Science and Technology Facilities Council

Antares - the first year in production

George Patargias on behalf of the Antares team

EOS 2023 Workshop
26/04/2023

Introduction

- March 2022: Antares entered production – take part in a data challenge
- May 2022: First CTA upgrade from 4.0.5 to 4.6.1-1
- June 2022: First Run 3 data
- November 2022 – now: Focus on merging CASTOR Facilities instance
- December 2022 – January 2023: Repeat of data challenge for LHCb
- February 2023: Second CTA upgrade from 4.6.1-1 to 4.7.14-1

Antares usage

VO	Data volume	Pledge
ATLAS	42.2 PB	53 PB
CMS	14.8 PB	24.4 PB
LHCb	14.3 PB	39.9 PB
ALICE	921.4 TB	2.5 PB
DIRAC	2.6 PB	3.0 PB
Other	11.3 PB	-

Antares: Main features

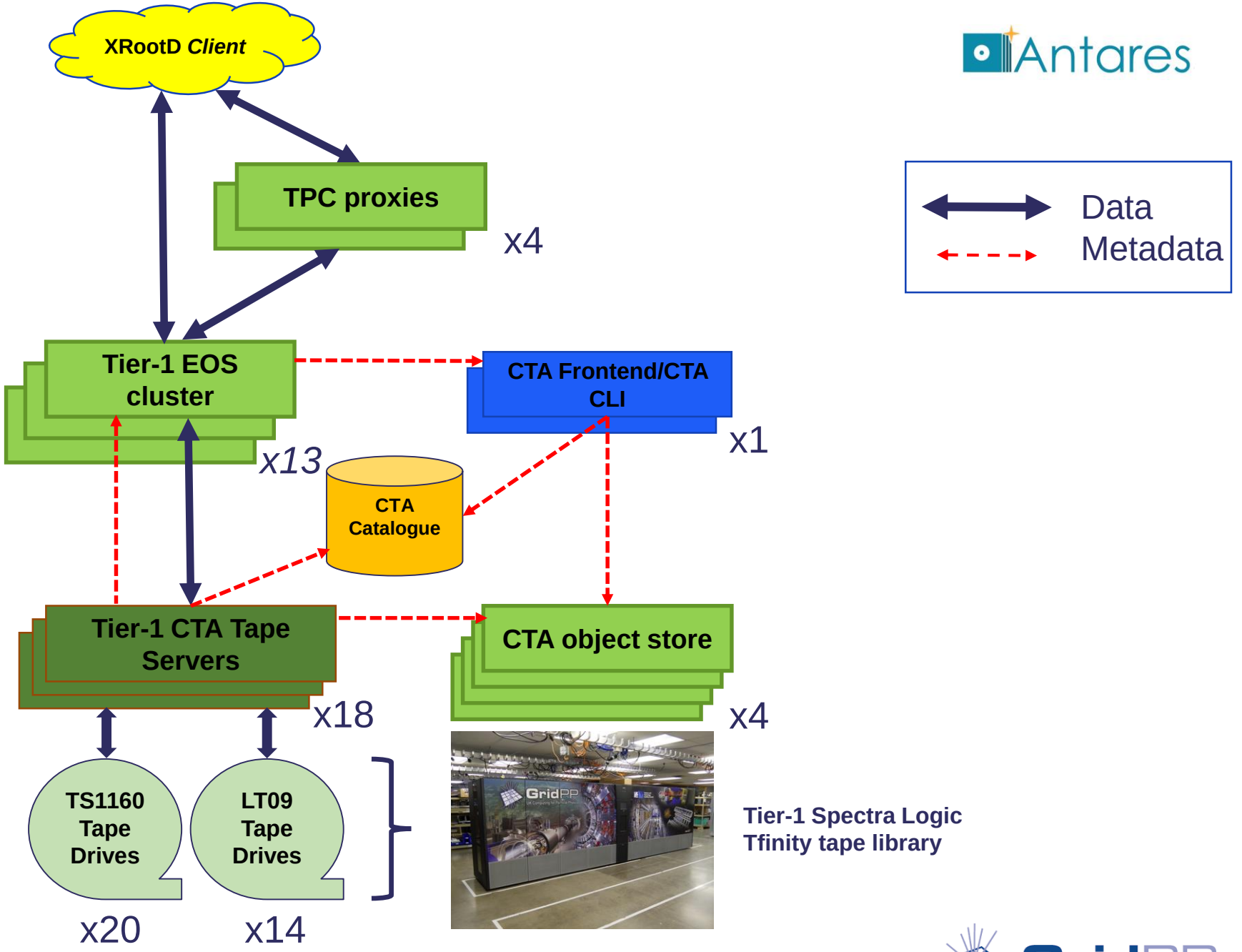
- Single EOS cluster (3 QDB/1 MGM/13 FST nodes) for LHC/non LHC
- Default space: 173TB (90 FS), Retrieve space: 200TB (105 FS)
- XRootD proxy servers for TPC transfers with delegation
- Full XRootD tape API - partial (only writes) WebDAV tape API
- Backpressure: 40TB (targeted free space), 1800s (sleep time)
- Tape aware GC: Min free space: 500GB/FS, 14400s < GC age < 86400s

