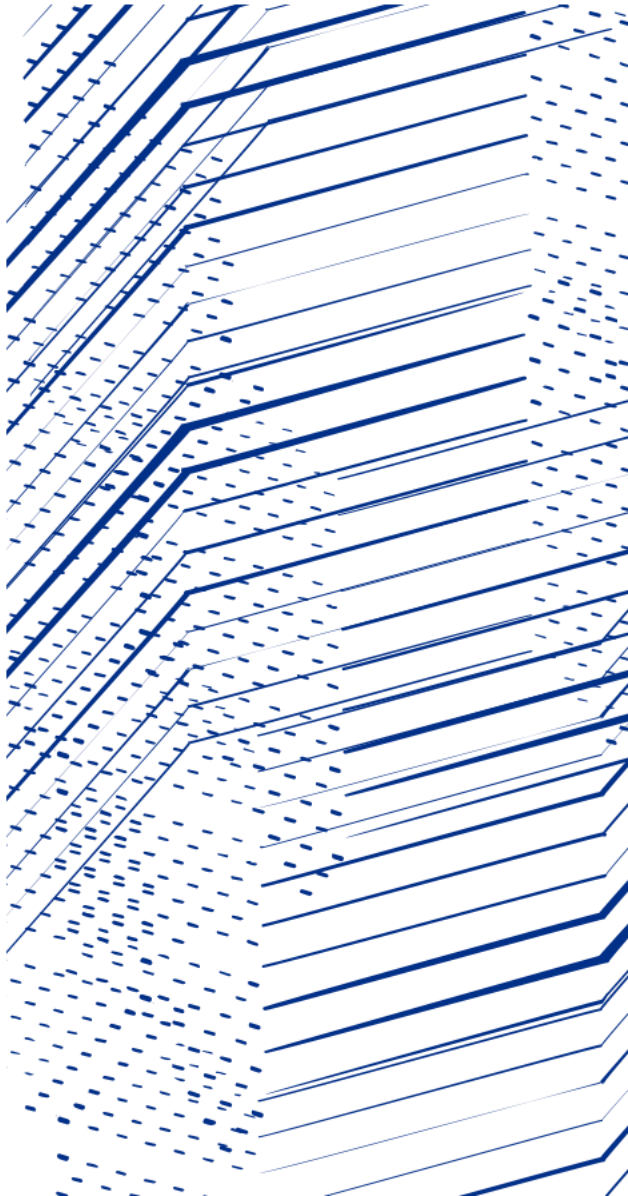
Science and Technology Facilities Council

WP1 Intelligent Data Management

Timothy Noble
GridPP53 / Swift-HEP
11 Apr 2025

Science and Technology Facilities Council



Work package 1 Vison

To optimise the use of the diverse Digital Infrastructure available to HEP experiments in the UK

- HEP experiments record and produce more than 100 PByte of data every year, and is expected to increase to more than 1 EByte/year soon.
  - Storage is evolving, and much like compute resources, we need to adapt storage to a more heterogeneous environment
  - Individual transfers need to be completed faster, and not just horizontally scale storage
- The data-lake will be based on the Rucio data management tool

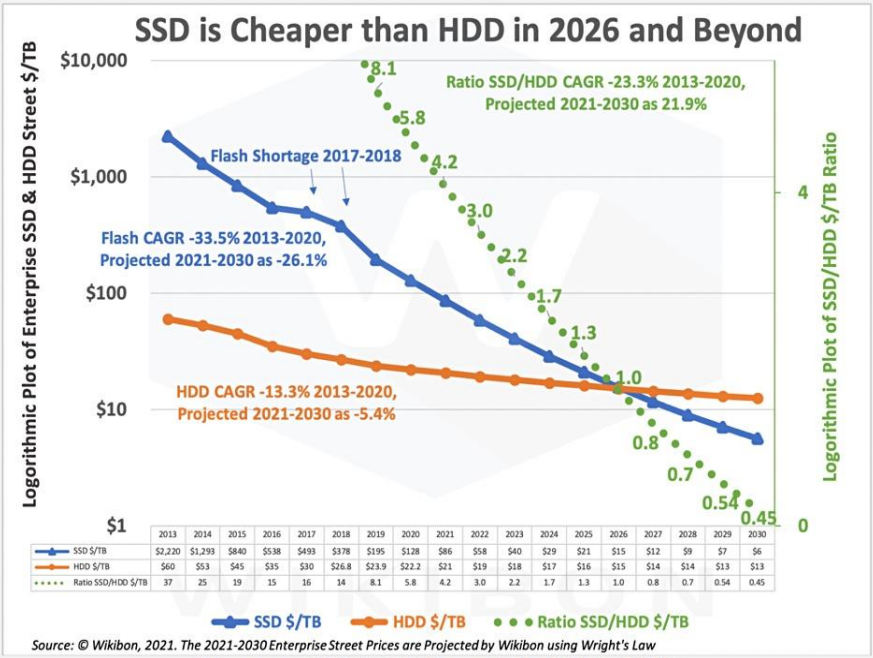


Figure 4 - SSD/HDD Pricing Ratio 2013 - 2030

Source: © Wikibon, 2021.

