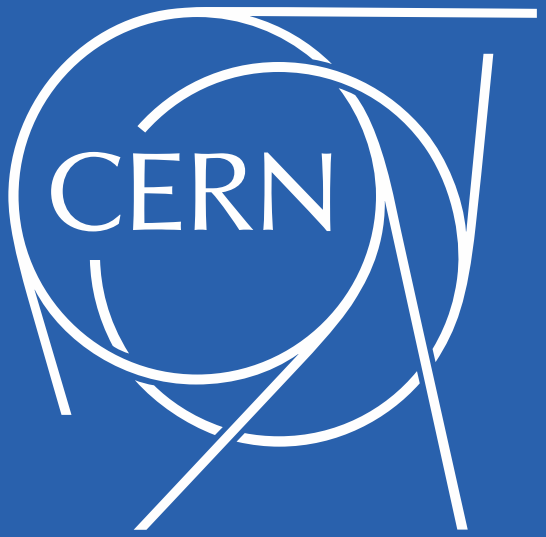

CERN

Smart WRITING for optimal READING

Vladimír Bahyl
CERN IT department

Purpose of this talk

- Experiment (recall) requirements as presented at WLCG DOMA general meeting 28 November 2018:
  - https://indico.cern.ch/event/767209/
- Describe the issue(s) and all aspects
- Make a proposal
- Trigger discussion
- *tape family = tape pool (in CERN terminology)*

Experiment tape storage requirements for Run 3

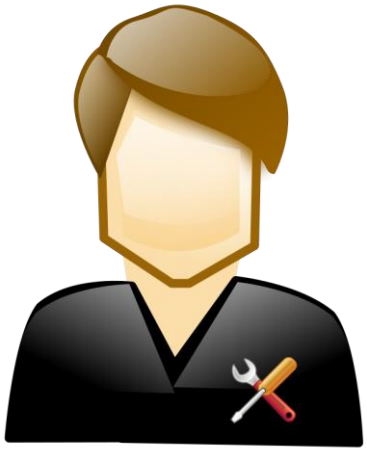
- ALICE – no plan presented at the meeting
- ATLAS
  - Looking at Tape Carousel model as expect to use tape more in the future
  - Now only care about RAW, but in the future they will also distinguish other data types (AOD, HITS, etc.)
  - They think they need finer granularity about tape families (not explained, but I guess to recall files together)
- CMS
  - 6 different data types presented (RAW, RECO, variousAOD, etc.)
  - Each type ~10 PB/year or less
  - Have Logical File Name structure and expect tape families to follow it
  - Usually do 2 deletion campaigns / year in the PB range
  - Reprocessing O(10)PBs of RAW data up to twice per year, other reprocessing insignificant (less than ~1 PB)
- LHCb
  - Mentioned performance (up to 10 GB/s), but not quantities of data
  - 2 recall campaigns during the year + 1 at the end of the data taking year
- **Not processing data from the previous runs**

General experiment expectations

- WRITING
  - Migrate new data as fast as it arrives

- READING
  - The relevant data sets should be recalled together
  - Data should be recalled as fast as possible so the re-processing can start as soon as the last file is available
  - Data recalls might be a continuous process for some experiments (ATLAS Tape Carousel)
  - LHC accelerator runs usually take ~3 years but experiments mostly re-process this or last year's data = data from the current run, not from the previous runs

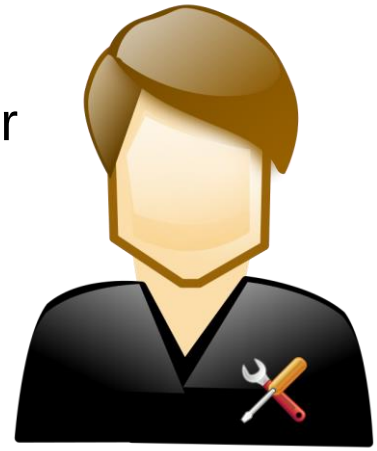
Site (CERN) Tape Admin viewpoint 1/2



- *Tape resources are limited and shared between experiments*
- WRITNIG has priority over READING
- READING:
  - Should be as efficient as possible (reaching >80% of native tape drive data transfer rates) so the tape drives are quickly liberated for WRITING
  - Maximum amount of data should be recalled during each mount – i.e. re-mounts of the same tapes must be avoided as much as possible
- Data should be processed as soon as it arrives because there will always be residual issues to get to the last byte

