
CERN

Storage preGDB : Summary

Data Steering Group

- Topics were chosen by the Data Management Steering Group.
  - Aim: facilitate the creation of common solutions across WLCG
- https://twiki.cern.ch/twiki/bin/view/LCG/WLCGDataSteeringGroup
- Need input on what discussions or forums are needed in the future.

Themes

- Exploitation of Object Stores
- Making the most of tapes
- Resource reporting and storage accounting

https://indico.cern.ch/event/578974/

Object Stores

- Understand…
  - The potential of object stores
    - What are they (not) good for??
  - Their integration into existing services
    - Use as an “SE”, adaptors, …
  - Their use as part of cloud procurements
  - Who is doing what with them

Echo @ RAL

Conclusions

15

- As object store is in production and functioning as a normal Grid Storage endpoint for ATLAS.
- VO should ensure that their software can work with different types of backend storage service.
- To fully integrate Object Stores, we are going to need to deploy services on top of them. Ideally one that:
  - Stored the secret key and provided authentication / authorization.
  - Could proxy the transfer in the event that third party WebDav transfers are not supported.

Alastair Dewhurst, 12th September 2017

Dynafed

- Role as “cloud storage integrator”
- RAL
- University of Victoria, Canada
- CERN volunteer computing platform

Ceph S3 at CERN: cs3.cern.ch

- 865TB cluster, 369TB RAW used.
  - 32 OSD servers, 3 virtual Ceph mons, 10 virtual S3 gateways.

- Running Ceph jewel 10.2.8, servers run CentOS 7.3.
- S3 data pool uses **erasure-coding**: 4 data + 2 parity stripes.

- Buckets accessible as <bucket>.cs3.cern.ch or cs3.cern.ch/<bucket>
- http or https; Comodo SSL certificate for (*.)cs3.cern.ch

Discussion…

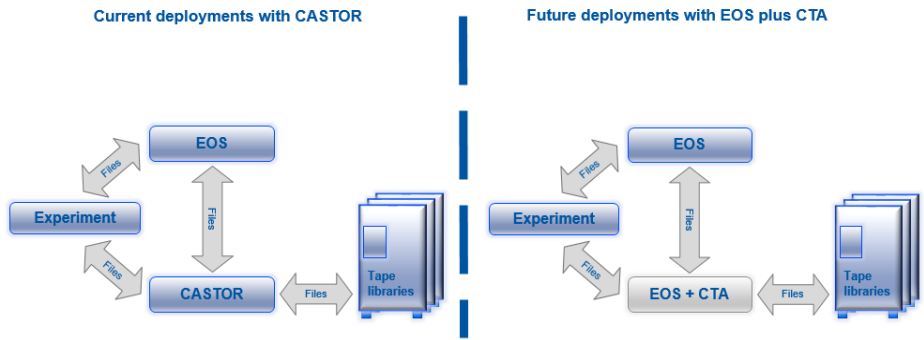
- Lots of interest in RAL's experience
- How best to integrate such systems (Ceph, essentially) into WLCG, by consolidating on one approach?
  - Gateways (over rados? S3?)
  - Dynafed – use as access federator, cloud adaptor, replica manager.

Making the most of tape

- In part, a follow-up of discussion in the FTS steering group
  - Broader scope
- How we can make the most of tape
  - What are the key metrics?
  - How can we optimise them?
    - Clients and interfaces
- How are experiments using it now?
- Updates from providers
  - dCache
  - CERN Tape Archive (CTA)

What is EOS+CTA?

- **EOS plus CTA is a “drop in” replacement for CASTOR**



RESTfull API examples

Query available transitions for a QoS

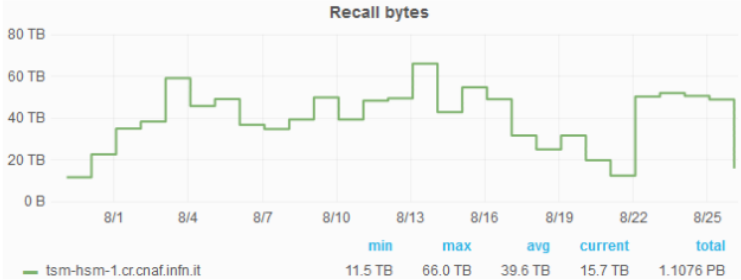
```
$ curl http://dcache-se/api/v1/qos-management/qos/file/disk
{
  "backendCapability" : {
    "name" : "disk",
    "transition" : [ "tape", "disk+tape" ],
  }
  "status" : "200",
  "message" : "successful",
}
```

dCache

Recent intense recall activity by CMS (cont.)