Science and Technology Facilities Council

Work package 1

Intelligent Data Management

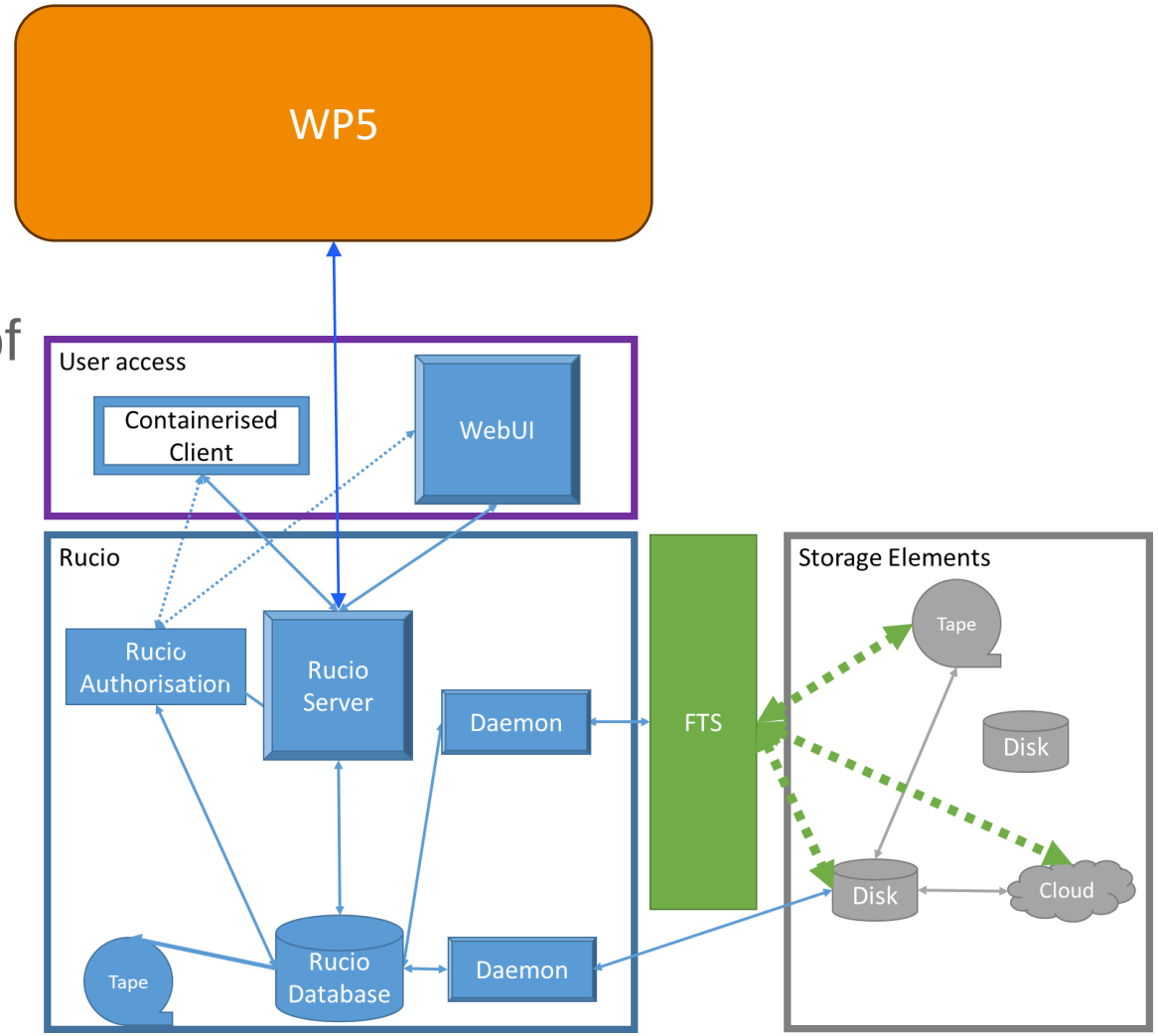
- Set up a UK data-lake prototype.
- Implement QoS information in Rucio
- Produce recommendations to sites for data access
- Integrate with the Analysis Facility

Science and Technology Facilities Council

Rucio

Distributed data management tool

- Used by several large experiments
  - Handles Petabytes of data and Exabytes of data movement
  - Treats different storage types with various behaviours
- Adds a layer of abstraction
- Distributed components work together to orchestrate data movement as required
- Tokens (Expected end Mar-2026)

