UKRI
Science and
Technology
Facilities Council

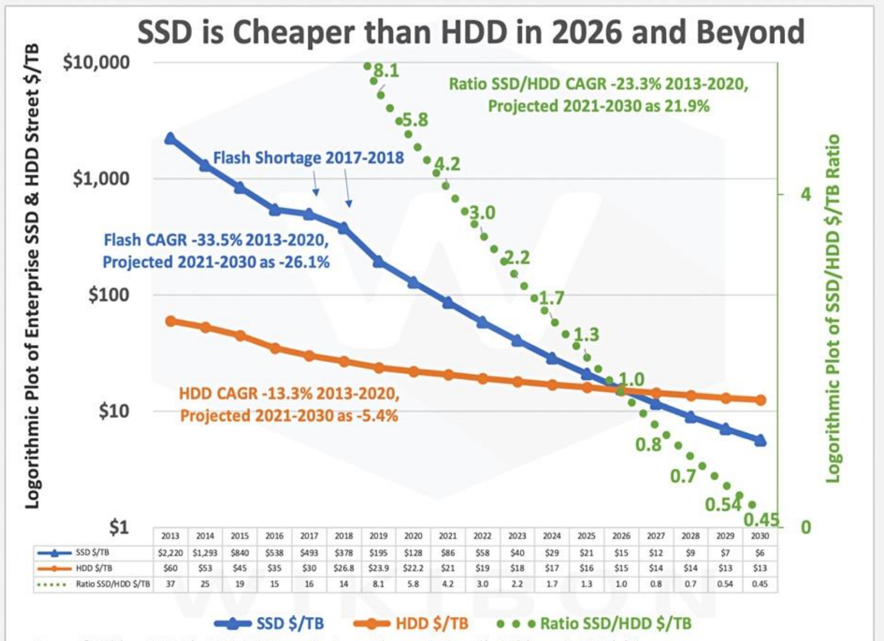
WP1.1 Intelligent Data Management

Timothy Noble
22 Nov 2023

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 in the next ~~five~~ three years
  - 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



	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
SSD $/TB	$2,220	$1,293	$840	$538	$493	$378	$195	$128	$86	$58	$40	$29	$21	$15	$12	$9	$7	$6
HDD $/TB	$60	$53	$45	$35	$30	$26.8	$23.9	$22.2	$21	$19	$18	$17	$16	$15	$14	$14	$13	$13
Ratio SSD/HDD $/TB	37	25	19	15	16	14	8.1	5.8	4.2	3.0	2.2	1.7	1.3	1.0	0.8	0.7	0.54	0.45

Source: © Wikibon, 2021. The 2021-2030 Enterprise Street Prices are Projected by Wikibon using Wright's Law

Figure 4 - SSD/HDD Pricing Ratio 2013 - 2030

Source: © Wikibon, 2021.