Antares: Current setup

Tier-1 EOS cluster in front of one CTA instance

EOS: 4.8.88-1

CTA: 4.7.14-1

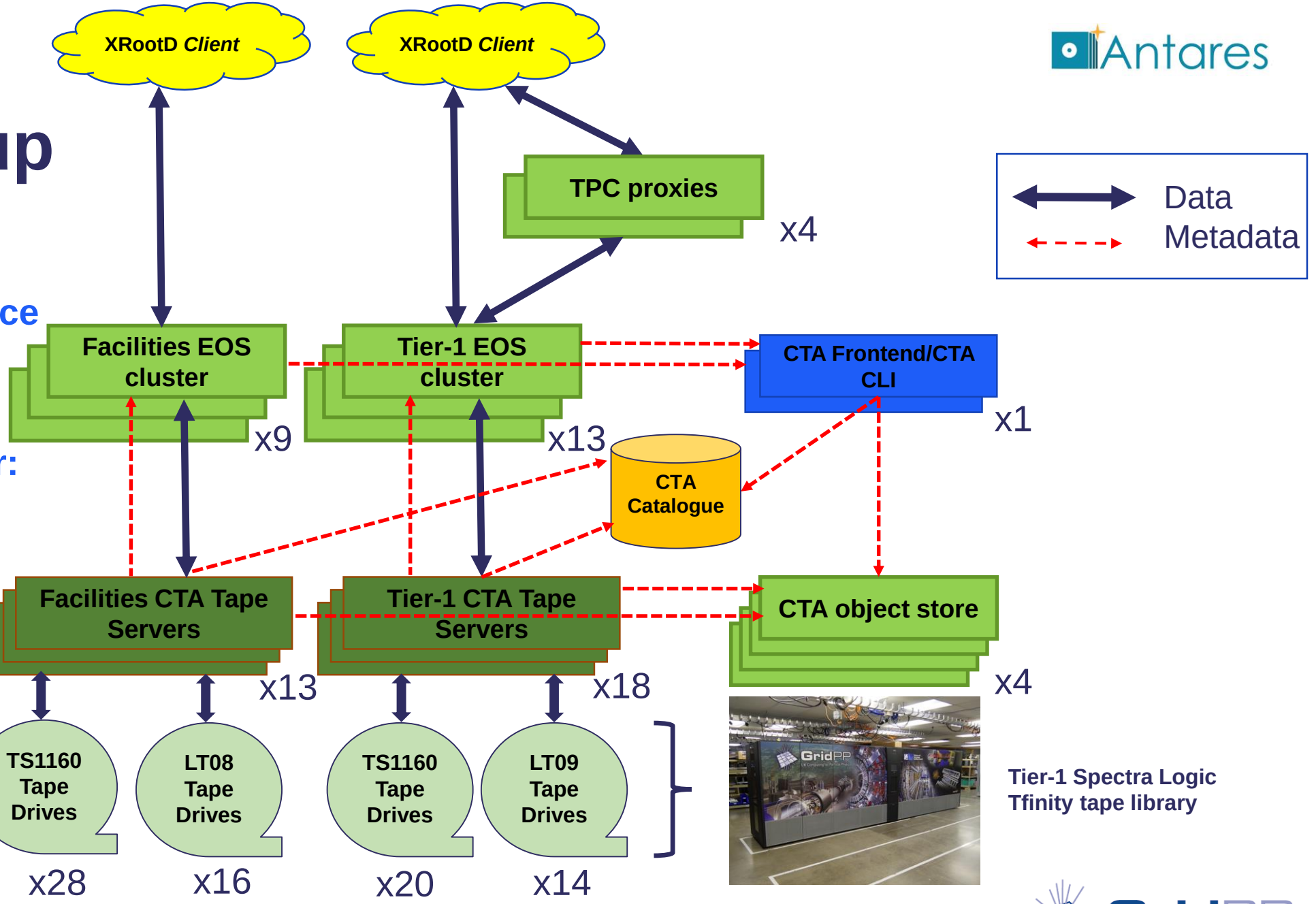


Tier-1 Spectra Logic Tfinity tape library

Antares: Future setup

Tier-1 and Facilities EOS cluster against the same CTA instance

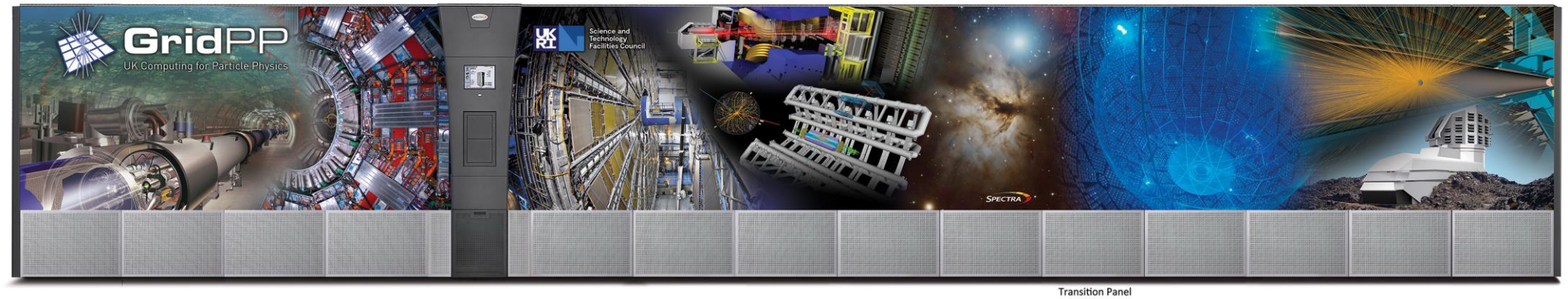
Facilities EOS cluster: 2 x 3.5TB + 7 x 1.5TB SSD nodes



Facilities SpectraTfinity tape library

Tier-1 Spectra Logic Tfinity tape library

Tier-1 tape library (Asterix)



- **Total of 15 frames - longest in Europe/UK!**
- **20 x TS1160, 16 x LTO9**
- **Capacity: 250PB**

Facilities tape library (Obelix)

- Total of 13 frames
- 28 x TS1160, 17 x LTO8, 6 x LTO9
- Capacity: 212 PB

Antares performance

- **Overall performance: very good**
  - **LHC Run 3 rates are comfortably handled**
  - **Rate limiting factor: number of mounted drives**
  - **Resolved issue with LTO9 drive performance**
- **Repeated LHCb tape challenge (Dec 2022 - Jan 2023)**
  - **Average writing speed: 2.6 GB/s (Run3 target 2.96 GB/s)**
  - **Average staging speed: 3 GB/s (Run3 target 1.93 GB/s)**
  - **Much better results compared with the previous challenge**