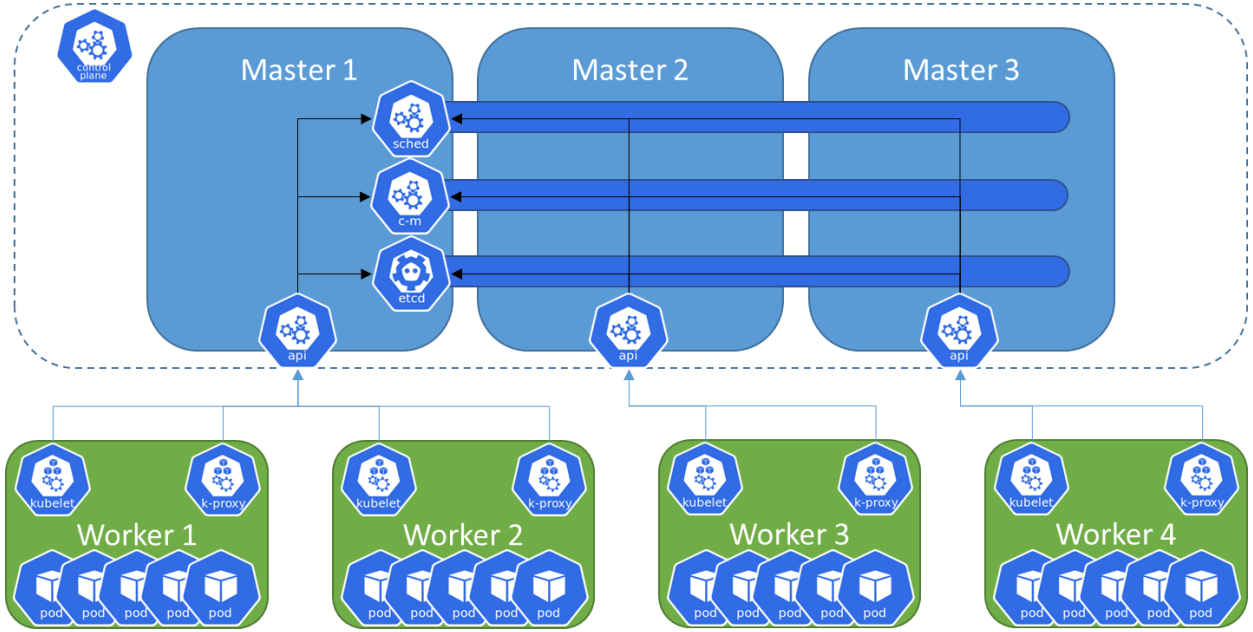


Science and Technology Facilities Council

Kubernetes

Container orchestration system for automating scaling and management

- Deployments environments described in easy-to-read and write YAML files
- Containers are self contained units of software, that allow the software to be deployed anywhere
- K8S orchestrates not just container deployment but networking, persistent storage, security, secret management



Science and Technology Facilities Council

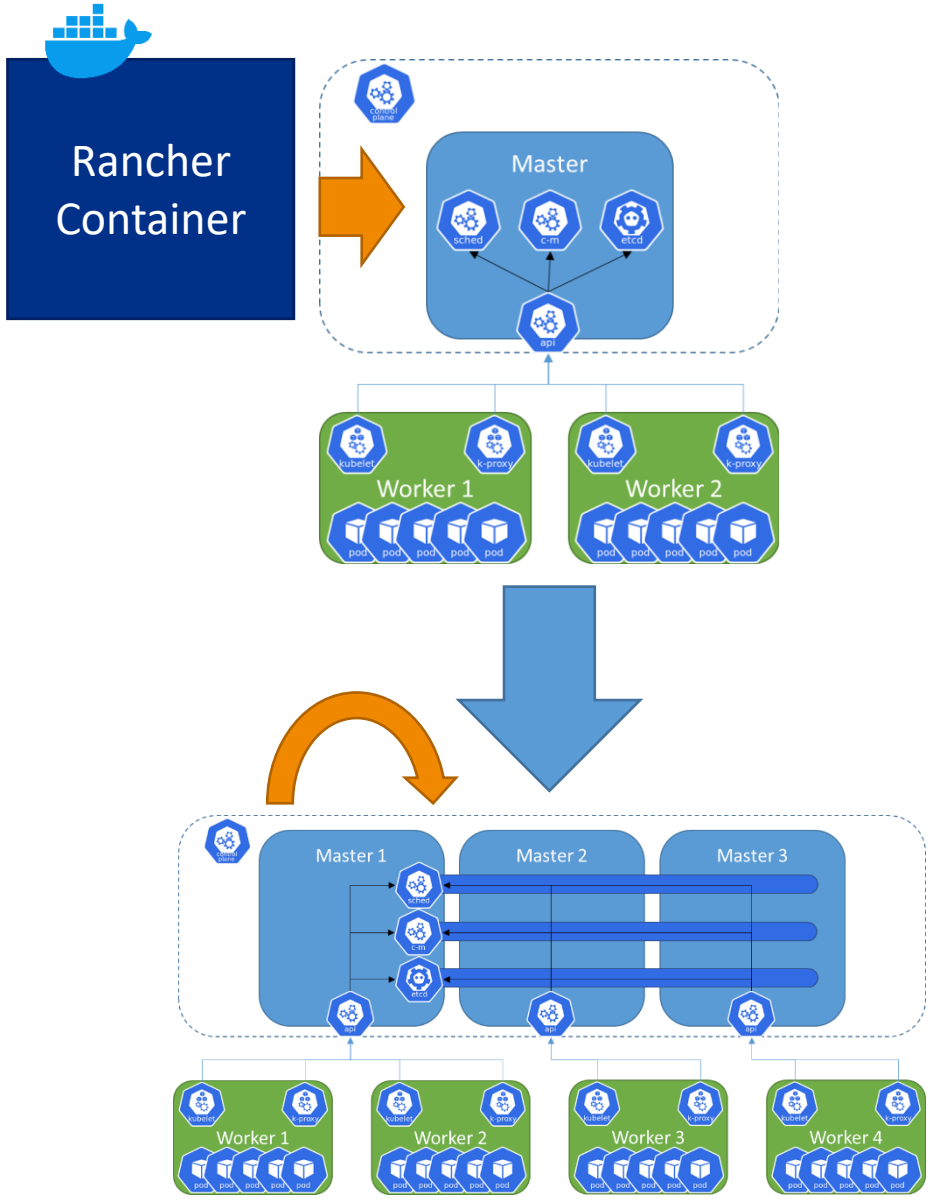
Challenges for WP1

- Deploying a Rucio instance that is flexible but robust for development and testing
- Token authentication for data management – extending to the Analysis Facility
- Maximising the SSD storage endpoints
- Prioritising the data movement depending benefits