Science and Technology Facilities Council

Data Challenge 2024 - DC24

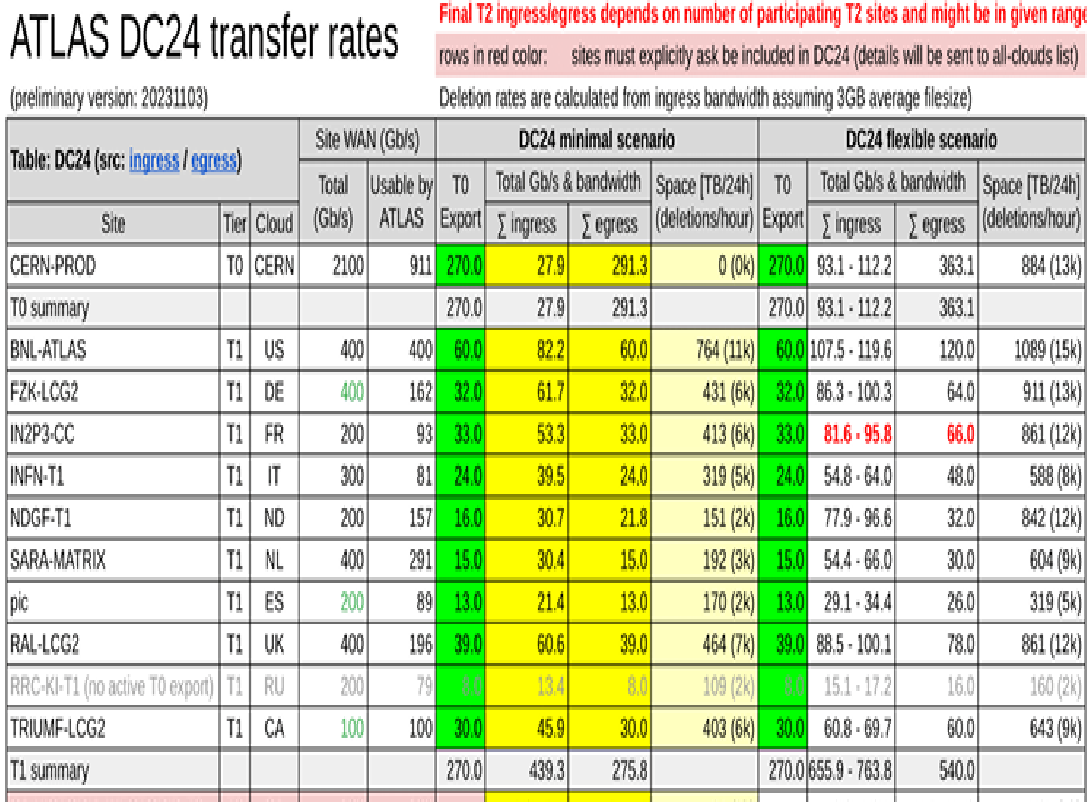
A challenge to participating sites to ensure sites are on track to being able to accept HL-LHC data rates.

ATLAS and CMS RAW data from CERN to T1 to be ~800 Gbps per experiment, with an additional 200 Gbps for other data types

ATLAS DC24 data rates

Very high data rates and this is only a fraction of the HL-LHC data rate.

Investigating QoS, especially for SSD endpoints could be critical to ensure data rates into sites, as well as I/O for jobs



ATLAS DC24 transfer rates

(preliminary version: 20231103)

Final T2 ingress/egress depends on number of participating T2 sites and might be in given range

rows in red color: sites must explicitly ask be included in DC24 (details will be sent to all-clouds list)

Deletion rates are calculated from ingress bandwidth assuming 3GB average filesize)

Table: DC24 (src: ingress / egress)			Site WAN (Gb/s)		DC24 minimal scenario				DC24 flexible scenario			
			Total (Gb/s)	Usable by ATLAS	T0 Export	Total Gb/s & bandwidth		Space [TB/24h] (deletions/hour)	T0 Export	Total Gb/s & bandwidth		Space [TB/24h] (deletions/hour)
Site	Tier	Cloud				∑ ingress	∑ egress			∑ ingress	∑ egress	
CERN-PROD	T0	CERN	2100	911	270.0	27.9	291.3	0 (0k)	270.0	93.1 - 112.2	363.1	884 (13k)
T0 summary					270.0	27.9	291.3		270.0	93.1 - 112.2	363.1	
BNL-ATLAS	T1	US	400	400	60.0	82.2	60.0	764 (11k)	60.0	107.5 - 119.6	120.0	1089 (15k)
FZK-LCG2	T1	DE	400	162	32.0	61.7	32.0	431 (6k)	32.0	86.3 - 100.3	64.0	911 (13k)
IN2P3-CC	T1	FR	200	93	33.0	53.3	33.0	413 (6k)	33.0	81.6 - 95.8	66.0	861 (12k)
INFN-T1	T1	IT	300	81	24.0	39.5	24.0	319 (5k)	24.0	54.8 - 64.0	48.0	588 (8k)
NDGF-T1	T1	ND	200	157	16.0	30.7	21.8	151 (2k)	16.0	77.9 - 96.6	32.0	842 (12k)
SARA-MATRIX	T1	NL	400	291	15.0	30.4	15.0	192 (3k)	15.0	54.4 - 66.0	30.0	604 (9k)
pic	T1	ES	200	89	13.0	21.4	13.0	170 (2k)	13.0	29.1 - 34.4	26.0	319 (5k)
RAL-LCG2	T1	UK	400	196	39.0	60.6	39.0	464 (7k)	39.0	88.5 - 100.1	78.0	861 (12k)
RRC-KI-T1 (no active T0 export)	T1	RU	200	79	8.0	13.4	8.0	109 (2k)	8.0	15.1 - 17.2	16.0	160 (2k)
TRIUMF-LCG2	T1	CA	100	100	30.0	45.9	30.0	403 (6k)	30.0	60.8 - 69.7	60.0	643 (9k)
T1 summary					270.0	439.3	275.8		270.0	655.9 - 763.8	540.0	