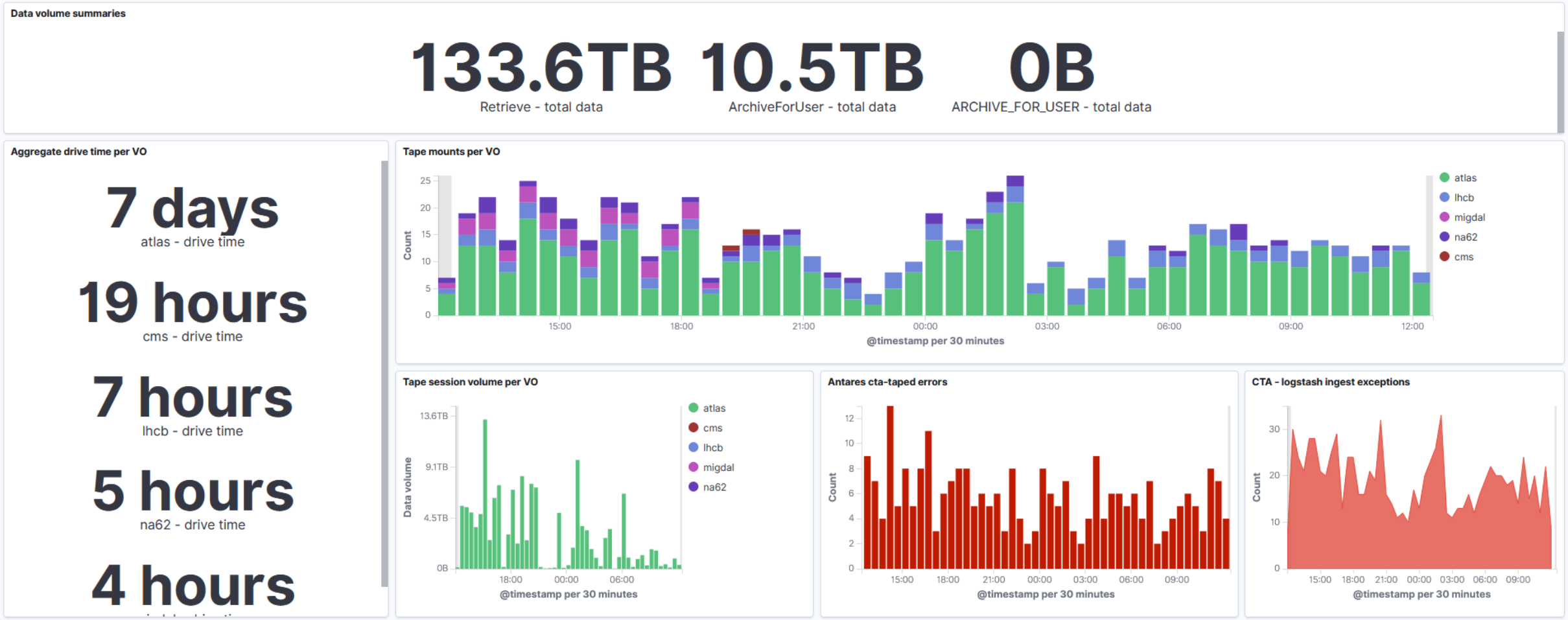
CTA event logging

- **Started working on centralised logging - beginning with the cta-taped logs**
  - **Filebeat/Logstash/OpenSearch**
- **The main technical challenge was extracting useable info from the log entries**
  - **cta-taped logs contains several fields that are used in multiple contexts**
  - **Logstash key-values plugin with custom ruby processor to cast fields to useable formats**


```
"cta_log_hostname": "cta-ts04",
"beat": {
  "hostname": "cta-ts04.scd.rl.ac.uk",
  "version": "6.8.23",
  "name": "cta-ts04.scd.rl.ac.uk"
},
"cta_kvs": {
  "mountId": "595492",
  "TID": "23544",
  "checksumingTime": 0,
  "waitFreeMemoryTime": 0.002546,
  "tapeVid": "CT1700",
  "fileId": "4312934729",
  "transferTime": 1.482578,
  "tapeDrive": "asterix_ts1160_00",
  "globalPayloadTransferSpeedMBps": 291.836633,
  "PID": "22560",
  "openingTime": 0.008003,
  "checkingErrorTime": 0.000207,
  "readWriteTime": 1.471807,
  "diskPerformanceMBps": 291.836633,
  "threadID": "9",
  "waitDataTime": 0,
  "openRWCloseToTransferTimeRatio": 0.998133,
  "MSG": "File successfully read from disk",
  "LVL": "INFO",
  "path": "root://antares-eos01.scd.rl.ac.uk//e
```