Kubernetes work

Evolution of Kubernetes cluster from a simple deployment to something that is production-ready

- Deployment and iteration on K8S cluster deployment to make the cluster more reliable
- Development of a production-ready K8S cluster, with Highly Available Master nodes
- Move from a container used to create a cluster to a self-monitoring cluster using RKE2

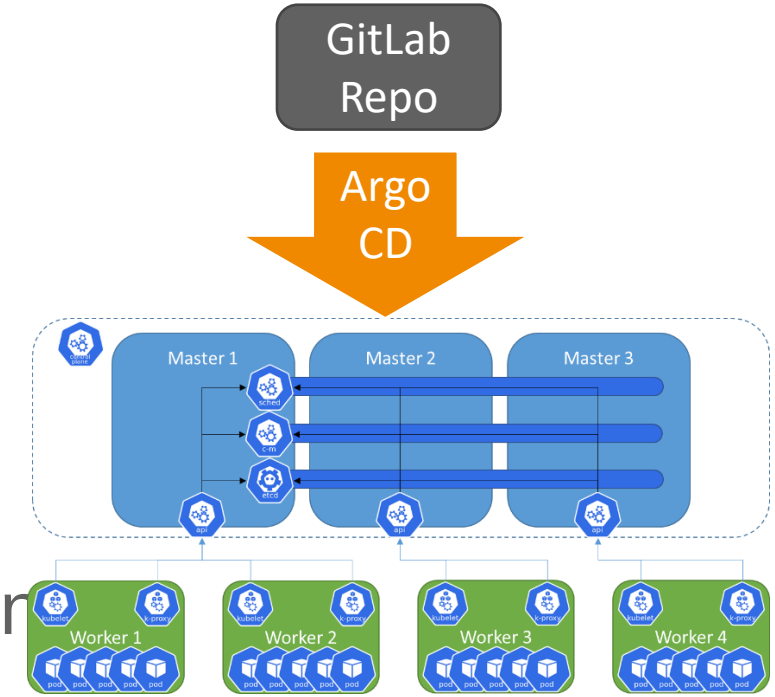


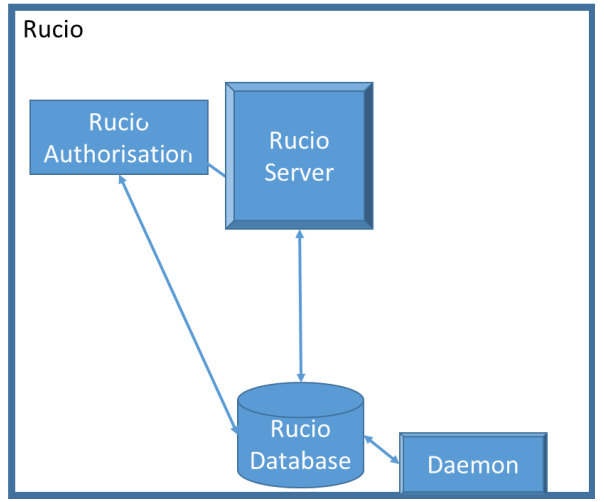
Science and Technology Facilities Council

Kubernetes Deployment

Rucio support shifted to K8S as the preferred way

- Rucio is now deployed on the K8S cluster
- Described in a GitHub Repository for a single source of truth and CI/CD integration to allow for development and deployment testing





