then transferred to the site belonging to the user's institution (**remote** storage)

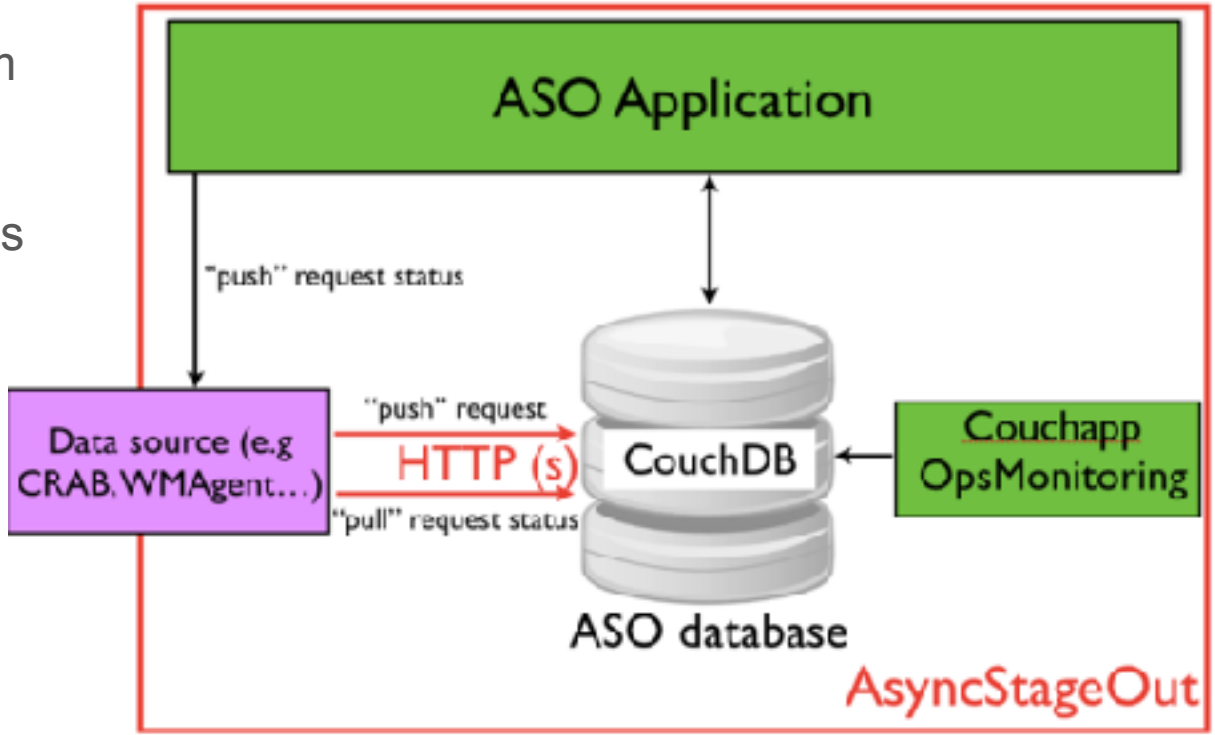


ASO general architecture

ASO uses **FTS** to perform the transfers

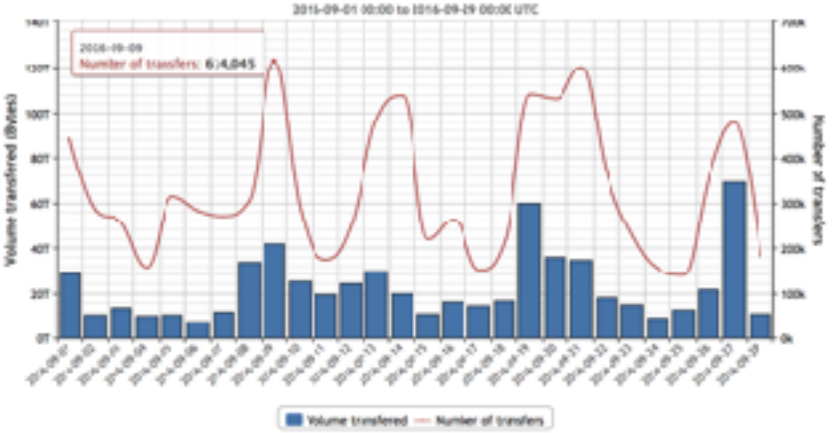
Requests from CRAB jobs are posted into a DB (CouchDB)