Work package 1.1

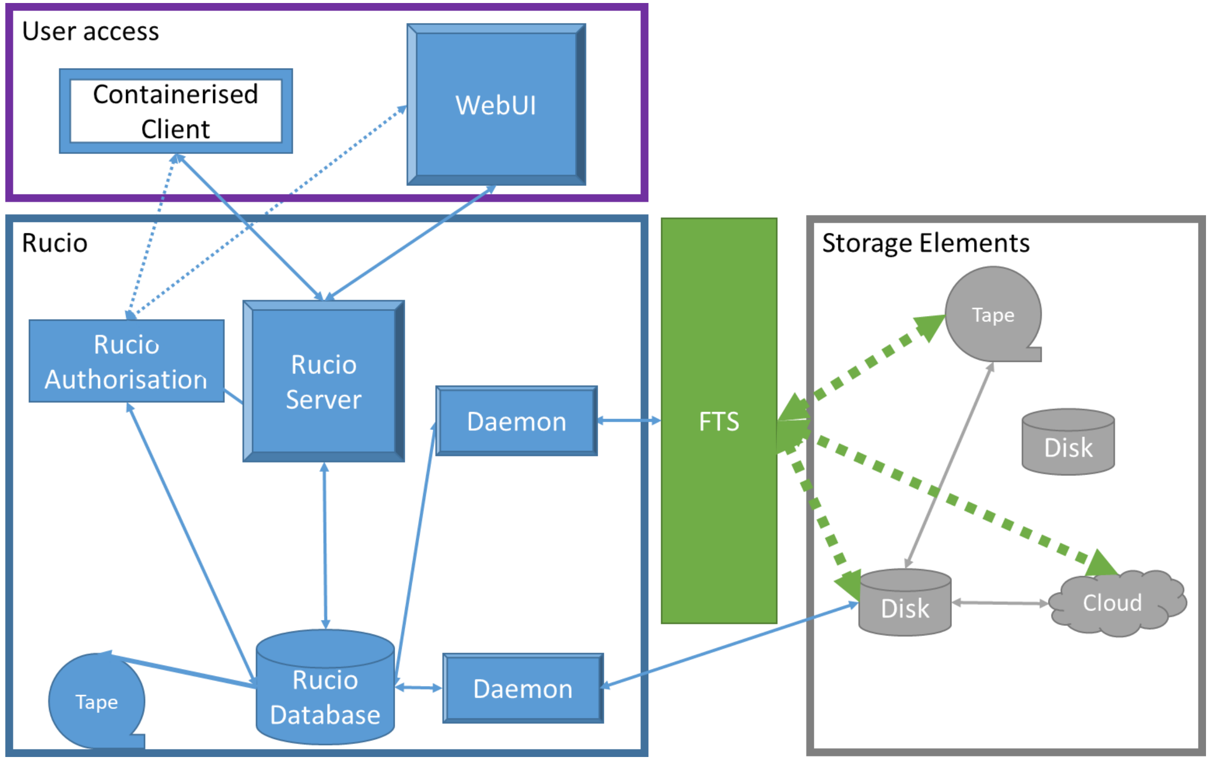
Intelligent Data Management

- Set up a UK data-lake prototype.
  - This will build on the DOMA prototype, with the intention that there will be one data lake per region/country.
- Setting up the Data-lake in the first instance consists of 3 steps
  - Configure core sites - current 10 RSEs configured in the UK
  - Configure additional – different storage sites
  - Generate metrics for comparison
- Implement QoS information in Rucio
  - Reliability of storage
  - High performance storage

