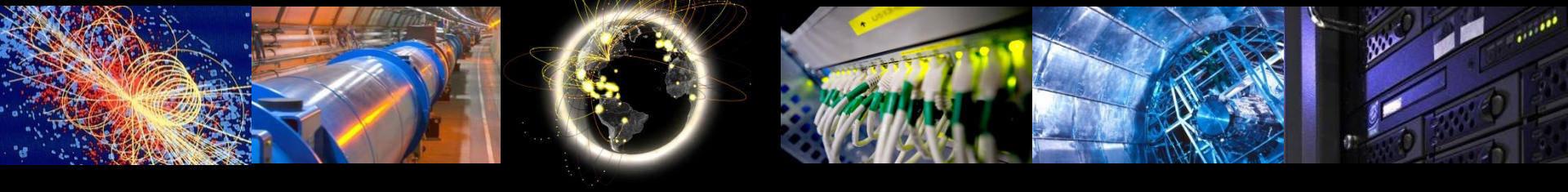
WLCG Archival Storage Group

Vladimír Bahyl, Oliver Keeble – CERN

Report for Spring HEPiX 2018



WLCG
Worldwide LHC Computing Grid

CERN

Raison d'être

- Formed to tackle two main themes:
  - Establish a knowledge-sharing community for those operating archival storage for WLCG
  - Understand how to monitor usage of archival systems and optimise their exploitation by experiments

Achievements

- Established contacts with T1 Archival Storage / Tape Experts
- Performed an archival site survey
  - An attempt to establish best practise for optimally exploiting the archival storage systems used in WLCG
  - A set of questions for each site designed to understand
    - Their advice to clients for efficient use of the system
    - The metrics available to spot anomalies and to track whether proposed improvements are effective
- Constructing common monitoring platform

Archival survey RECALL recommendations

- Inform the site with as much warning as possible about recall plans
  - Allows synchronisation with local activities such as repack
- Use bulk requests, group by creation time or tape family if possible
  - Bulk size not unified, ~1000 is a good reference
- Submit as far in advance as possible, keep the queue full
- Run with no timeouts
- Consider queue size to be unlimited (4 sites have limits)
- Back off on SRM_INTERNAL_ERROR and SRM_FILE_BUSY
- Interaction rates under 10Hz typically acceptable
- Ignore disk buffer occupancy (except: CNAF)
- Understand how priority requests are handled
  - Only partially implemented; no standard interface
- Synchronise data use with recalls to avoid purge/recall loops

- Pinning is not supported (in general)
- Multiple sites support staging via XROOT daemon

- Details with results: https://twiki.cern.ch/twiki/bin/view/HEPTape/Survey
- Survey is still open!

