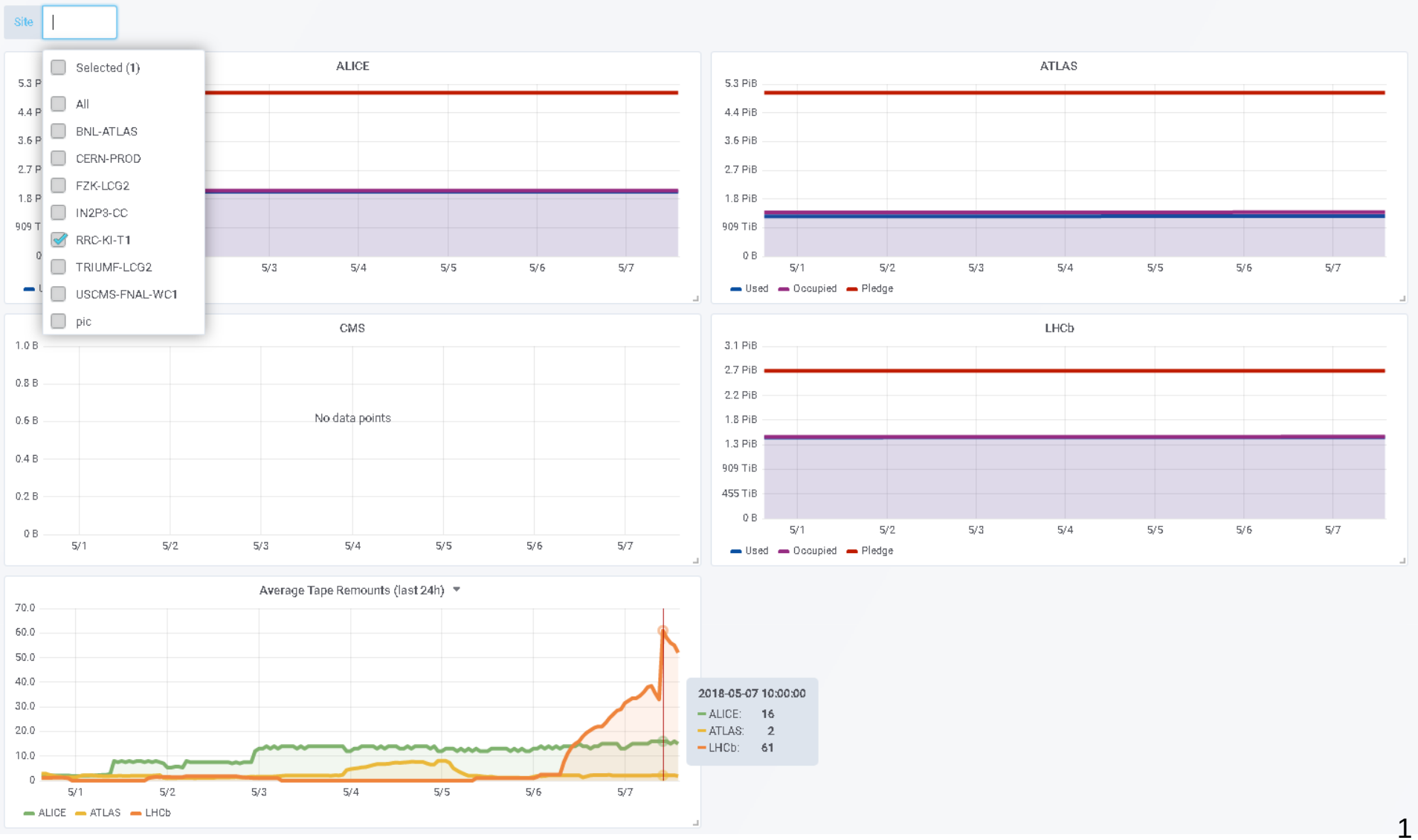
Tape Metrics dashboard

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



Tape Metrics

- https://twiki.cern.ch/twiki/bin/view/HEPTape/TapeMetricsJSON
- 9 sites out of 15 report
- Others?
  - ”Working on it…” – what is the status?
  - Completely silent – why?

- Next step – Tape Drive Performance comparison:
  - Do sites measure the transfer speeds of each individual mount session?
  - Do sites compare the performance of different VOs (i.e. ATLAS is using tape drives better than CMS)?
  - Do sites compare the performance of the tape drives with respect to their native speed (i.e. mount session performed at 150 MB/s, but the drive can do 300 MB/s so the efficiency is 50%)?

Dedicated Tape session @ HEPiX

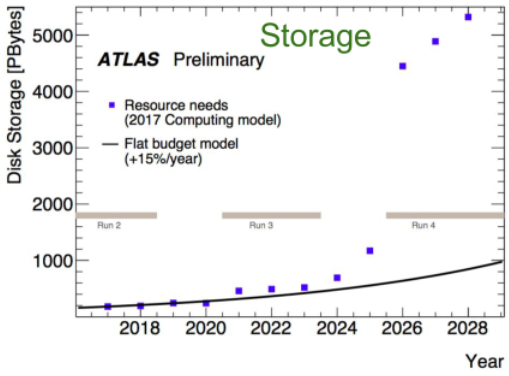
- When:
  - 17 May 2018 (Thursday), starting 17:30
- Where:
  - Spring HEPiX 2018 Workshop
  - Chamberlin Hall
  - University of Wisconsin-Madison
  - USA
- What:
  - Discuss tape topics / issues / solutions
  - CERN will present hear-term plans
  - Other contributions welcome

“Tape Carousel” concept

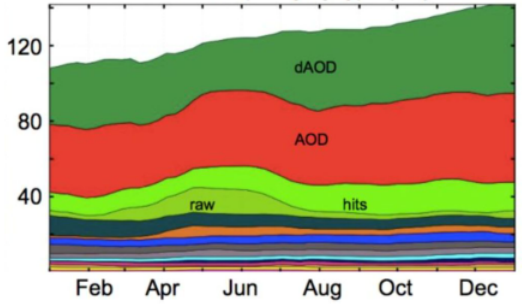
Scaling to HL-LHC: Storage

- ~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

- Tape seen as 3-4x cheaper than disk, can the overall storage cost be lowered by a similar factor?
- Only a (small) subset of the data will fit on disk, the rest has to be periodically recalled from tape
- Will require changes in the workload and storage systems of the experiments
- Mainly needed for HL-LHC, but a new data access model could already be tried for Run-3

“Tape Carousel” reality check (today)

- Tape infrastructure today designed for storage space & data taking rates
  - @CERN: 1 drive per ~400 tapes
- Tape recalls today highly inefficient
  - ATLAS@CERN: 50% of the drive efficiency, 33% of the time spent positioning
  - CMS@CERN: 60% of the drive efficiency, 20% of the time spent positioning
  - CMS@CERN example (3 May 2018): recalling *408* TB of data (could fit on ~30 tapes) spread over *1200* tapes, *~10* days of backlog (drives occupied writing + repack)
- Tape transfers to disk (including overhead) well below native drive transfer rates:
  - ATLAS@CERN: 112 MB/s on 360 MB/s tape drive = 31%
  - CMS@CERN: 123 MB/s on 360 MB/s tape drive = 34%
  - With larger disks number of spindles will decrease even further
- Tape infrastructure:
  - Usually shared between VOs so no guarantee in access sequencing can be expressed
  - Resources (such as drives and tapes) are typically partitioned across multiple independent tape libraries and subject to different workloads

“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
- 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