Science and Technology Facilities Council

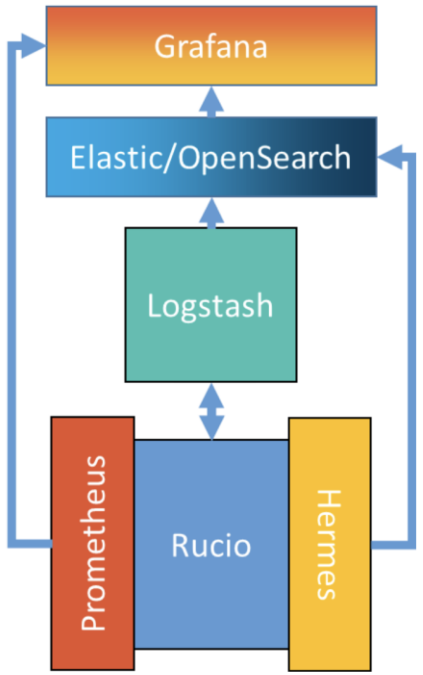
Monitoring of Functional Tests

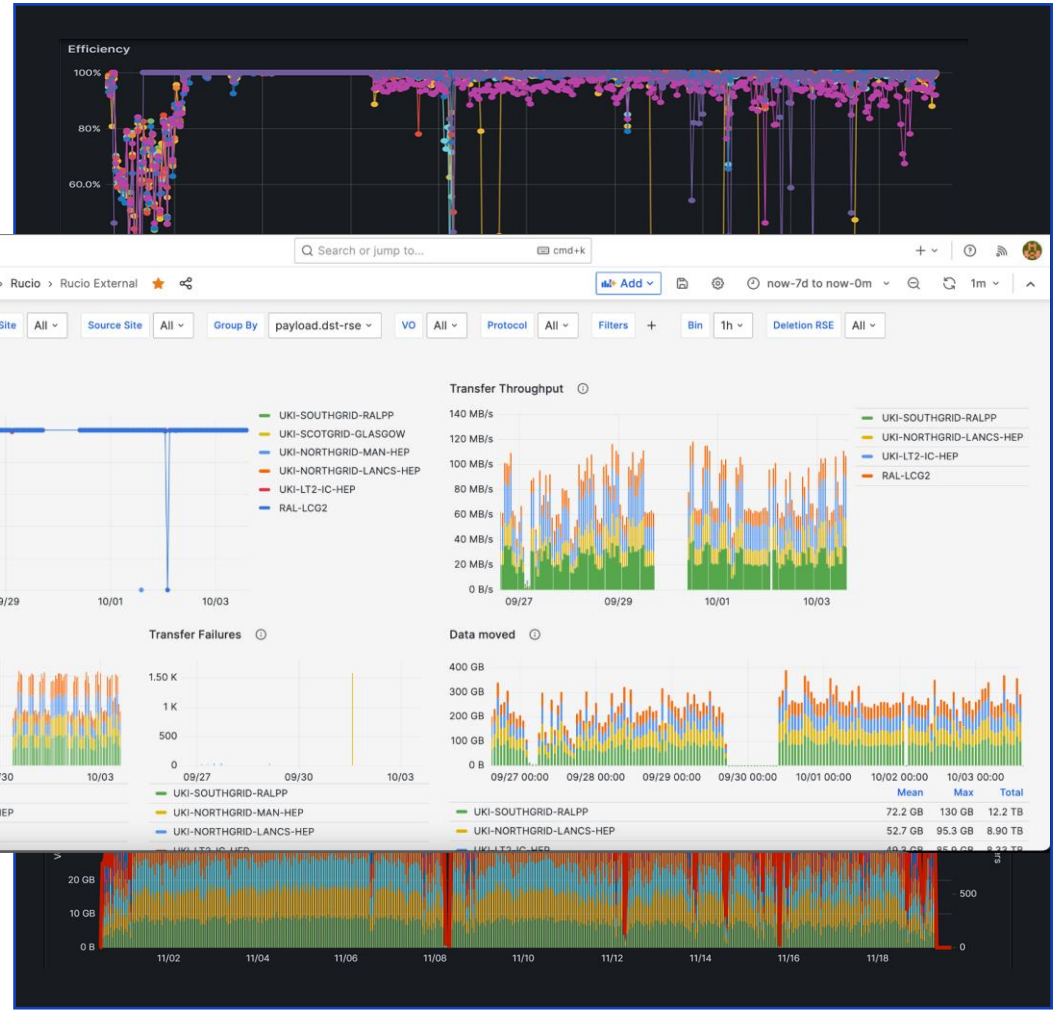
Monitoring deployment written and included as a minimum viable monitoring stack in the Rucio repository

As few components as possible to get metrics

Modelled on the FTS monitoring for the base dashboard

Deployed for LSST





Science and Technology Facilities Council

https://indico.cern.ch/event/1343110/contributions/6096503/

Storage pools at RAL

ECHO: Ceph-based (RADOS) object store with data access provided through XRootD:

- XrdCeph OSS plugin – originally developed by S. Ponce (CERN) using ceph's libradostriper
- Provides the interface between XRootD and ceph at the OSS layer
- Also deployed for UK Tier-2 site: Glasgow for ATLAS
- 98PB raw storage
- 8+3 Erasure Coding
- ~ 240 Storage Nodes (SN), with ~ 5000 OSDs
- Host level failure domain (i.e. OSDs from placement group placed across different SNs).
- 2 service nodes (e.g. XRootD Gateway, Ceph Mon) + several storage nodes per rack, with ToR routers.
- RAL also provides CephFS, S3 and SWIFT endpoints, etc.