Discussion points & Next steps

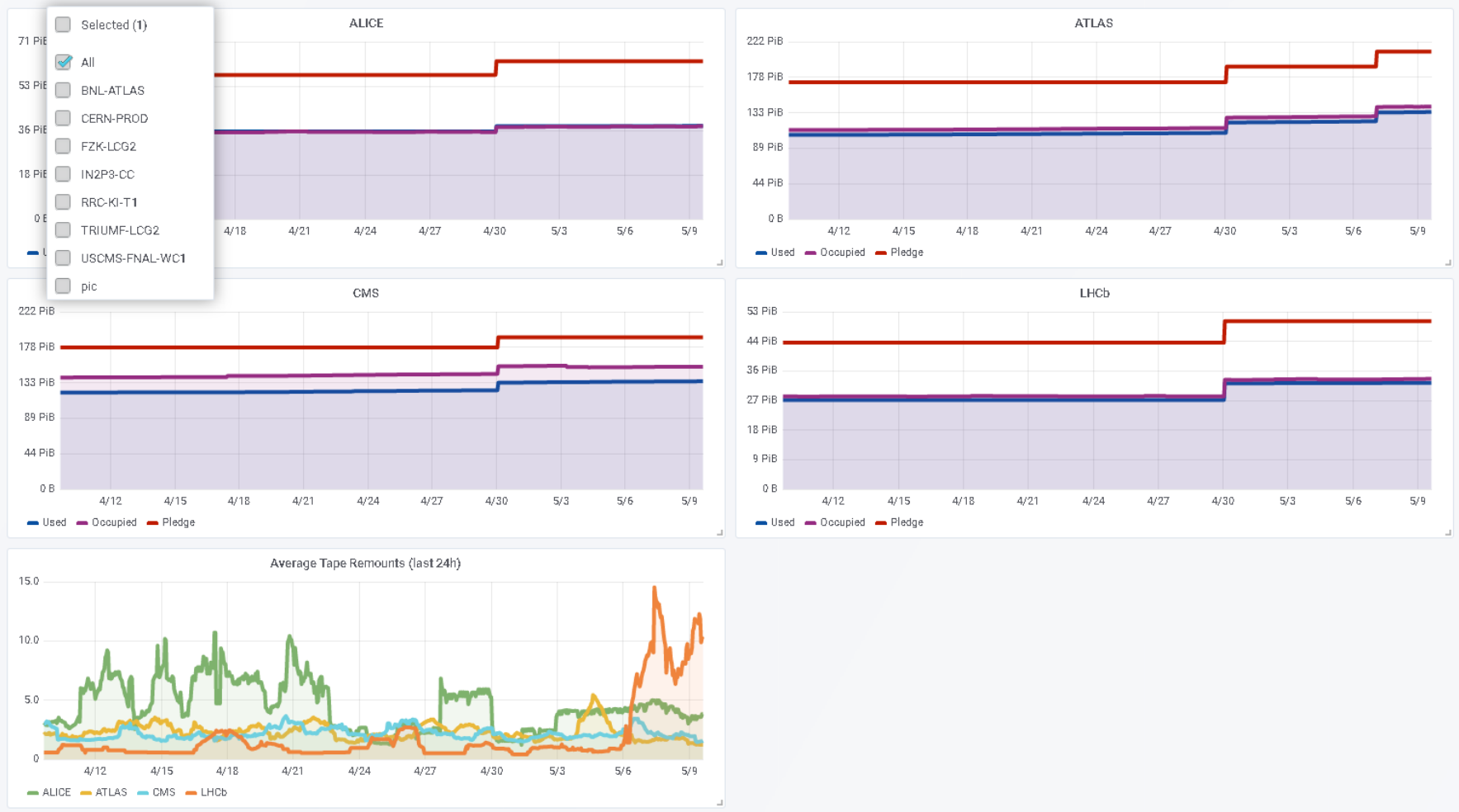
- The survey exposes the significant diversity in these systems. What should the strategy be?
  - Produce some "lower common denominator" advice – do the following... it may help, and will never hurt
  - Enumerate a small number of basic site characteristics and describe each site individually
- Should we try to make any recommendations on writing strategy?

- Recommendations do not give any indication where to make performance gains!

- Finalise the advice to clients in a “mode d’emploi” for tape systems
- Understand what actions, if any, need to be taken client side
- Track progress using the reported metrics
- Other activities
  - Investigate buffer dimensioning and drive allocation strategies
    - RAL & PIC at least
  - Presentation at June GDB to discuss advice with experiments
  - Make contact with the “technology watch” working group

