Archive Metadata for efficient data collocation on tape

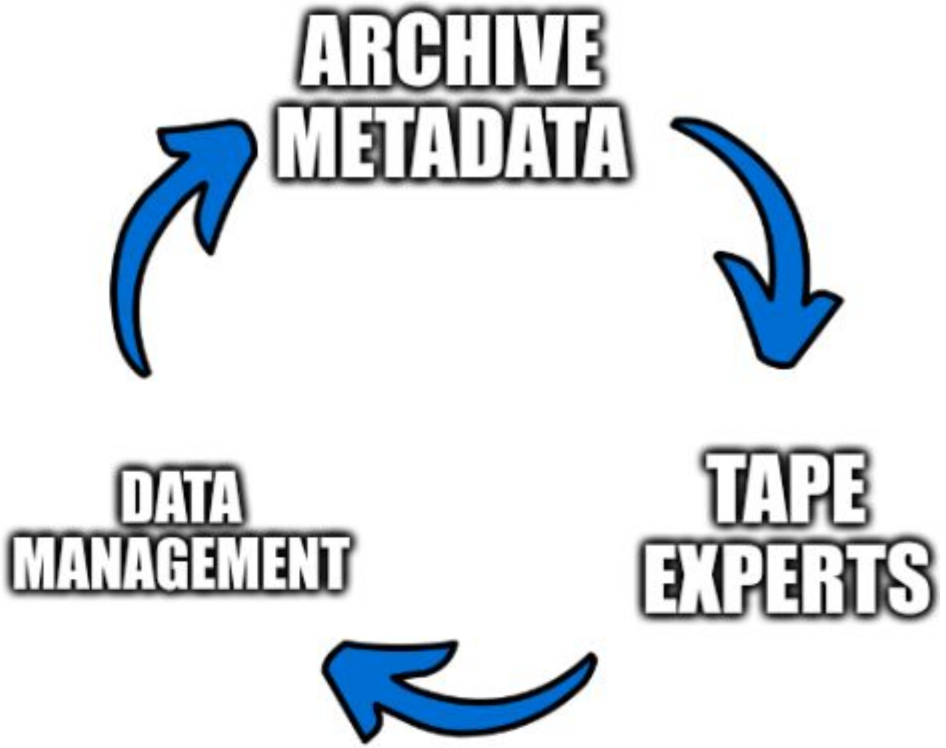
Julien Leduc - CERN Storage group

Archive metadata

Adding metadata along to data stream to solve tape specific issues

What is Archive metadata?

- **HTTP only specific ArchiveMetadata header added to http data transfers**
  - comes along with data transfer hence **outside of WLCG HTTP Tape REST API scope**
- **json object with specific keys that defines a common language that allows**
  - Separation of concerns
    - **Data management express their constraints**
    - **Tape experts decide which hints are relevant**
  - Continuous improvement
    - During tape reads tape experts provide constructive feedback to Data managers

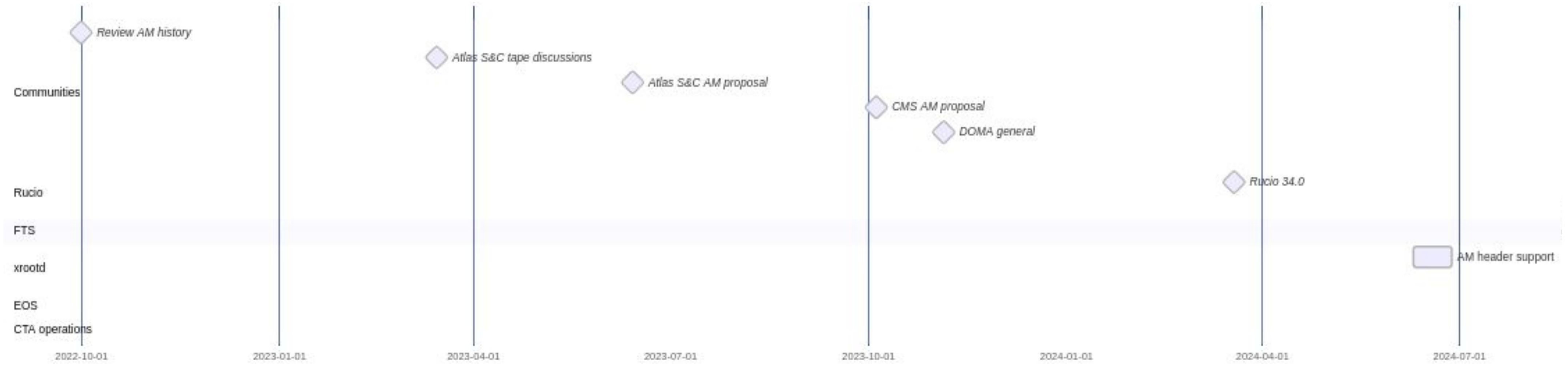


Passing Archive metadata

- **Archive metadata travels through the full experiment data management stack to reach tape storage endpoints**