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 between user and storage endpoints
- Distributed components work together to orchestrate data movement as required

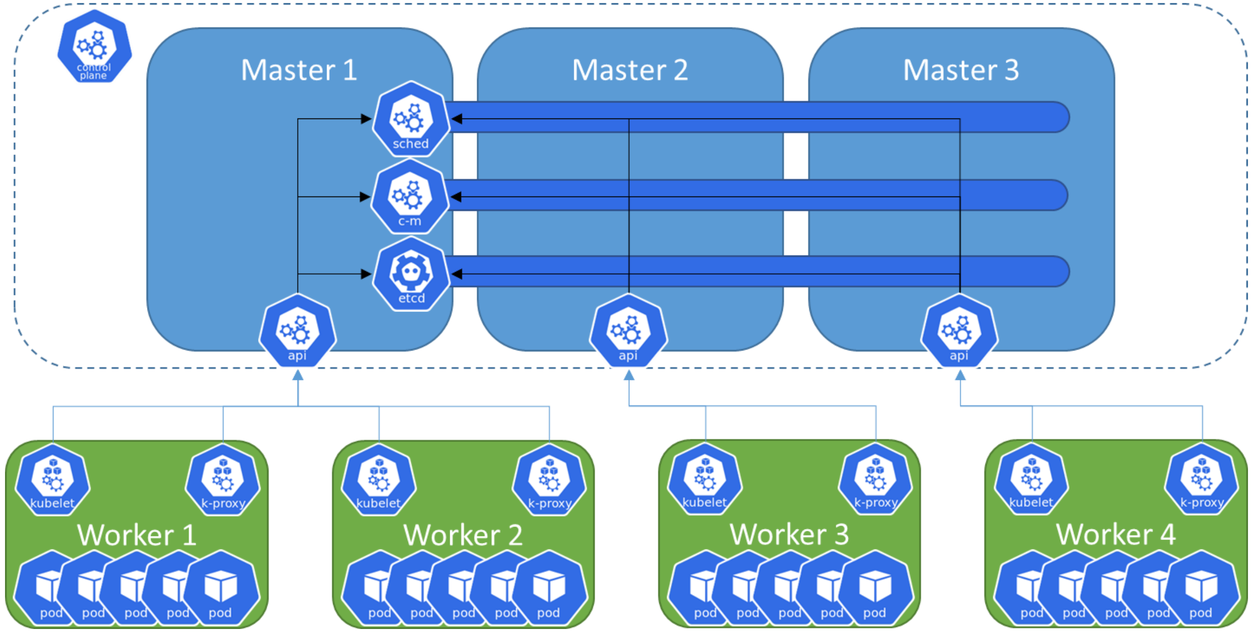


Science and Technology Facilities Council

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

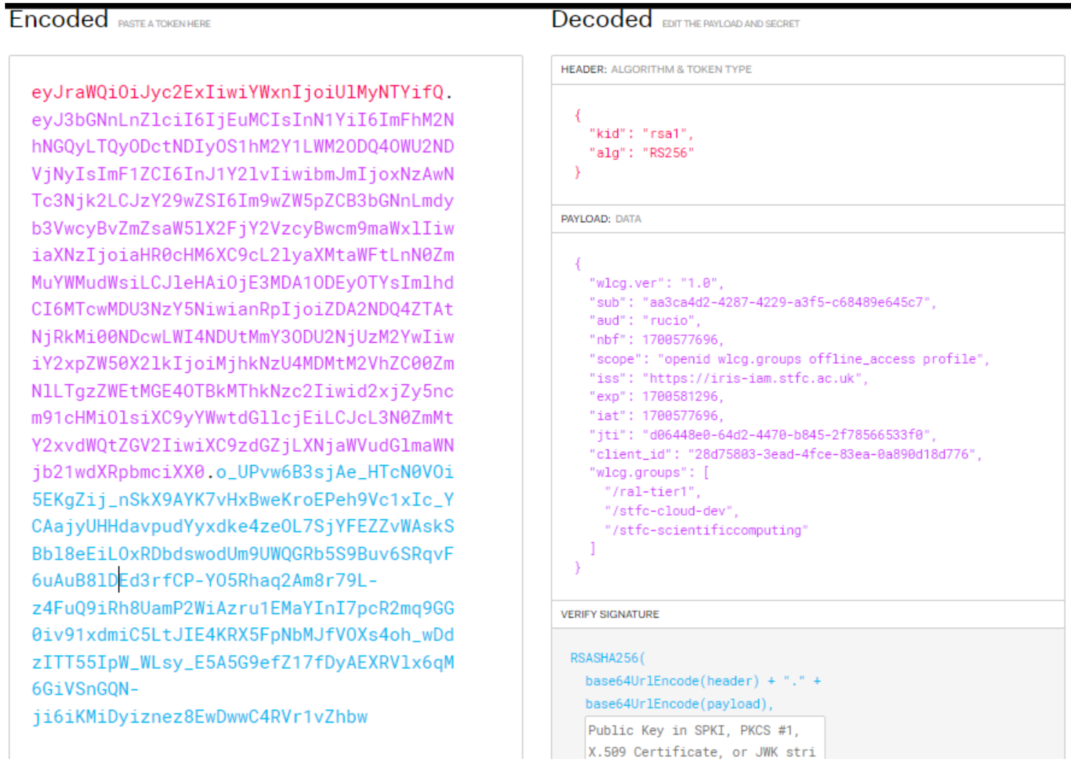


Science and Technology Facilities Council

Tokens

WLCG is aiming for a transparent replacement of user X509 certificates + VOMS with tokens

- CMS looking to use tokens for DC24 tests
- With X509 becoming the optional / non-favoured authentication method for users in March 2024
- Token authentication uses your institutional information to act as a assurance, then depending on the groups you are part of, give you authentication to perform specific actions



Encoded PASTE A TOKEN HERE

```
eyJraWQiOiJyc2ExIiwiYWxnIjoiUlMyNTYifQ.