Site Monitoring Dashboard

- http://cern.ch/go/bTt6

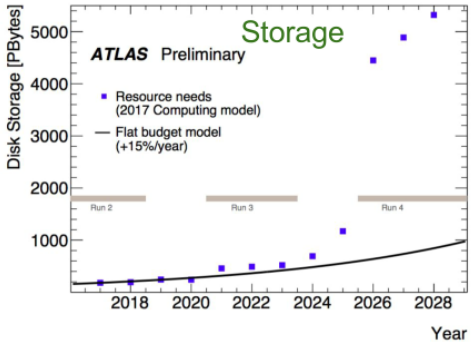


“Tape Carousel” concept

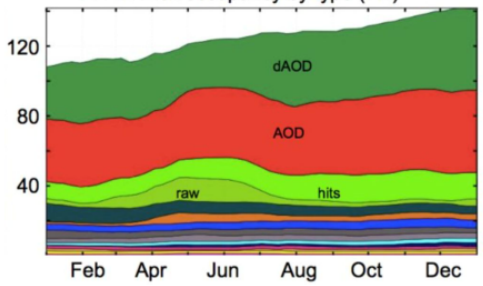
Scaling to HL-LHC: Storage

- ~6x shortfall by today’s estimate, a level that has held ~steady
- ‘Opportunistic storage’ basically doesn’t exist
- Working on format size reductions, but hard to achieve large gains
- Replica counts already squeezed, hard to achieve large gains
- **Storage shortfall is our biggest problem. We need new approaches!**
- ATLAS disk usage is currently ⅔ analysis formats and ⅓ everything else
- Within the ‘⅓ everything else’ are samples that reside mostly on tape, staged onto disk cache when needed for processing
- A way to **dramatically reduce our storage footprint** is to **grow the use of tape** (it looks like our cheap storage will remain tape)
  - Use a ‘**tape carousel**’ approach for the analysis formats
  - A moving window of say ~10% staged to disk at any one time
- **This is hard:** tape is slow and complicates workflow orchestration
  - Analysis workflows are time critical and already complex
- Tape is geographically limited, while processing happens everywhere
- **Fertile ground for R&D...**