- **ATLAS + Rucio example:**
  - Rucio generates Archive metadata
  - Add it to every FTS transfer to tape
  - Tape endpoint receives ArchiveMetadata HTTP header **before the first byte of data is received**

Tape software can decide what to do with upcoming file content



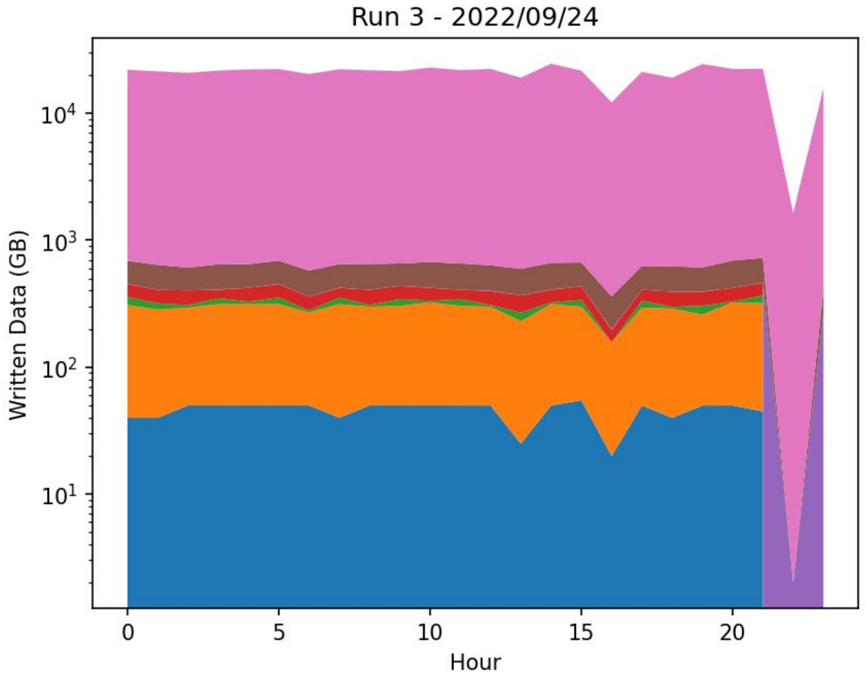
Archive Metadata proposal *In Real Life*?

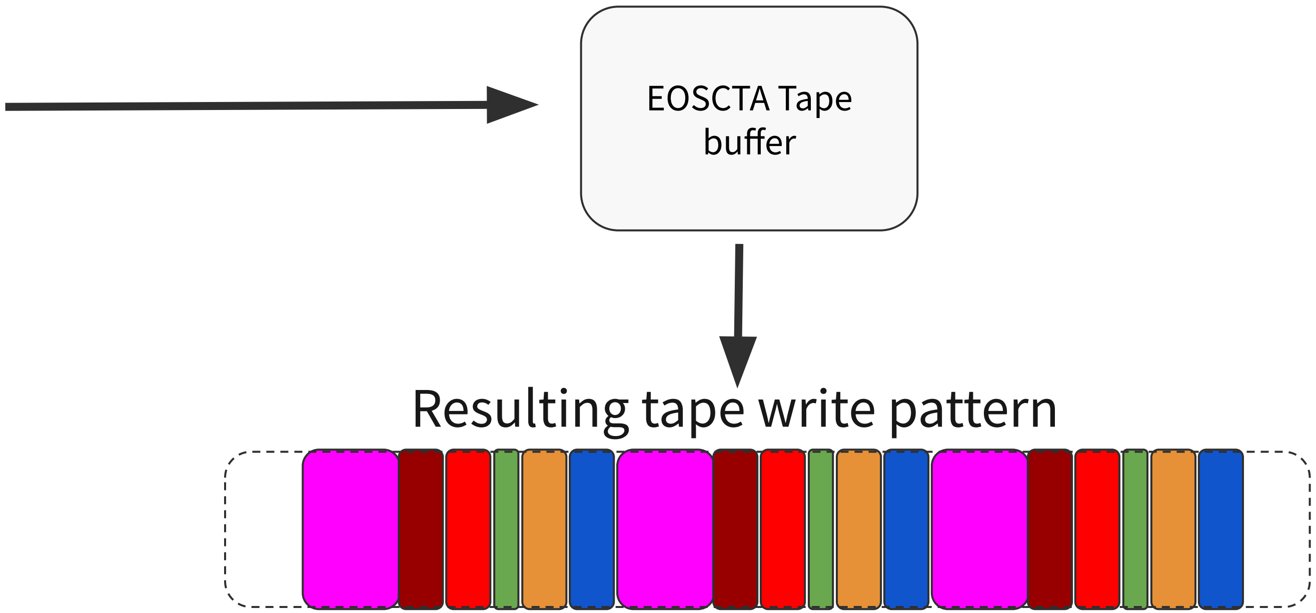
Agreement between Experiments data management, RUCIO, FTS, CTA/dCache, various tape sites