eyJ3bGNnLnZlciI6IjEuMCIsInN1YiI6ImFhM2N
hNGQyLTQyODctNDIyOS1hM2Y1LWM2ODQ4OWU2ND
VjNyIsImF1ZCI6InJ1Y2lvIiwibmJmIjoxNzAwN
Tc3Njk2LCJzY29wZSI6Im9wZW5pZCB3bGNnLmdy
b3VwcyBvZmZsaW5lX2FjY2VzcyBwcm9maWxlIiw
iaXNzIjoiaHR0cHM6XC9cL2lyaXMtaWFtLnN0Zm
MuYWMudWsiLCJleHAiOjE3MDA1ODEyOTYsImlhd
CI6MTcwMDU3NzY5NiwianRpIjoiZDA2NDQ4ZTAt
NjRkMi00NDcwLWI4NDUtMmY3ODU2NjUzM2YwIiw
iY2xpZW50X2lkIjoiMjhkNzU4MDMtM2VhZC00Zm
NlLTgzZWEtMGE4OTBkMThkNzc2Iiwid2xjZy5nc
m91cHMiOlsiXC9yYWwtdGllcjEiLCJcL3N0ZmMt
Y2xvdWQtZGV2IiwiXC9zdGZjLXNjaWVudGlmaWN
jb21wdXRpbmciXX0.o_UPvw6B3sjAe_HTcN0VOi
5EKgZij_nSkX9AYK7vHxBweKroEPeh9Vc1xIc_Y
CAajyUHHdavpudYyxdke4zeOL7SjYFEZZvWAskS
Bbl8eEiLOxRDbdswodUm9UWQGRb5S9Buv6SRqvF
6uAuB8lDEd3rfCP-YO5Rhaq2Am8r79L-
z4FuQ9iRh8UamP2WiAzru1EMaYInI7pcR2mq9GG
0iv91xdmiC5LtJIE4KRX5FpNbMJfVOXs4oh_wDd
zITT55IpW_WLsy_E5A5G9efZ17fDyAEXRVlx6qM
6GiVSnGQN-
ji6iKMiDyiznez8EwDwwC4RVr1vZhbw
```