The DB stores the state of transfers and other relevant info

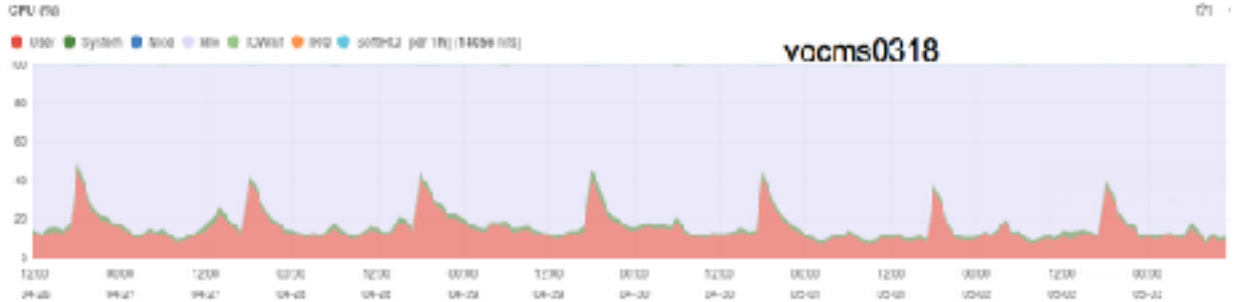


ASO usage

- In production since June 2014
- Key component of CMS computing
  - September: managed over 600k file transfers
- Improves previous model where transfers were done from WN and:
  - Sometimes could cause DDoS of CMS Tier 2 storage systems
  - Transfer failure was the primary reason of job failures
  - Resulted in loss of cpu because of failures (need to rerun the job) and because WN CPU is sitting idle during the transfer
- Critical parameter in operation is the number of transfers requests
  - User transfers consist of many small transfers (differently from central production)
  - Pressure to the ASO transfer CouchDB database instance



ASO CouchDB



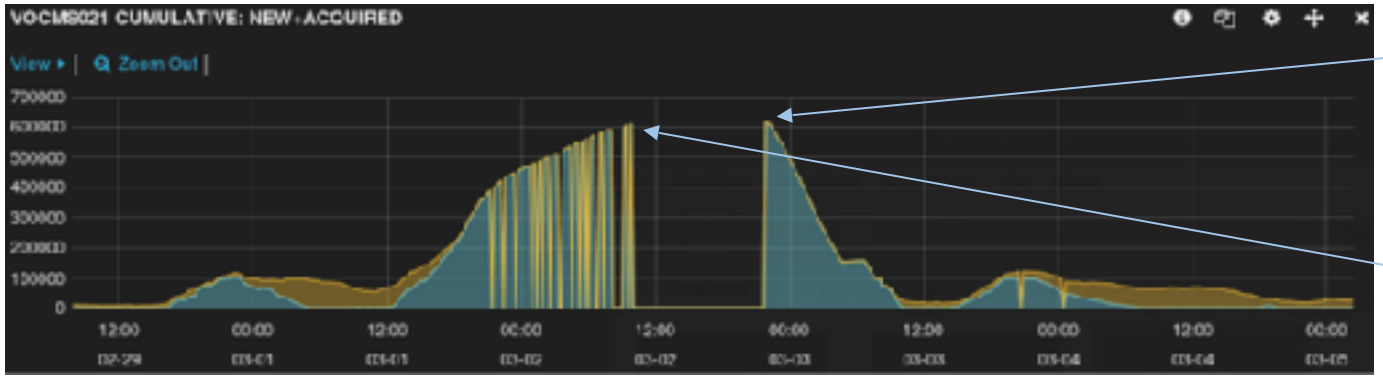
Suffers of performance limitations: the particular choice (CouchDB) is not well suited for frequently changing data