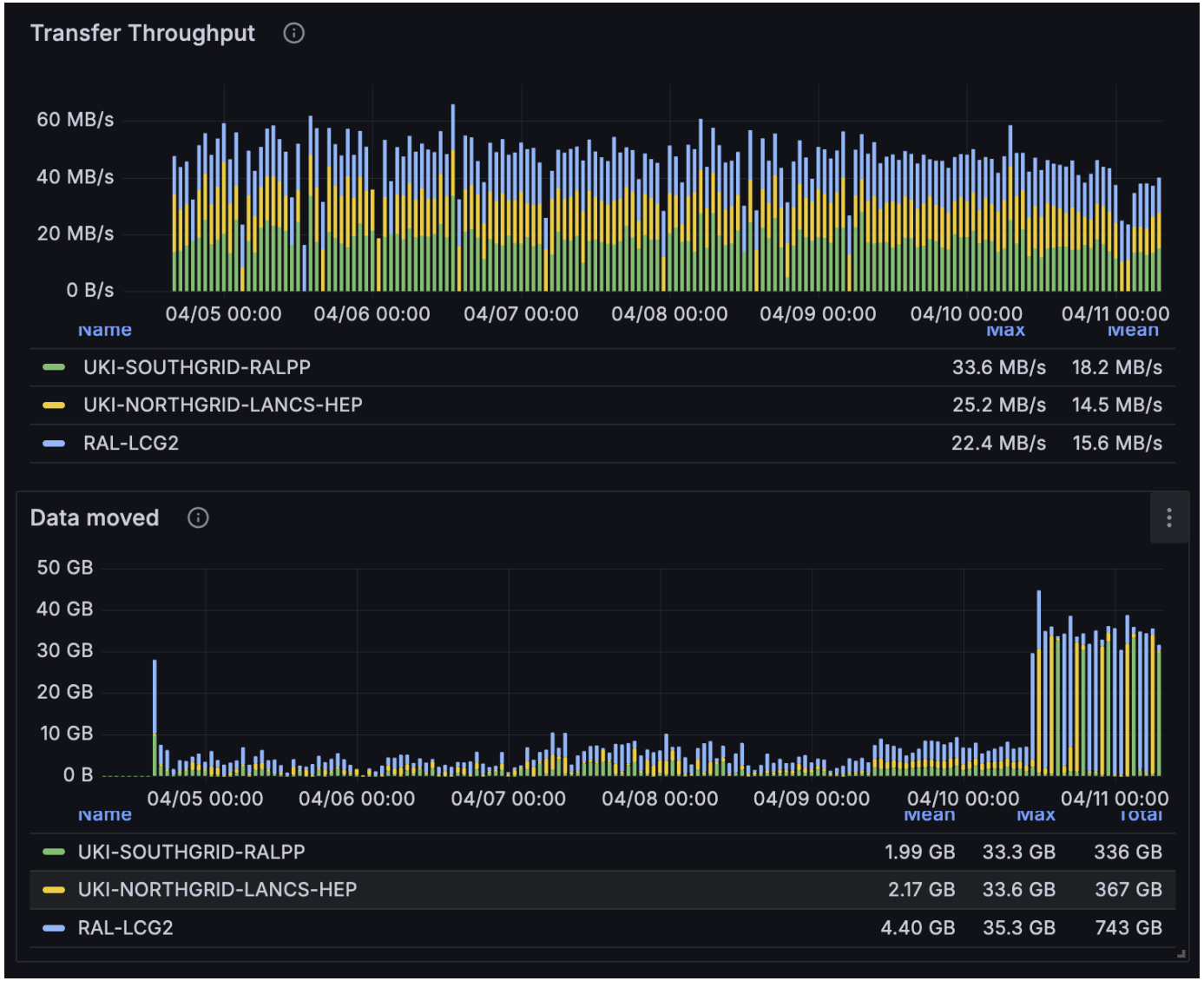
- SSDs
  - 138 SSDs
  - 480 TB
- HDDs
  - 6162 HDD
  - 98PB

Science and Technology Facilities Council

More information on Echo
https://indico.jlab.org/event/459/contributions/11324/

Testing the SSD

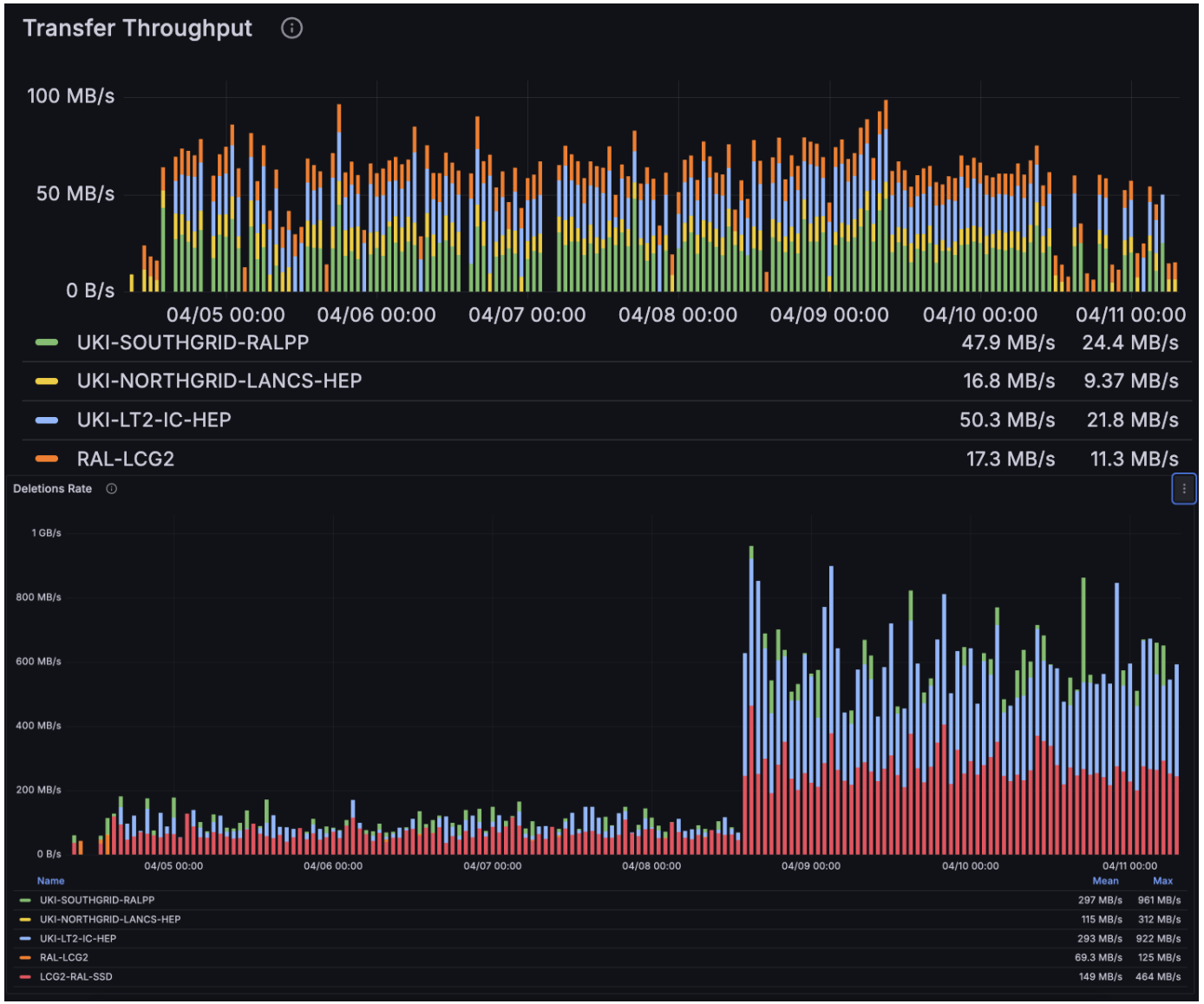
- To SSD pool
  - 81.2Mb/s Max
  - 47.7Mb/s Mean
- To Echo
  - 22.4 Mb/s Max
  - 15.2 Mb/s Mean
- To Lancs
  - 40.1 Mb/s Max
  - 27.07 Mb/s Mean