Decoded EDIT THE PAYLOAD AND SECRET

HEADER: ALGORITHM & TOKEN TYPE

```
{
  "kid": "rsa1",
  "alg": "RS256"
}
```

PAYLOAD: DATA

```
{
  "wlcg.ver": "1.0",
  "sub": "aa3ca4d2-4287-4229-a3f5-c68489e645c7",
  "aud": "rucio",
  "nbf": 1700577696,
  "scope": "openid wlcg.groups offline_access profile",
  "iss": "https://iris-iam.stfc.ac.uk",
  "exp": 1700581296,
  "iat": 1700577696,
  "jti": "d06448e0-64d2-4470-b845-2f78566533f0",
  "client_id": "28d75803-3ead-4fce-83ea-0a890d18d776",
  "wlcg.groups": [
    "/ral-tier1",
    "/stfc-cloud-dev",
    "/stfc-scientificcomputing"
  ]
}
```

VERIFY SIGNATURE

```
RSASHA256(
  base64UrlEncode(header) + "." +
  base64UrlEncode(payload),
  Public Key in SPKI, PKCS #1,
  X.509 Certificate, or JWK stri
```

Science and Technology Facilities Council

WLCG Token Transition Timeline working doc

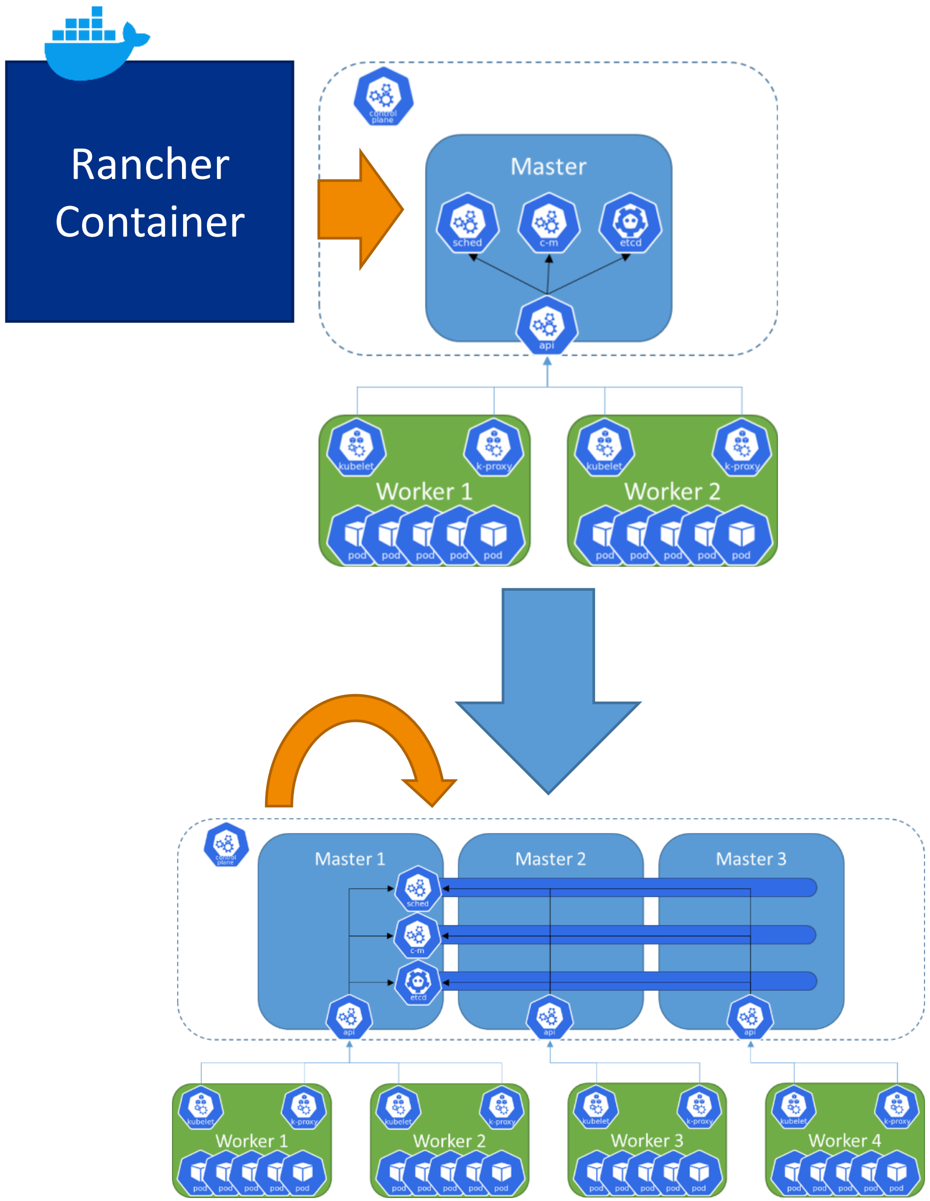
Challenges for WP1.1

- 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 on job priority

Kubernetes work

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

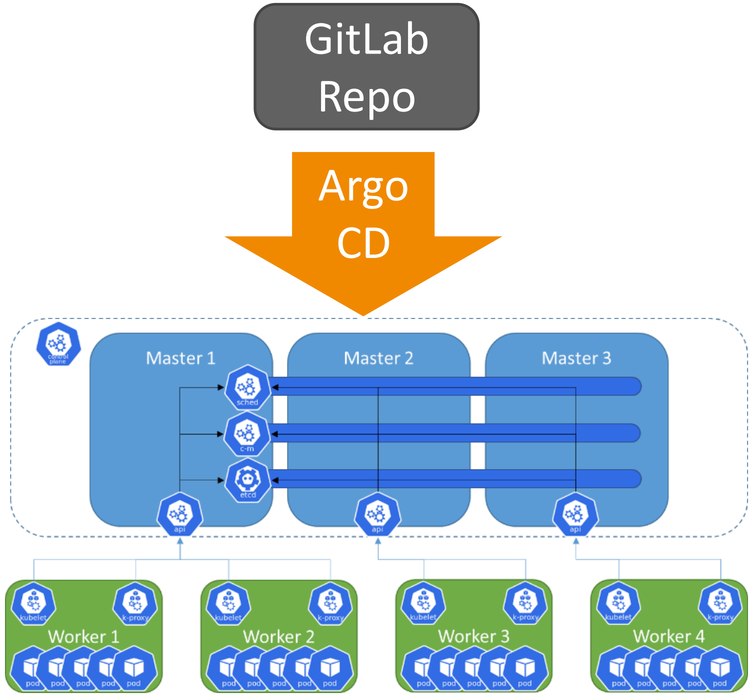
- Deployment and iteration on K8S cluster deployment to make the cluster more reliable
- Development of a production-ready K8S cluster, with Highly Available Master nodes
- Move from a container used to create a cluster to a self-monitoring cluster using RKE2
- Next step utilising the RAL Cloud K8S training to improve to a Cluster to integrate with OpenStack, and spin up and down Prod and Dev clusters