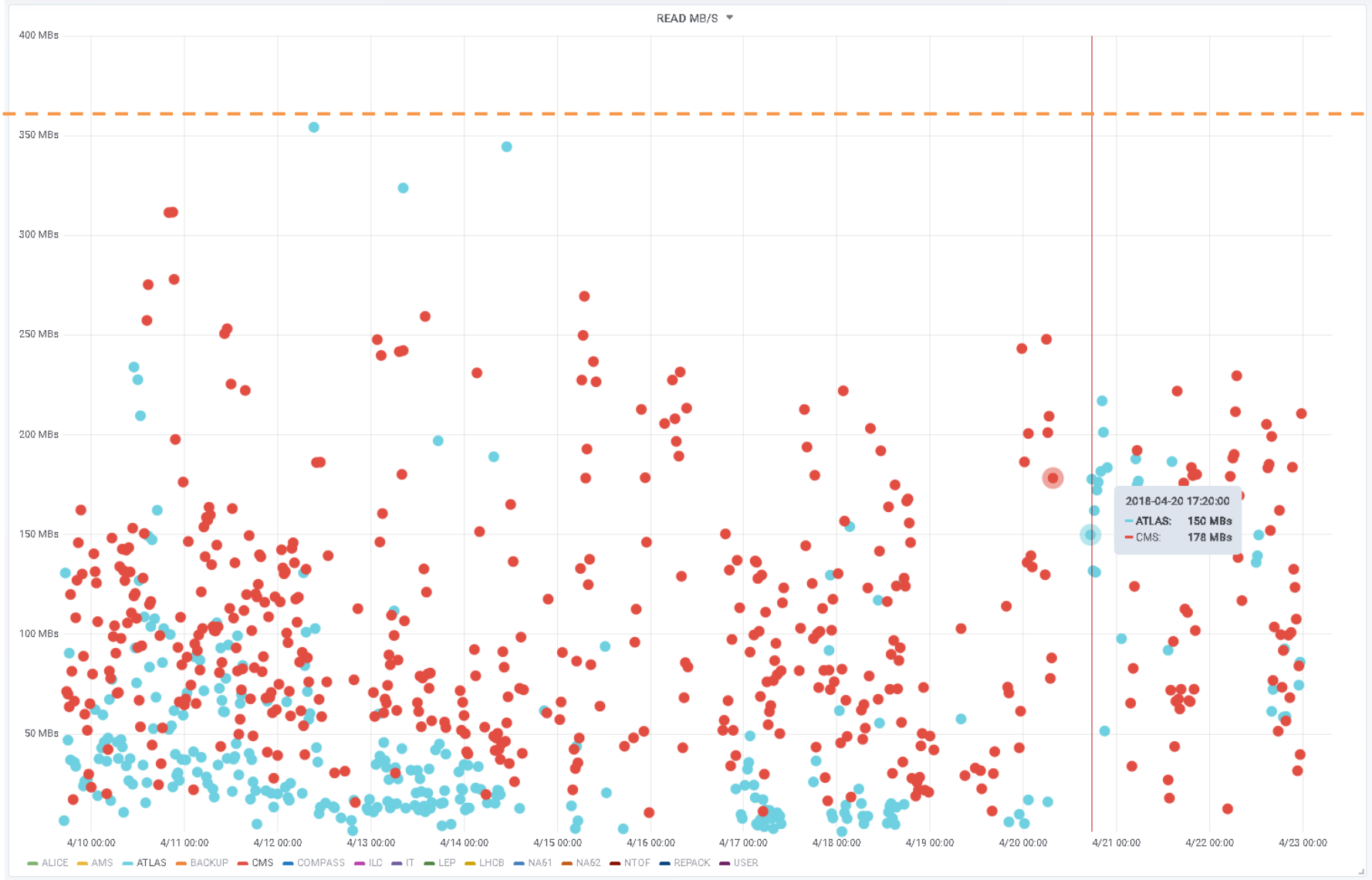
Disk storage ~6x short at HL-LHC



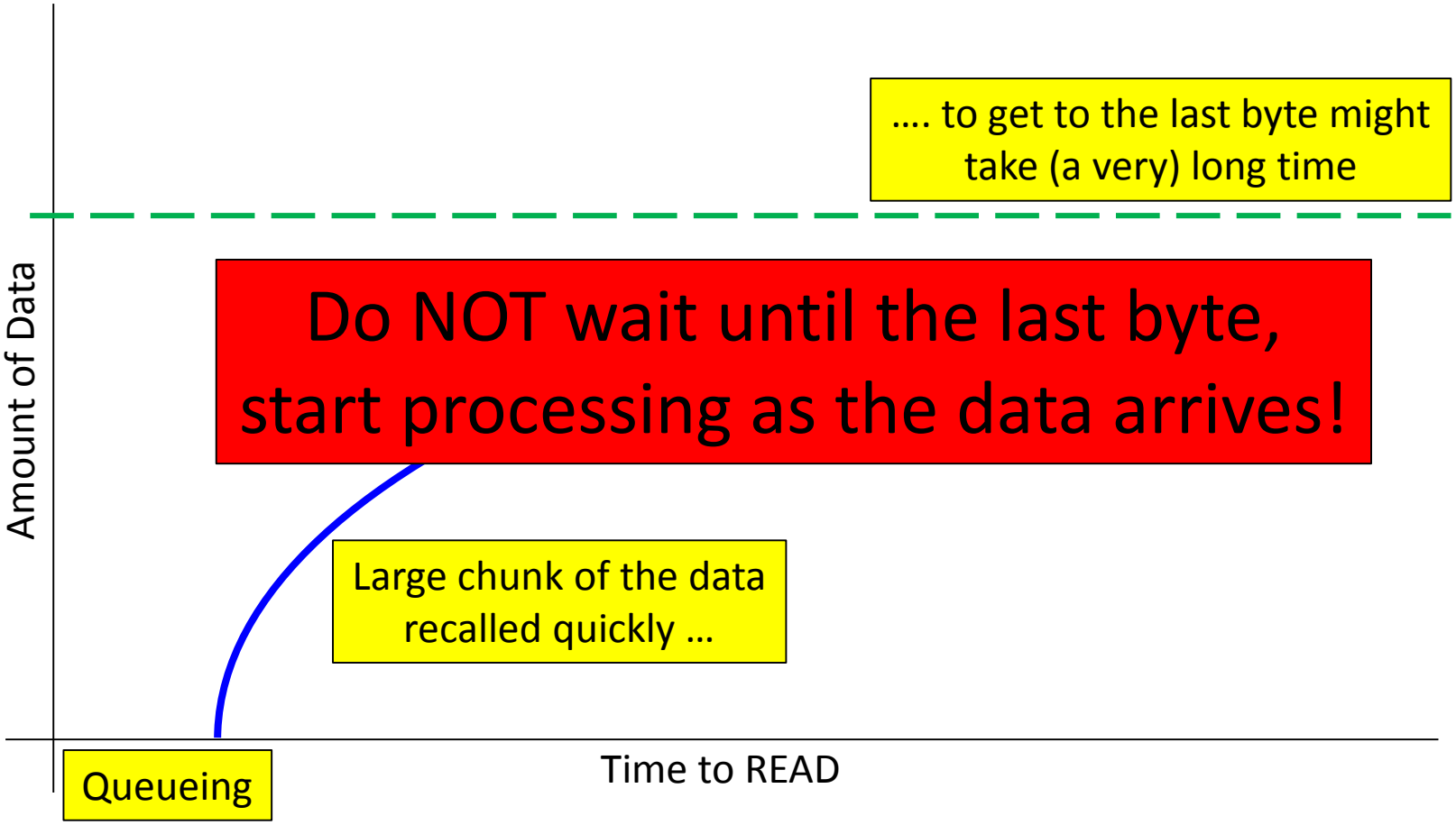
ATLAS disk usage 2017



“Tape Carousel” expectations vs. reality



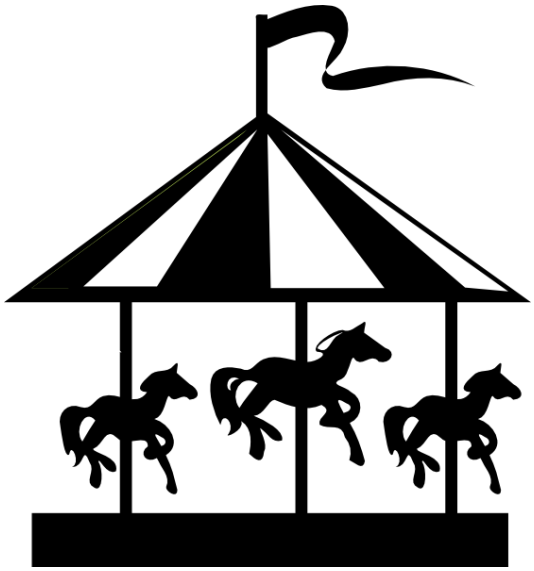
Understanding the RECALLs



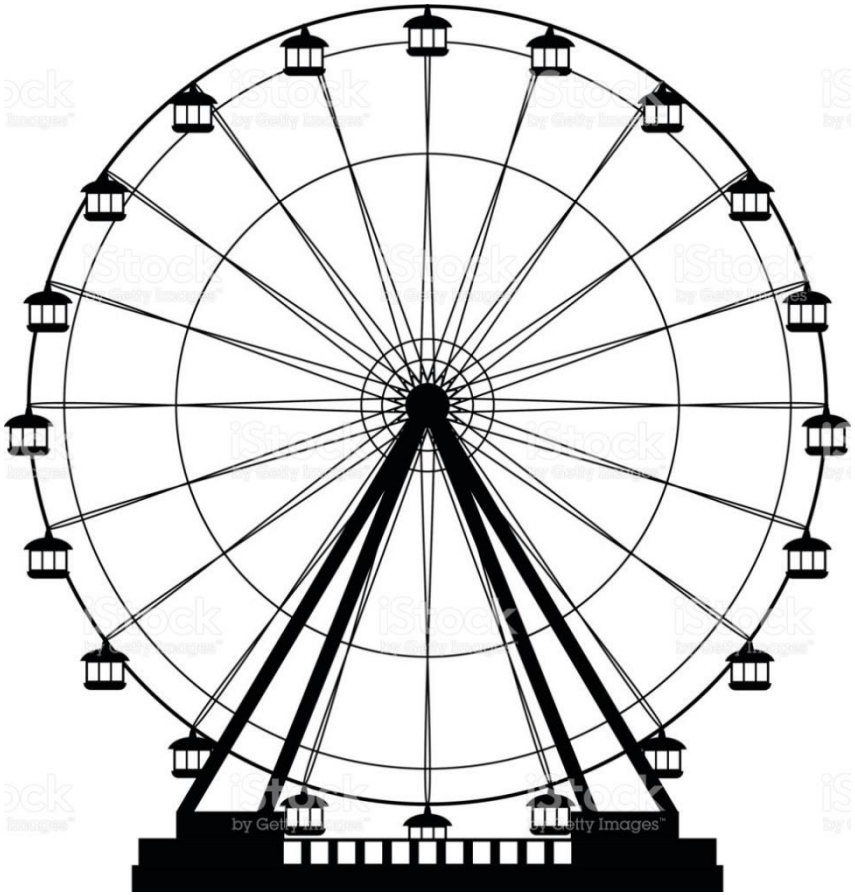
“Tape Carousel” discussion points

- Experiments should not get direct access to tapes
- *We* (the tape site managers) should try to improve the overall throughput of the infrastructure over time, but the experiments shall not make any assumptions on collocation, ordering nor time to completion when retrieving data from tape