cta-taped log event in Opensearch

High level system overview from cta-taped information



Deriving accounting stats

- **Aggregate cta-taped logs contain the information required to derive extra fields required for the Tape Metrics Accounting**
  - **Aggregate values (tape mounts, read/write bytes per VO) can be calculated directly in OpenSearch, work in progress to generate the various accounting reports from this data**

```
GET tier1*/_search
{
  "size": 0,
  "query": {
    "bool": {
      "must": [
        {"match": { "type": "cta" }},
        {"match": { "cta_kvs.MSG": "Tape session finished" }},
        {"match": { "aq_cta_instance": "antares" }},
        {"range": {"@timestamp": {"gte": "now-24h", "lt": "now"}}}
      ]
    }
  },
  "aggs": {
    "vo": {
      "terms": {
        "field" : "cta_kvs.vo.keyword"
      },
      "aggs":{
        "unique_mounts": { "cardinality": { "field": "cta_kvs.tapeVid.keyword" } },
        "total_mounts": { "cardinality": { "field": "cta_kvs.mountId.keyword" } },
        "mountType": {
          "terms": {
            "field" : "cta_kvs.mountType.keyword"
          },
          "aggs": {
            "total_bytes" : { "sum": { "field": "cta_kvs.dataVolume" } }
          }
        }
      }
    }
  }
}
```

```
{
  "took" : 1999,
  "timed_out" : false,
  "_shards" : {...},
  "hits" : {...},
  "aggregations" : {
    "vo" : {
      "doc_count_error_upper_bound" : 0,
      "sum_other_doc_count" : 0,
      "buckets" : [
        {
          "key" : "atlas",
          "doc_count" : 20149,
          "unique_mounts" : {
            "value" : 127
          },
          "total_mounts" : {
            "value" : 533
          },
          "mountType" : {
            "doc_count_error_upper_bound" : 0,
            "sum_other_doc_count" : 0,
            "buckets" : [
              {
                "key" : "Retrieve",
                "doc_count" : 17984,
                "total_bytes" : {
                  "value" : 1.61649337743067E14
                }
              },
              {
                "key" : "ArchiveForUser",
                "doc_count" : 2165
```

Operational issues

- **EOS MGM becomes unresponsive**
  - **Mitigation: a cron script restarts eos@mgm if necessary**
  - **Monitor metadata operations to identify cause of load**
- **cta-taped stops after repeated tape mount failures**
  - **Because we have 2 drives/tape server, both drives are brought down**
  - **Suggested mitigation: run cta-taped in a container**
- **ALICE operational issues**
  - **Can’t get the TPC test/transfers to work**
  - ***absolute_max_age_secs* (86400s) too small for the ALICE workflows**

CASTOR Facilities migration

- The second and last CASTOR to CTA migration
- Three groups of data (“VOs”) to migrate: CEDA, Diamond and JASMIN

  ~13.3 million files (~101.5PB)
- More complex than the Tier-1 data migration
  1. Integrate Facilities client code with EOS-CTA
  2. Remove FileID clashes with some Tier-1 files (Tom’s talk at 11:30 today)
  3. Test large file ingest and recall
- Migration procedure will be the same as with Tier-1 migration
  - ✓ Sequence of steps defined
  - ✓ Two tape pools migrated to Antares dev
  - ✓ Data from the migrated pools recalled successfully

Future plans

- Tape accounting
- Monitoring
  - MGM metadata operations
  - Garbage collection in the retrieve space
- Configure Tape REST API
- Upgrade to EOS/CTA 5
- Hardware: Add 4 TPC nodes
- Network: Enable IPv6

Antares

THANK YOU!

Science and Technology Facilities Council

GridPP
UK Computing for Particle Physics