Archive Metadata for Tape Collocation

CERN tape collocation constraints

- **T0 legacy tape collocation mapped on experiment directory structure**
- **T0 tape is low latency very high throughput**
  - 1 tape family for RAW using time based collocation
- **At T0 strict separation or RAW data by tape family by dataset would add too many constraints**



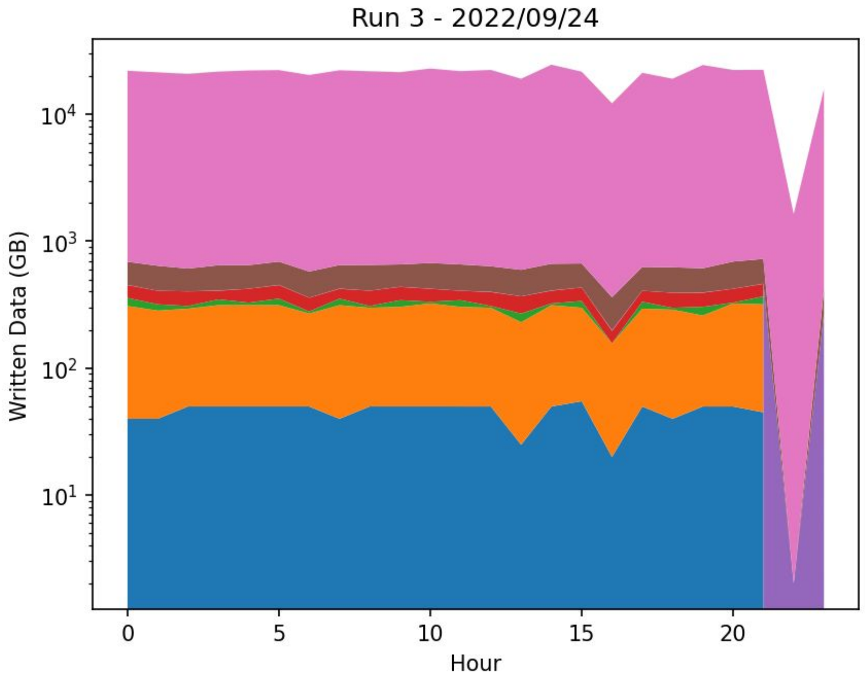


EOSCTA Tape buffer

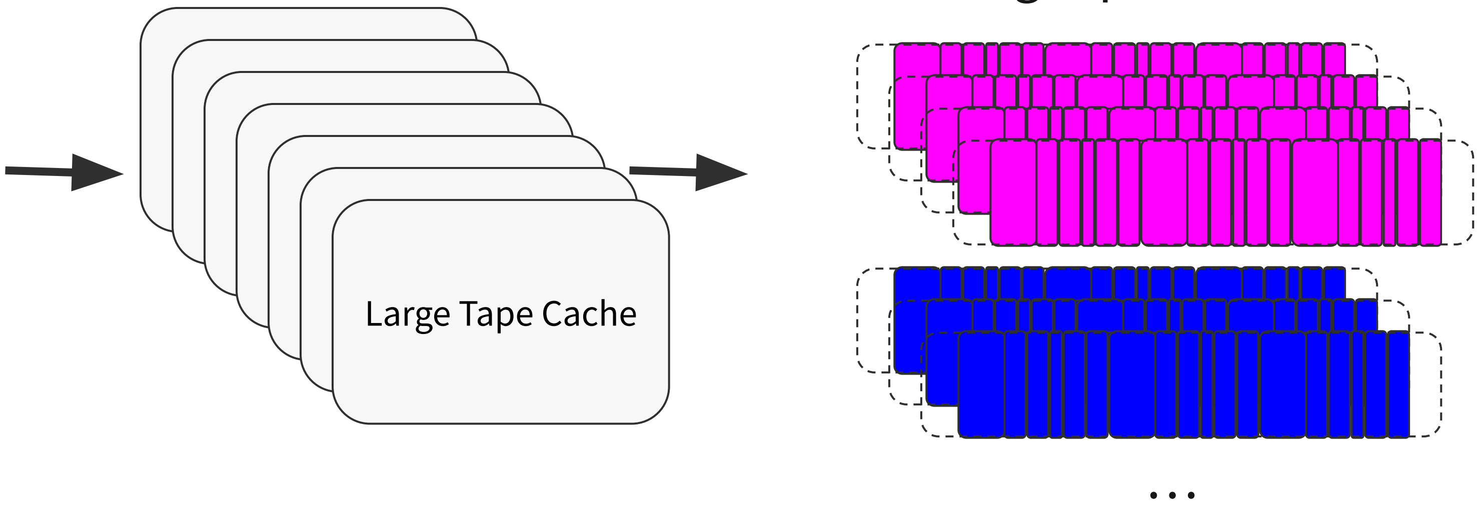
Resulting tape write pattern

Tier1s tape collocation constraints

- **T1s receive out of order delayed transfers from T0**
- **T1s rely on strict tape families to demultiplex streams**