- In fact “temporal collocation” of data does not hold in the long run
  - Can be configured and achieved in year X, but X+5 or X+7 years later:
    - The hot data becomes cold
    - Data is usually reshuffled across different tapes due to repack
    - The only collocation that might work is the one within the (larger) file
- More tape drives will be needed for sustained recall activity
  - Total cost of the tape infrastructure will increase
- Fast (but small-ish) buffer in front of tape (SSD based?) will be needed
  - SSD price decrease will continue, but unlikely to get significantly close to the HDDs
- Bulk prestaging is key for allowing internal optimisation of requests (maximising throughput, minimising tape mounts)
- *We* (the tape site managers) need to define what “tape carousel” means depending on what is feasible

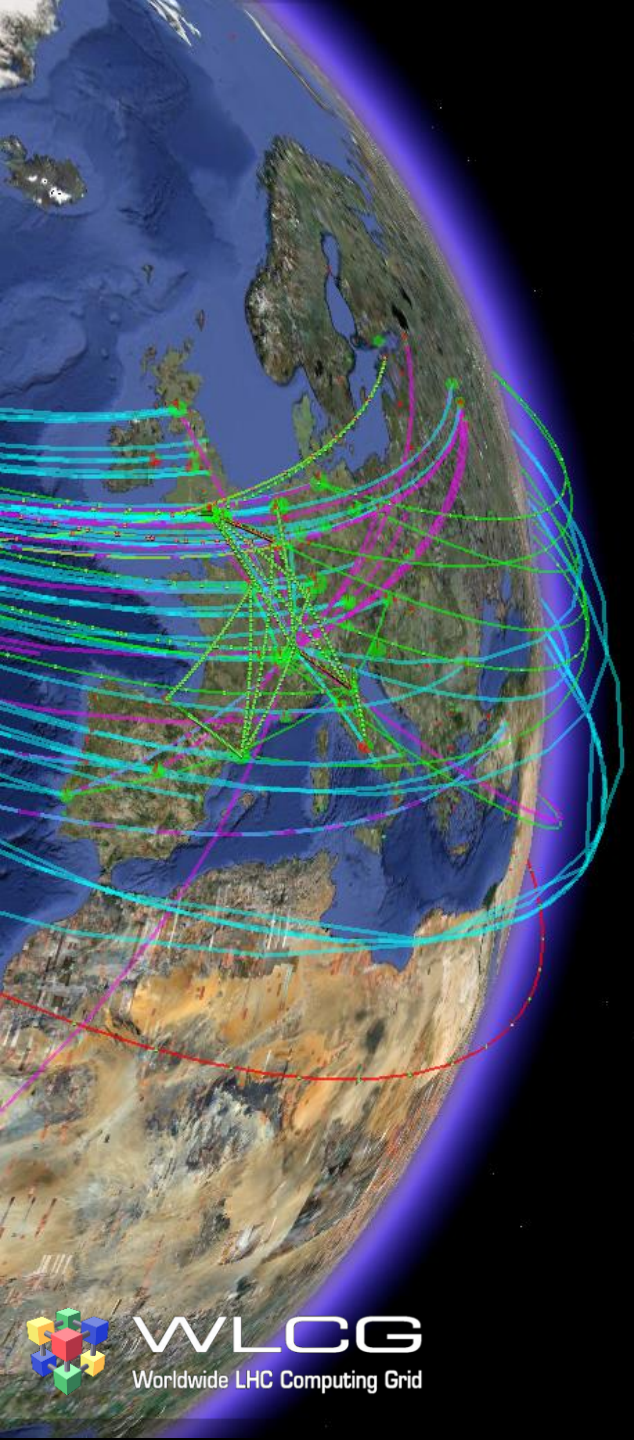
BTW: It's more like a Ferris wheel …



Not allowed to jump in/out.



Continuous process.



Dedicated Tape Storage discussion session

Thursday, 17 May 2018, starting 16:25

WLCG
Worldwide LHC Computing Grid