Site (CERN) Tape Admin viewpoint 2/2

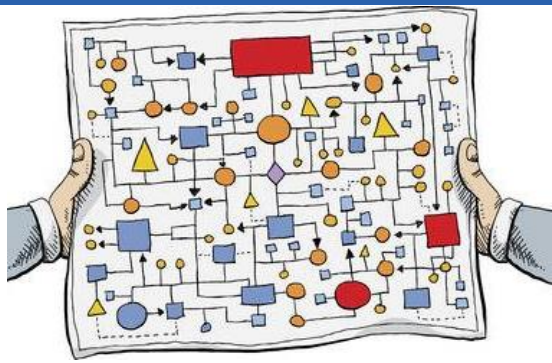
- REPACK
  - Is a migration of data from one tape technology to another
  - Usually needs to be executed fast (= spreading the data across many WRITE streams) in order to gain space / liberate library slots / save money
  - The need to run repack is orthogonal to the LHC run-stop schedule as vendors have their own schedule to put the products on the market



- Preserve Temporal Collocation
  - Files written together should stay together (i.e. on the same tape) (even after repack) as the expectation is that they will also be recalled together

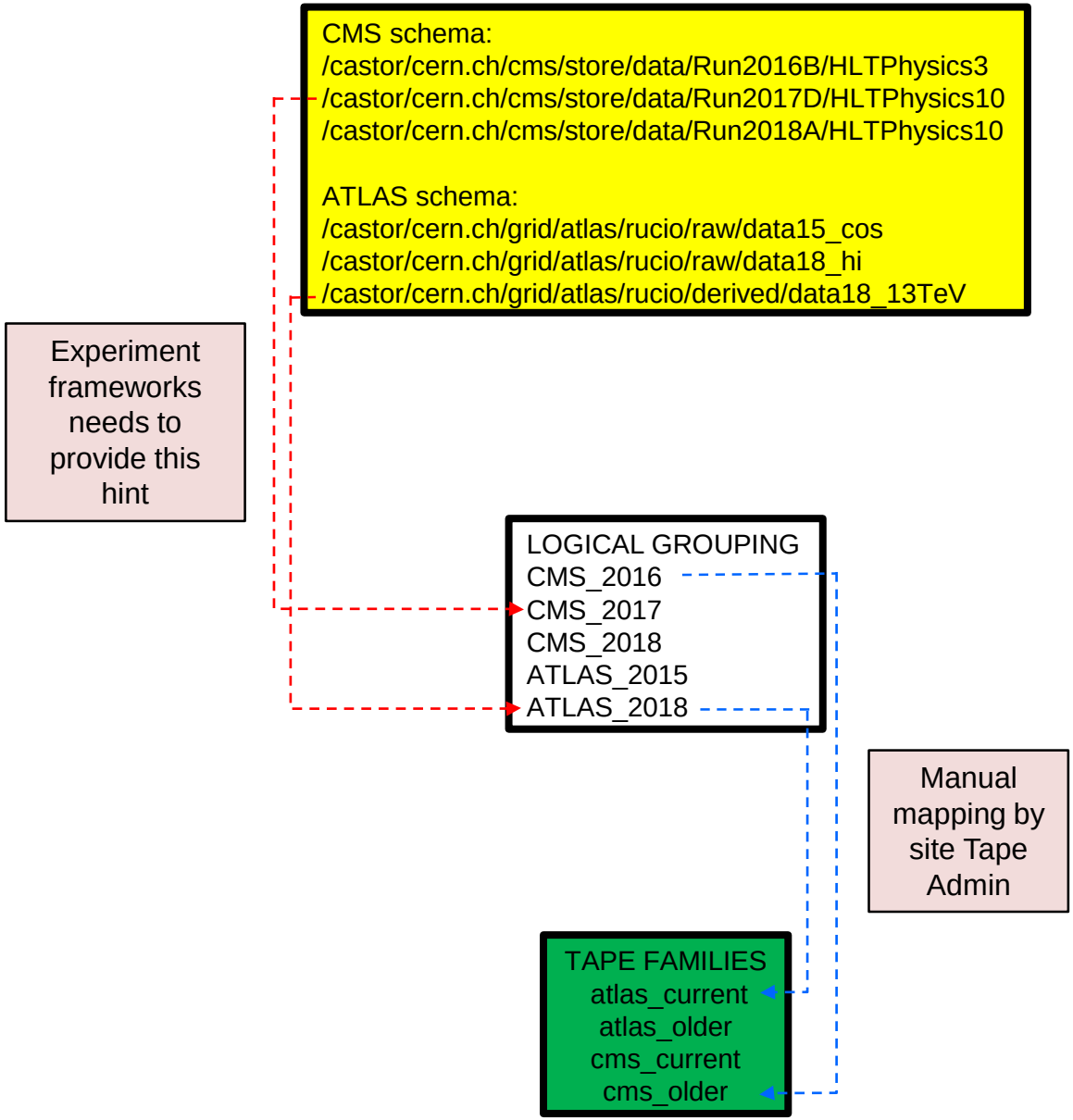
- The number of tape families should be low O(10)
  - If too many O(100), space on partially filled tapes is wasted
  - Management overhead