  - Many more tape families needed than T0
  - Logically grouping data transferred over multiple days requires
    - large and expensive tape caches in HSM
    - additional hints that signal logical set completion to trigger flush to tape


Run 3 - 2022/09/24

Resulting tape collection

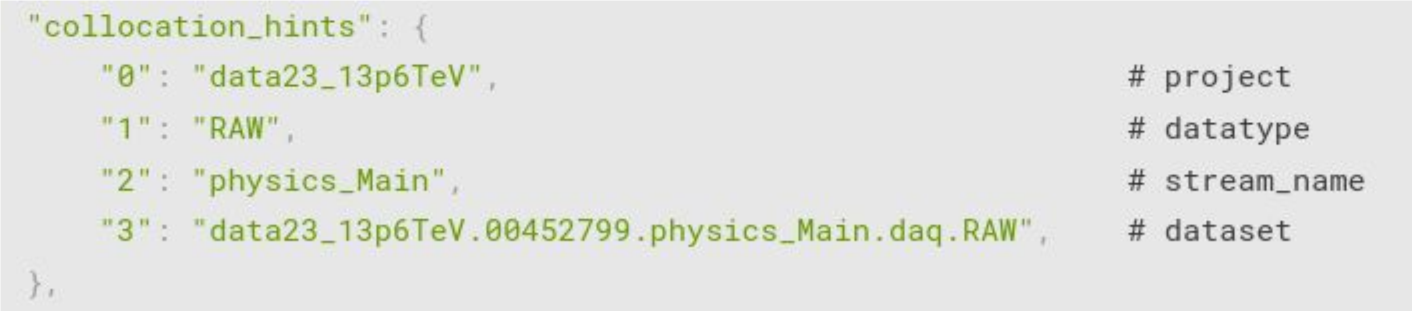

Large Tape Cache

Improving data collocation on tape

Strictly mapping experiment directory structure to tape collocation does not work:

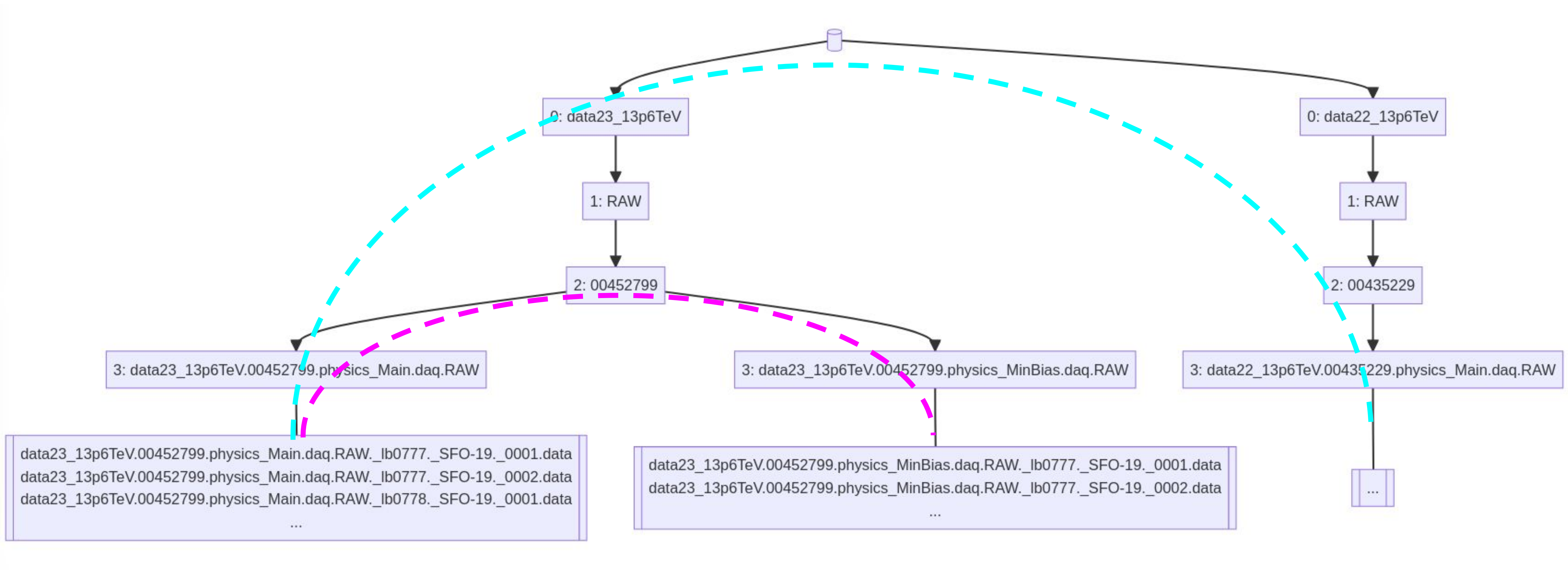
- **Experiment conventions evolve over time**
  - Multi VO sites complexity explodes
- **Flat namespace based on UUIDs like ALICE?**