In **28** days (from 30/7 to 26/8):

- bytes recalled: **1.11**PB (of 12PB total) - **9.2** %
- tapes used in recall: **1103** (out of 1544 total) - 72 %
- files recalled: **400k**
- number of mounts (tot): **9858**
- mount/day (average): **352**
- mount/tape (average): **9**, (max): **17**
- number of tape drives used (avr): **6** (max): **7**
- Waiting time: **5.6** days (average of max/day)

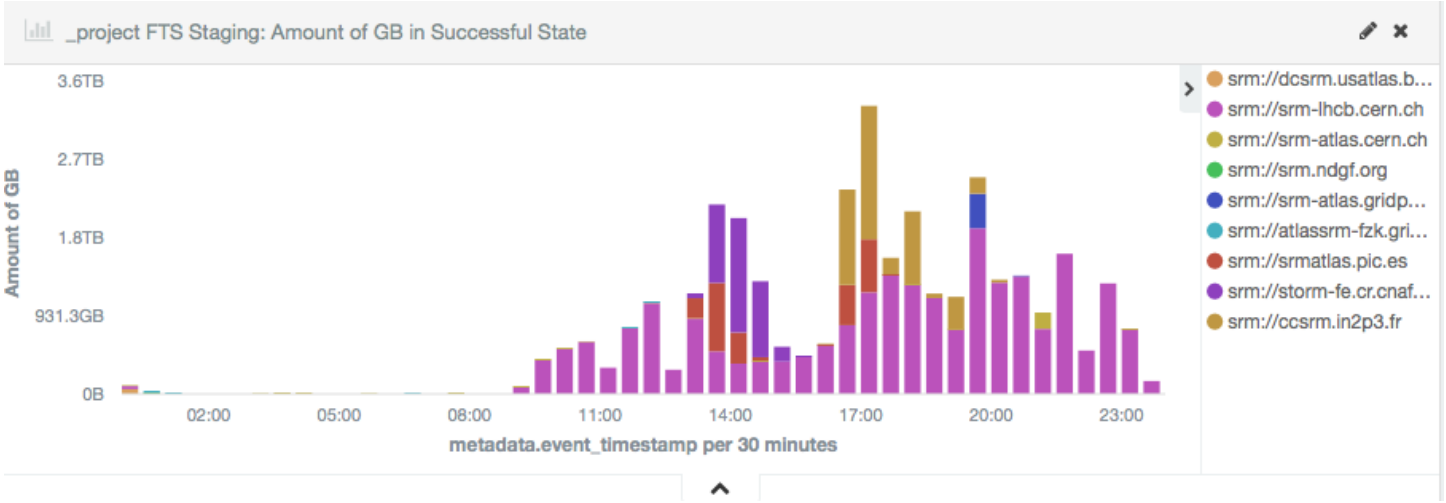


After all, we've got re-read ~10% of all CMS data

- 1 read error
  - recovered after tape copy to new media
- 1 robot arm (out of 4) broken
  - Took more than a week to repair

FTS provides comparable metrics

- D. Amount of GB of successful staging files



CMS local staging and transfers details

- Data are routed from the available sources to the destination, taking into account the link costs, performance and subscription priorities, using priority-window up to a fixed data

download the

o check for data
he files and
s available for

asynchronously.

Fermilab

Issues/Discussion (Atlas)

- Optimising throughput
  - How to balance large requests vs flexibility in priority
- Timed-out bring online requests
  - A “failure” reported by FTS leads to repeated requests -> backlog
- Different policies/behaviour in different sites
- Is the best policy to reduce tape reads as much as possible by intelligence on experiment side?
  - Ideally only reading RAW and recovery of lost files on disk
- What about the HL-LHC era?
  - Does tape have a role?

Tape: outcome

- What are the figures of merit we care about for tape systems?
  - Sites and experiments
- In order to optimise these, what information do the experiments or FTS need about the systems?
  - E.g. how many requests should be submitted at once?
  - What interface should provide this information?
- How do we progress this?
  - Some systematic data-gathering and consensus-building is needed

Resource Reporting: Reminder

- Ongoing process to converge on a set of requirements (or “requests”) to storage providers to support, in the non-obligatory-SRM-era,
  - WLCG-level storage accounting
  - Experiment operations
- https://docs.google.com/document/d/1yzCvKpxsbcQC5K9MyvXc-vBF1HGPBk4vhjw3MEXoXf8/edit?usp=sharing
- Hopefully, we’re on the final iteration

Resource Reporting

- Minimal feedback to last iteration
- Final chance!
- Is this now good enough to request prototypes from the storage providers?
  - This does not represent much work for them
- Aim to finish this by 10th Oct (GDB)

Sign up for WLCG data announcements

- https://e-groups.cern.ch/e-groups/EgroupsSubscription.do?egroupName=wlcg-data
- Please post topics of interest here too.