Science and Technology Facilities Council

Testing the SSD

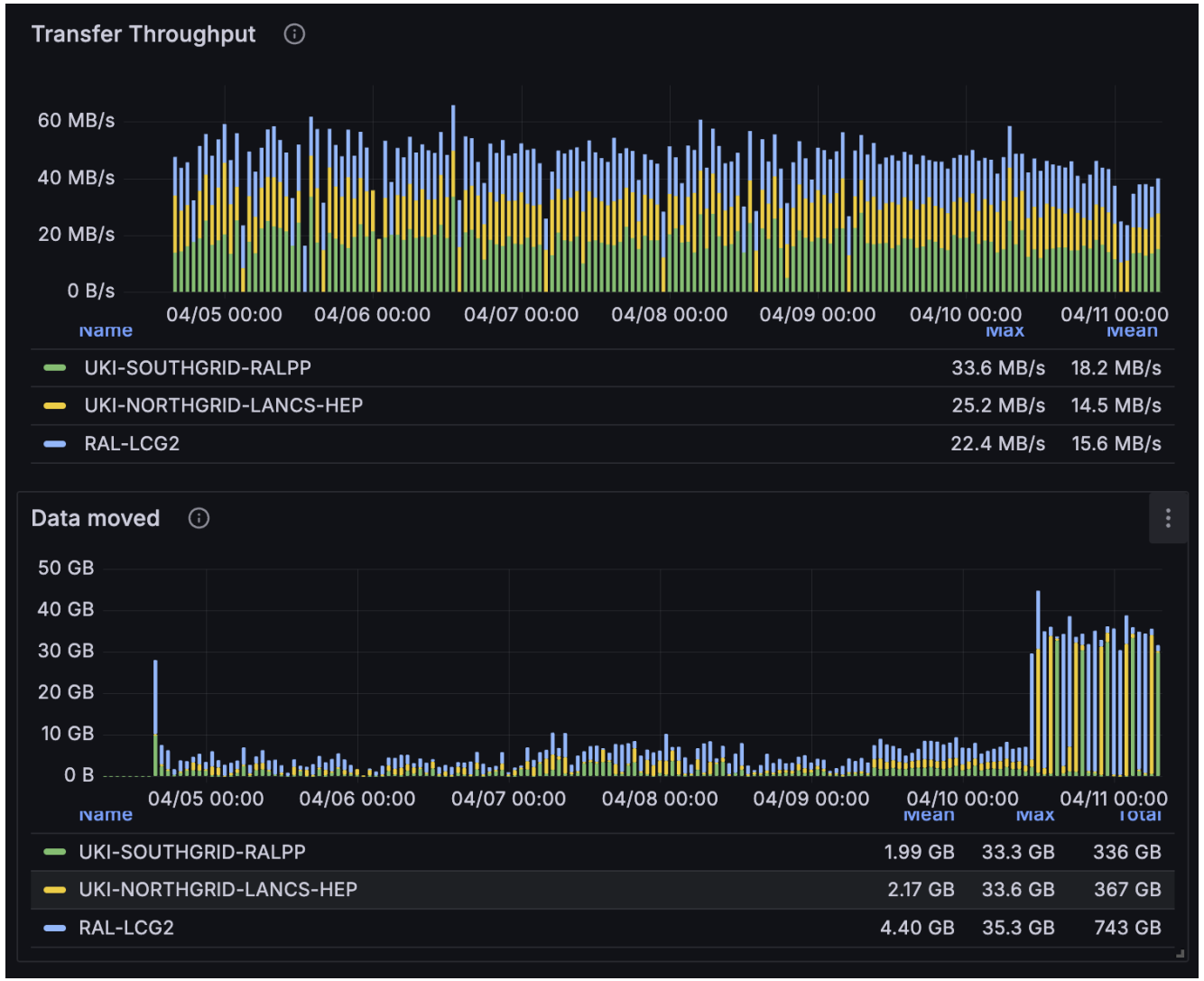
- From SSD pool
  - Mean of 67 Mb/s out
  - 62mb/s from RAL
  - 62Mb/s from Lancs
- Deletions
  - 274Mb/s Mean SSD
  - 590Mb/s from RALPP
  - 213Mb/s from Lancs
  - 97.5Mb/s from Echo
- What does this mean?