We need to standardize archive metadata collocation hints independently of experiment namespace structure



```
"collocation_hints": {
    "0": "data23_13p6TeV",                                      # project
    "1": "RAW",                                                 # datatype
    "2": "physics_Main",                                        # stream_name
    "3": "data23_13p6TeV.00452799.physics_Main.daq.RAW",        # dataset
},
```

Archive Metadata defines a mathematical distance



0: data23_13p6TeV

1: RAW

2: 00452799

3: data23_13p6TeV.00452799.physics_Main.daq.RAW

data23_13p6TeV.00452799.physics_Main.daq.RAW._lb0777._SFO-19._0001.data
data23_13p6TeV.00452799.physics_Main.daq.RAW._lb0777._SFO-19._0002.data
data23_13p6TeV.00452799.physics_Main.daq.RAW._lb0778._SFO-19._0001.data
...

3: data23_13p6TeV.00452799.physics_MinBias.daq.RAW

data23_13p6TeV.00452799.physics_MinBias.daq.RAW._lb0777._SFO-19._0001.data
data23_13p6TeV.00452799.physics_MinBias.daq.RAW._lb0777._SFO-19._0002.data
...

0: data22_13p6TeV

1: RAW

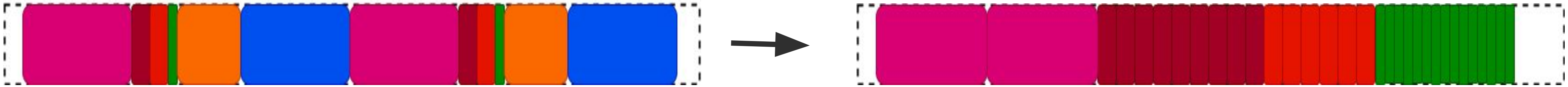
2: 00435229

3: data22_13p6TeV.00435229.physics_Main.daq.RAW

...

Converge toward a flexible and sustainable solution?

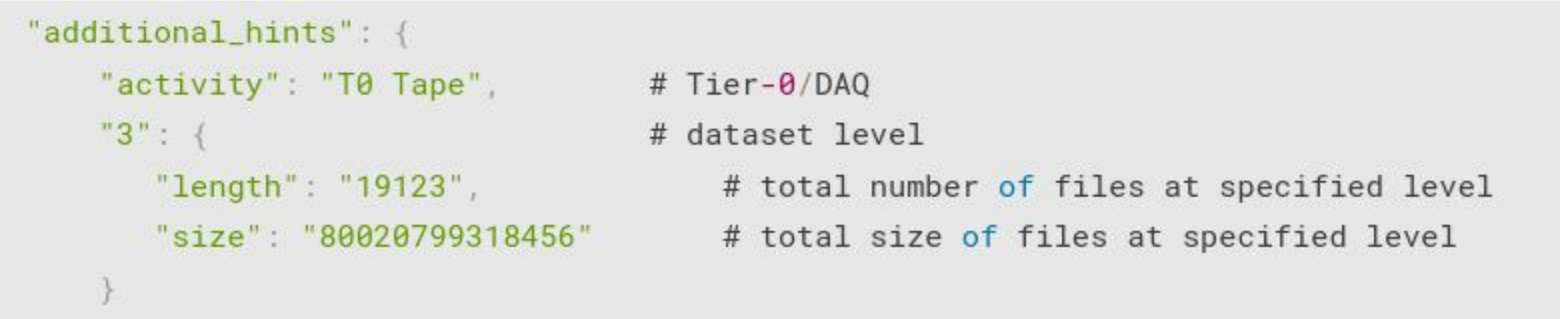
- **collocation_hints**
  - Define tree structure using tree depth as key
  - Define a mathematical distance between files on the tree structure: for example using node distance in collocation hints tree
    - Use it to improve collocation during tape repack operations
    - Evaluate worst collocated tapes