Internally created a new document every time you need to update some data (e.g. need to change the state of a transfer)

Size of DB and caches grows too fast and a daily compaction run

Burst of transfer requests that arrives during compaction can create a downward spiral where documents cannot be processed by ASO

Problems with ASO CouchDB



Queue of 600k documents ready to be transferred

Drops indicates that the database could not answer to the request

During that period transfers ready to be processed by the ASO application piles up and database become unresponsive

Delay in time to provide physicist results of their jobs

Waste of resources since eventually workload management system gives up and jobs have to be run again

Explored solutions

Horizontal scalability

The ASO tool is structured to be horizontal scalable (more CouchDB, ASO instances), but…

ASO CouchDB instances requires special hardware that is difficult to deploy (not from standard offers from IT)

Aim to find a way to reduce load on CMS operators, not to increase it!

Use a different technology

Couch used for historical reasons, never profited from the map/reduce paradigm typical of no-SQL databases

How to choose among the different database management systems available?

Other components of the workload management system already uses **Oracle**

No need of special hardware since it is among the services offered by CERN IT

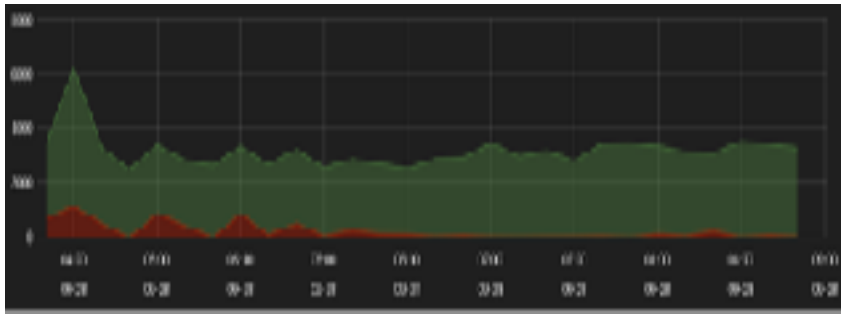
Current state

Oracle solution has been implemented. Required changes :

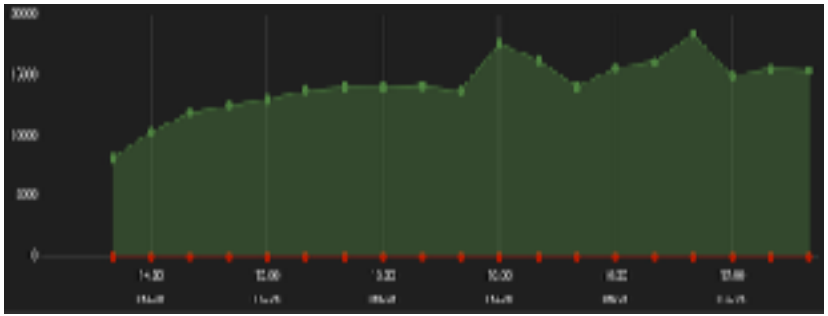
- ASO backend software
- Job wrapper that insert the transfer request in the database and other parts of the workload management system (CRAB)
- New REST interface on top of the Oracle schema to also allows authen/authz and insert data from the WN using user's certificate

Early scale tests at the production rate do not show any problem

Production rate of 3k jobs/10m



15k jobs/10m average for scale test



Conclusions

The code for the migration of the CMS transfer database to Oracle is in production since the October release

Solved problems in Oracle connection management

- By grouping multiple SQL queries into single Oracle request

Currently using both CouchDB and Oracle at the same time to make sure that:

- Functionally all the corner cases and potential bugs have been addressed (in case integration test suite does not cover them all)
- There are no hidden scalability bottlenecks

Plan to parasitically increase Oracle share over the time until CouchDB is switched off

Backup slides

ASO detailed architecture