The Proposal



- At the START of a new LHC Run:
  - The Experiment coordinators together with site Tape Admins will agree on set of logical groups
    - = File/Storage classes in CASTOR/CTA
  - The Experiment coordinators will then make sure that their framework tools (e.g. Rucio) provide a hint with their data to indicate to which logical group it belongs
  - The site Tape Admins will configure mapping of the logical group to limited set of tape families
- At the start of each following data taking year (January – February):
  - Site Tape Admins and the Experiment coordinators will check annually the above mentioned structure
- At the END of the LHC Run (or latest at the start of a new LHC Run):
  - Merge the tapes with the data from the just finished Run with the older tapes in previous tape families, keep the number of tape families low
- Data from the previous LHC Runs
  - Unlikely to be recalled together with the current Run data

The grouping mechanism overview



The grouping mechanism discussion

- The Experiments:
  - Need to understand that proving a grouping hint is in their interest to optimize the recalls later on
  - The list of logical groups has to be periodically reviewed making sure it still matches the directory structure
  - The configuration should be done at the top directory level and sub-directories and all files inherit from it
- The Tape Admins
  - Ensure multiple logical groups map into smaller set of tape families
  - At the end of each Run (or year), they merge the used current tapes with older data and prepare for the next Run (or year)

- This would give freedom of management to both experiments as well as tape admins
- Common solution covering all experiments is needed
- Comments?

What about REPACK and RECALLS?

- REPACK:
  - Files from the current Run might be reshuffled, but will stay mostly together on smaller set of tapes
  - Files from the previous Runs are likely to be mixed with old(-er) data

- RECALLS:
  - The files from the current Run will be placed together so the recalls should be optimal
  - Data from the previous Runs will be more sparsely placed
  - Experiments will certainly understand what if they want to re-process data from previous Runs it will take (a bit) longer to recall

Conclusion

- The Experiments need to provide overview how much and how often they are going to recall the data
  - Table: Data type | Recall Quantity | Recall period
- Tape storage systems need some hints which data belongs / will be recalled together
- Set of Logical Groups should be configured between tape families and the experiment logical data structures (directories)
  - Agreed between The Experiment coordinators and site Tape Admins
  - Reviewed on annual basis
  - New people in both teams need to be trained to understand this need
- The Experiment coordinators will map the data structures to logical groups
- The site Tape Admins will map the logical groups to tape families
- The presented proposal:
  - Fulfills what experiments want
  - Keeps the tape admin management overhead low
  - Is universal enough to work with various tape management systems