Naïve first order approximation: Could we use total geometrical distance between file read sequentially as cost model for data placement on tape?

"additional_hints" for collocation

- **additional_hints**
  - Expresses collocation subtree properties at specified level:
    - "length": "number of files"
    - "size": "size in bytes"
    - For ATLAS level 3 is the dataset level
    - Allows Tier1s to understand how the data will fit in the tape cache according to site flush to tape policy
  - "activity"
    - suggested by Rucio but not for all experiments…



```
"additional_hints": {
    "activity": "T0 Tape",         # Tier-0/DAQ
    "3": {                         # dataset level
       "length": "19123",              # total number of files at specified level
       "size": "80020799318456"        # total size of files at specified level
    }
```

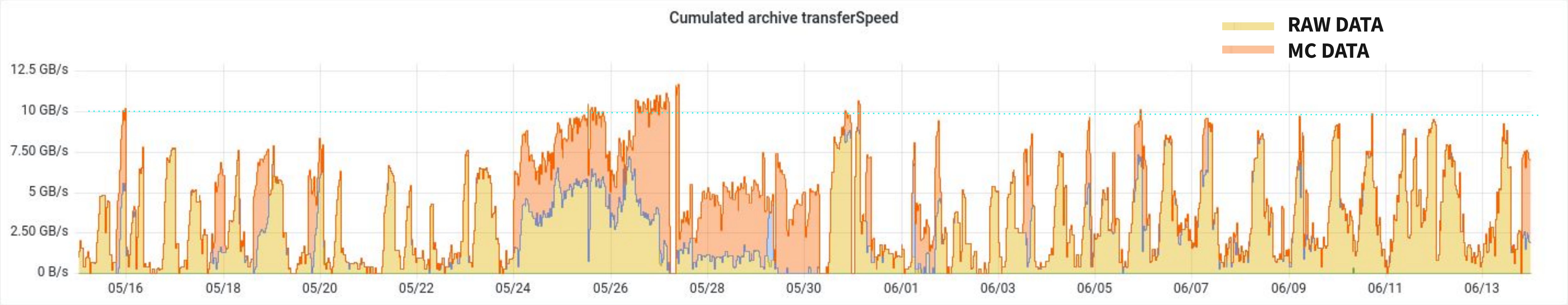
Archive Metadata for Traffic Arbitration

aka Archive backpressure

Improve tape scheduling

- **CMS:**
  - RAW must go ASAP to tape
  - MonteCarlo can wait for beam dump

Would allow to move large chunk of total T0 traffic outside of peak



Improve overall efficiency with less tape hardware

Improving tape efficiency does not only mean increasing tape bandwidth peaks during stable beam

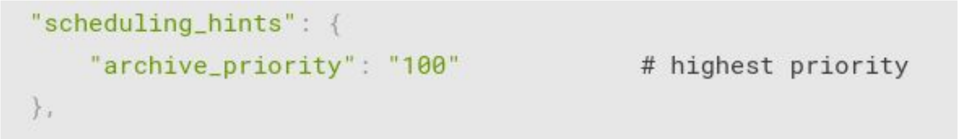
- **DAQ infrastructure tied with detectors**
  - Fixed max throughput for the run is defined by DAQ buffer hardware choices set at the very end of previous LS for all experiments
- **Secondary traffic is very likely to increase during the run**
  - Reschedule secondary traffic to increase operational margins
  - Improve overall tape efficiency using more tape hardware when there is no beam

Provide scheduling_hints to allow tape site to reschedule secondary traffic

Toward a common solution for Archive Backpressure?

