



# Archival Storage WG Update

Oliver Keeble on behalf of the WG

# Archival WG - overview

- Experiments have been asked to exploit archival storage as much as possible
  - Cost advantages
- This independently triggered two threads
  - “Bottom up” discussions among those responsible for operating tape systems
  - “Top down” discussions in experiment data management activities
- WLCG Archival Storage Group created to follow up
  - Launched with two main topics
    - Optimal exploitation of archival systems by experiments
    - Metric identification and reporting (see Vlado’s slides)
  - Since added
    - Atlas Carousel tests
    - Cost modelling

# Archival site survey

## Queue

What limits should clients respect?

---> Max number of outstanding requests **in number of files or data volume**

---> Max ~~requests submitted at one time~~ **submission rate for recalls or queries**

---> Min/Max bulk request size (**srmBringOnline or equivalent**) **in files or data volume**

Should clients back off under certain circumstances?

---> How is this signalled to client?

---> For which operations?

Is it advantageous to group requests by a particular criterion (e.g. tape family, date)?

---> What criterion?

## Prioritisation

Can you handle priority requests?

---> How is this requested?

## Protocol support

Are there any unsupported or partially supported operations (e.g. pinning) ?

## Timeouts

What timeouts do you recommend?

Do you have hardcoded or default timeouts?

## Operations and metrics

Can you provide total sum of data stored by VO in the archive to 100TB accuracy?

Can you provide space occupied on tapes by VO (includes deleted data, but not yet reclaimed space) to 100TB accuracy?

How do you allocate free tape space to VOs?

What is the frequency with which you run repack operations to reclaim space on tapes after data deletion?

## Recommendations for clients

Recommendation 1

---> Information required by users to follow advice



# Archival site survey - conclusions

- Results for each question have been summarised on the twiki
  - Discussed at June GDB - <https://indico.cern.ch/event/651354/>
- Advice can be classified under
  - Campaign planning
    - Planning, grouping, sync with processing, managing priorities
  - Client behaviour
    - Fill up queues and use bulk submission
  - Write operations
    - Write the way you want to read.
    - Investigating dataset awareness.
- [https://twiki.cern.ch/twiki/bin/view/HEPTape/Survey\\_Conclusions](https://twiki.cern.ch/twiki/bin/view/HEPTape/Survey_Conclusions)
  - Archival sites are diverse and universal advice is hard to identify
- Actions
  - FTS
    - [fts3-pilot.cern.ch](https://fts3-pilot.cern.ch) has been reconfigured to use larger bulks (1000)
      - Trying to quantify the effect
    - Optimisations are planned to allow better filling of queues
      - Updating logic on queue state detection

# Atlas Carousel R&D

- The WG has acted as a forum for discussion on the ongoing Atlas R&D on Carousels
  - A “Carousel” is a moving window of disk residency for a large tape archive
    - Atlas has tested this workflow on all 10 of their T1s
    - Xin Zhao presented results at WG meetings
  - Focused discussions between sites, Atlas and storage devs (dCache in particular)
  - Comparison of site performance, configuration, disk buffers, tape families, hardware etc
    - Understanding bottlenecks, improving metrics
  - WG will follow up on further tests and try to track performance improvements

# Preliminary Results

- Throughput

Site	Tape Drives used	Average Tape (re)mounts	Average Tape throughput	Stable Rucio throughput	Test Average throughput
[1]BNL	31 LTO6/7 drives	2.6 times	1~2.5GB/s	<a href="#">866MB/s</a>	545MB/s (47TB/day)
FZK	8 T10KC/D drives	>20 times	~400MB/s	<a href="#">300MB/s</a>	286MB/s (25TB/day)
INFN	2 T10KD drives	Majority tapes mounted once	277MB/s	<a href="#">300MB/s</a>	255MB/s (22TB/day)
PIC	5~6 T10KD drives	Some outliers (>40 times)	500MB/s	[2] <a href="#">380MB/s</a>	400MB/s (35TB/day)
[1]TRIUMF	11 LTO7 drives	Very low (near 0) remounts	1.1GB/s	<a href="#">1GB/s</a>	700MB/s (60TB/day)
CCIN2P3	[3]36 T10KD drives	~5.33 times	2.2GB/s	<a href="#">3GB/s</a>	2.1GB/s (180TB/day)
SARA-NIKHEF	10 T10KD drives	2.6~4.8 times	500~700MB/s	<a href="#">640MB/s</a>	630MB/s (54TB/day)
[4]RAL	10 T10KD drives	n/a	1.6GB/s	<a href="#">2GB/s</a>	1.6GB/s (138TB/day)
[5]NDGF	10 IBM Jaguar/LTO-5/6 drives, from 4 sites	~3 times	200~800MB/s	<a href="#">500MB/s</a>	300MB/s (26TB/day)

Slide from Xin Zhao  
– more info in his  
HEPiX talk.

<https://indico.cern.ch/event/730908/contributions/3153161/>

[1] dedicated to ATLAS

[2] with 5 drives, later increased to 6 drives

[3] 36 is the max number of drives, shared with other VOs who were not using them during the test

[4] 8 drives dedicated to this test. Will have 22 shared with other VOs in production.

[5] federated T1, 4 physical sites have tapes



# System performance and cost modelling

- Established a connection to the “SPACM” WG.
- The “make the most of tape” story is motivated by its lower cost.
  - The advantage must be demonstrated before sites will make significant further investments
- Have started to discuss how to model the costs of tape systems
  - Driven by IN2P3 (Renaud Vernet)
  - Many questions about what to take into account and what timescales to amortise over
  - All T1s will be asked for a report
    - Discussion will be scheduled with the WG
- WG will evaluate the Carousel model in this context and collaborate with “SPACMWG”.



# Topic status

- Client optimisation and survey – closed
  - FTS modifications are on their way
- Metrics – finalising
  - Summarised in Vlado's contribution
  - More metrics will be added as required
- Carousel R&D – ongoing
  - New rounds of testing scheduled
- Cost modelling – starting up
- Open for other topics and contact with DOMA activities where required

# More Information

- <https://twiki.cern.ch/HEPTape>
- Join the group –
  - <https://e-groups.cern.ch/e-groups/EgroupsSubscription.do?egroupName=wlcg-data-archival-storage>
- Contact Oliver Keeble (CERN) or Vlado Bahyl (CERN)