Science and Technology Facilities Council

Testing the SSD

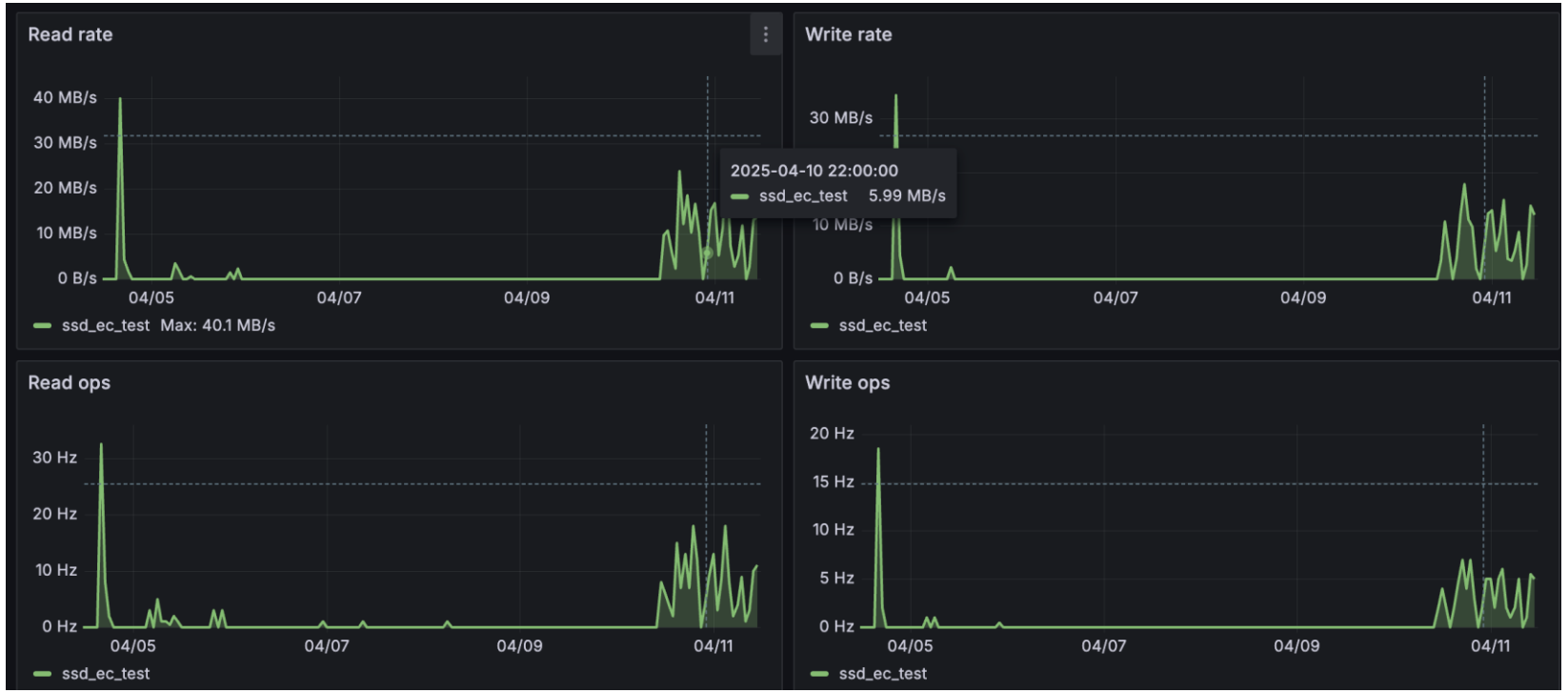
- Increased data throughput did not increase the data rates
- Should be two data sizes, of 32Mb and 2Gb files



Science and Technology Facilities Council

Testing the SSD

- These metrics from Echo directly match the Rucio / FTS monitoring



Science and Technology Facilities Council

Interesting twist

- A 32M file was generated despite configuration changes
- dd is used to create the test files in the Rucio
- dd buffer, preventing larger files from being created by the Rucio Automatix daemon (I discovered this yesterday morning).

```
{
        "LargeFew": {
                "probability": 0.5,
                "nbfiles": 6,
                "filesize": 5000000000,
                "metadata":
                {
                        "project": "SSD Test",
                        "prod_step": "ATLAS"
                }
        },
        "SmallMany": {
                "probability": 0.5,
                "nbfiles": 1000,
                "filesize": 33000000,
                "metadata": {
                        "project": "SSD Test",
                        "prod_step": "LSST"
                }
        }
}
```

```
[user@f69bcec8e865 ~]$ dd if=/dev/urandom of=test bs=3G count=1
0+1 records in
0+1 records out
2147479552 bytes (2.1 GB, 2.0 GiB) copied, 24.2586 s, 88.5 MB/s
```

```
[root@daemons-automatix-d867449c5-n2wlw tmp]#  dd if=/dev/urandom of=test bs=3G count=1
0+1 records in
0+1 records out
33554431 bytes (34 MB, 32 MiB) copied, 0.814298 s, 41.2 MB/s
```

Science and Technology Facilities Council

Testing the SSD

- Small files are not great for this kind of test
- Connections being maxed out in FTS and transfer speeds not being reached

Science and Technology Facilities Council

Plans for future testing

- Create a patch for Automatix to create larger files as needed (and cleaning up after itself)
  - Actually having large files
- S3 endpoint testing for comparison (webdav vs s3 protocols)
- Looking at I/O from jobs that may require higher access speeds
- Create a report/recommendation on when SSD storage is useful
- Look at any development needed in Rucio to support the SSD usage better (or maybe its enough already)

Work package 1

Intelligent Data Management

- Set up a UK data-lake prototype. - ☑
- Implement QoS information in Rucio - ☑
- Produce recommendations to sites - In progress
- Integrate with the Analysis Facility - In progress

Science and Technology Facilities Council