There is no point accepting files in tape buffer/cache if their time to tape is expected to exceed agreed Service Level Agreements or compromise site tape operations

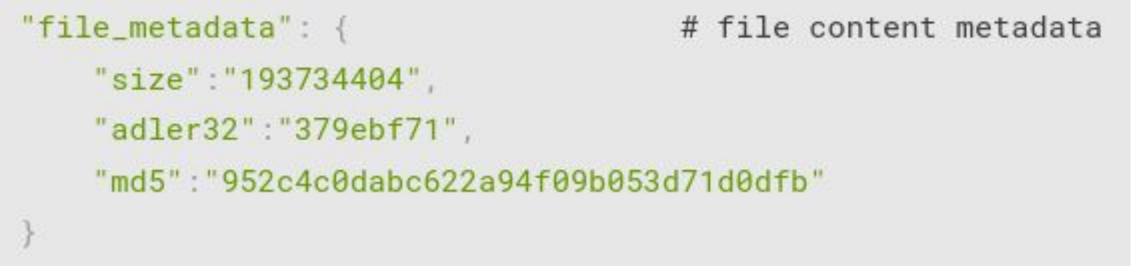
- **scheduling_hints**
  - **archive_priority: "0" to "100"**
    - "0" is lowest priority, 100" is highest
    - Rucio policy deduces value from activity
  - If bandwidth to tape is too constrained
    - Exceeding allocated experiment pledge
    - Sudden loss of bandwidth (tape hardware failure on site,...)
  - Allow to apply *backpressure* on archive transfers
    - Protect RAW data transfers



```
"scheduling_hints": {
    "archive_priority": "100"            # highest priority
},
```

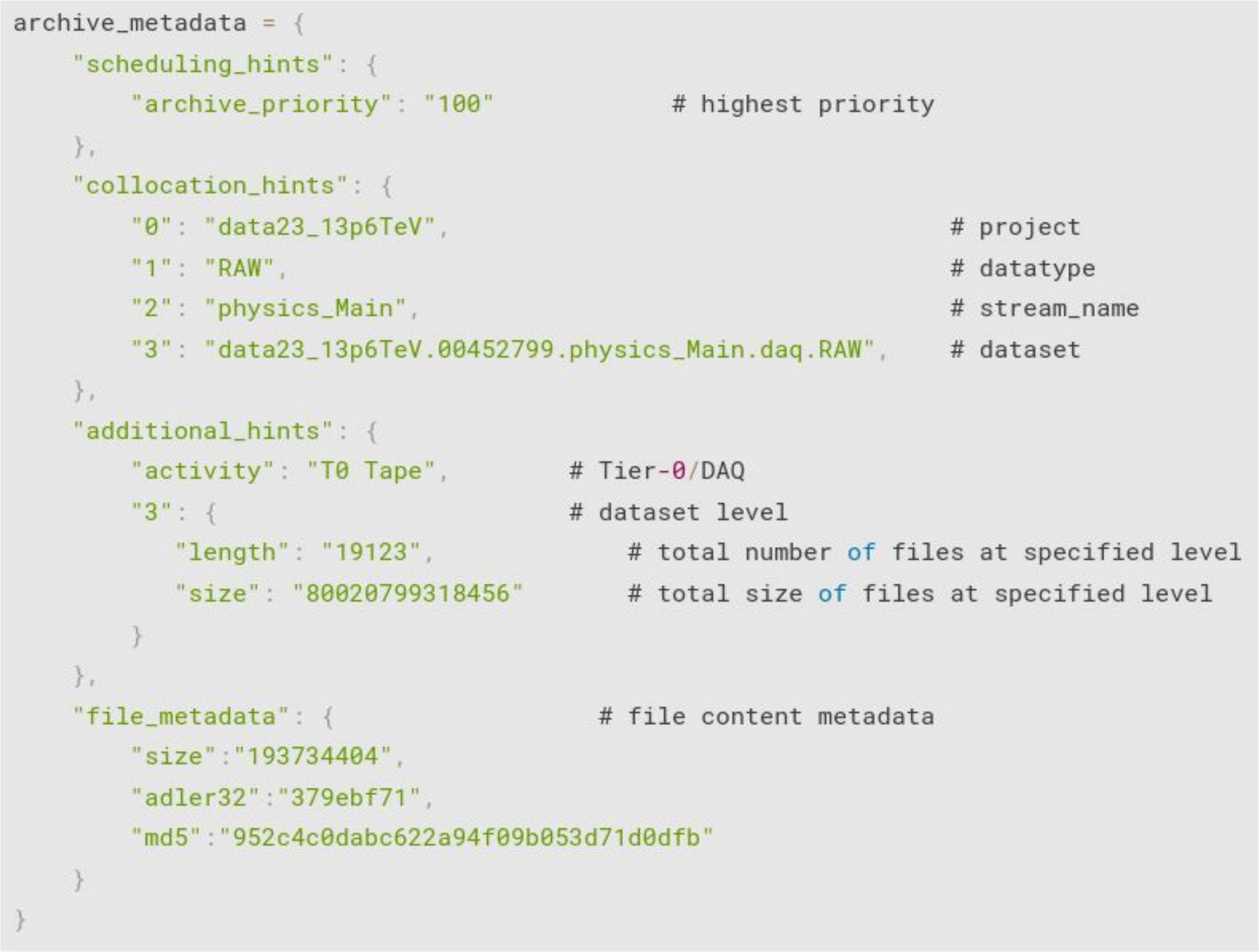
"file_metadata"

- **Provide file size and multiple checksums in archive metadata**
  - Allows tape sites to evaluate file integrity
    - delete on close was coming for free with *xrootd*
      - this is over in HTTP world
  - If at close time file size and checksum do not match file Archive Metadata it should be *deleted on close*



```
"file_metadata": {                        # file content metadata
    "size":"193734404",
    "adler32":"379ebf71",
    "md5":"952c4c0dabc622a94f09b053d71d0dfb"
}
```

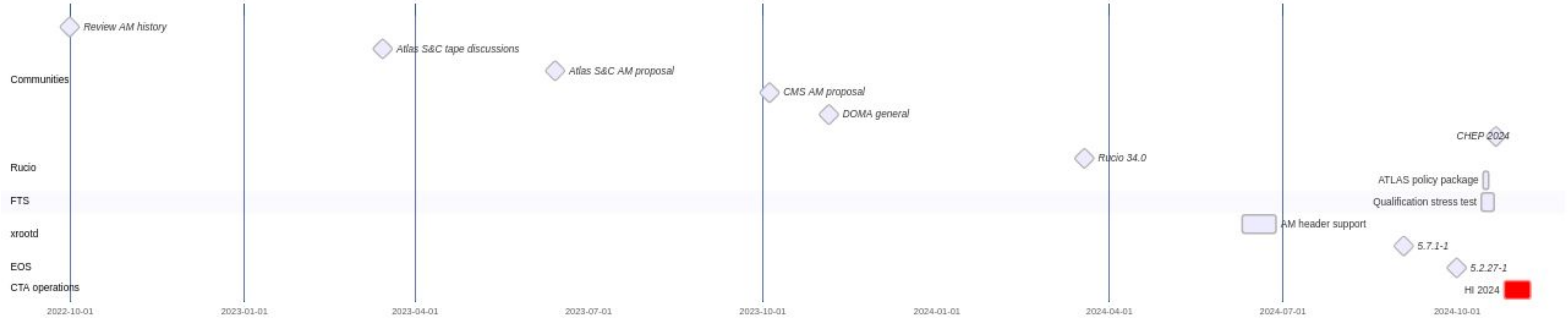
Archive Metadata example



```
archive_metadata = {
    "scheduling_hints": {
        "archive_priority": "100"            # highest priority
    },
    "collocation_hints": {
        "0": "data23_13p6TeV",                                  # project
        "1": "RAW",                                             # datatype
        "2": "physics_Main",                                    # stream_name
        "3": "data23_13p6TeV.00452799.physics_Main.daq.RAW",    # dataset
    },
    "additional_hints": {
        "activity": "T0 Tape",        # Tier-0/DAQ
        "3": {                        # dataset level
            "length": "19123",            # total number of files at specified level
            "size": "80020799318456"      # total size of files at specified level
        }
    },
    "file_metadata": {                  # file content metadata
        "size":"193734404",
        "adler32":"379ebf71",
        "md5":"952c4c0dabc622a94f09b053d71d0dfb"
    }
}
```

Archive Metadata for ATLAS Heavy Ion run 2024

- **How this works in practice?**
  - Rucio extracts and generates Archive metadata via a Rucio plugin
    - Base Archive Metadata plugin released in Rucio 34.0
    - ATLAS specific plugin implemented and deployed in production via *atlas-rucio-policy-package*
  - Passes file specific AM for every FTS transfer where destination RSE is tape AND RSE property archive_metadata=True
    - in –*archive-metadata* FTS option
  - eosctaatlas instance receives and collect ArchiveMetadata for later tape placement analysis



Outlook

- **CTA delivers nominal archival performance for Run3 with significant write efficiency improvements**
  - Initially limited data placement features
- **Ongoing WLCG Tape software and protocol consolidation**
  - Opportunity to formalize and consolidate tape dataflows should not be missed
- **NEXT STEP clearly oriented toward monitoring and improving data placement for tape data reads**
  - HTTP only
  - Currently supported in Rucio+FTS+EOS/CTA chain
  - **Tape sites need something simple quickly**
    - Working on tape data placement improvements requires:
      - Better understanding of experiment archive traffic constraints
      - Gives feedback to DM teams better understanding of traffic conditions

Will be exercised during ATLAS HI 2024 data taking later this month