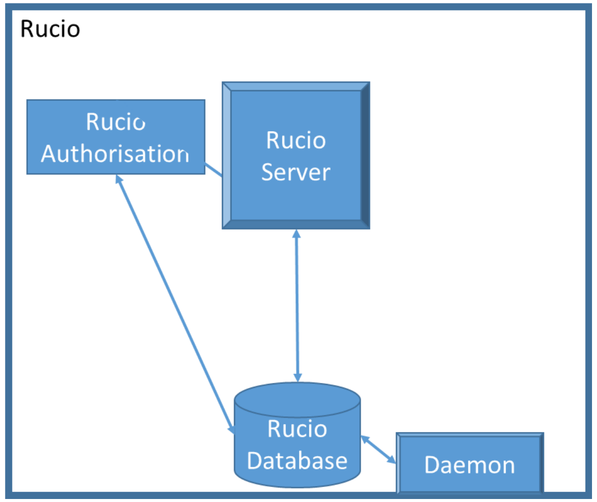


Kubernetes Deployment

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
- Allowed for jump in versions to 1.29LTS bringing many features to Multi-VO Rucio
- Working on upgrade to 32LTS (versioning format change, not a huge jump, the next LTS version)



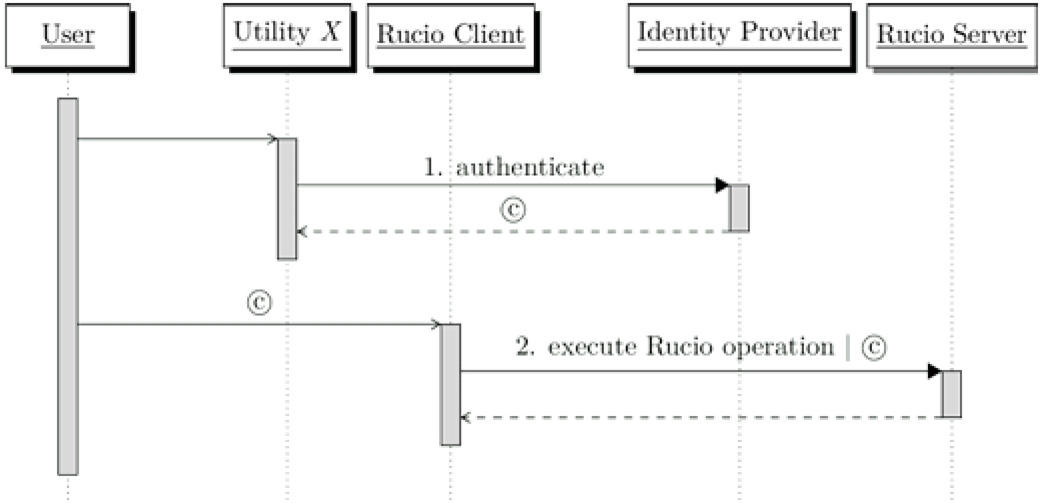


Science and Technology Facilities Council

Token Authentication into Rucio

Tokens are coming, are we ready?

- Token transition March 2024
- Authentication with tokens to Rucio
- Integrated into Rucio at RAL
  - To be used for Functional tests as sites become token enabled
- Token Concerns:
  - specificity vs. rate of request
  - Length of token life for FTS jobs
  - Complexity of flows



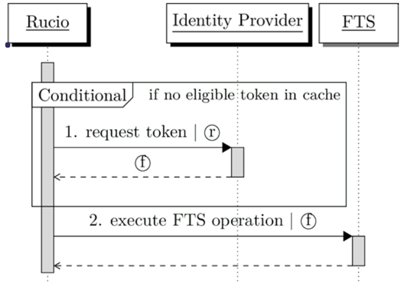
Science and Technology Facilities Council

Token Authentication into Rucio

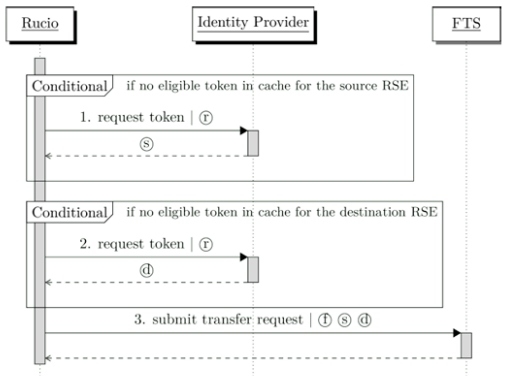
Rucio performing Third Party Copy using Tokens becomes quite a complicated workflow

1. Ensure authentication to use FTS
2. Obtain tokens to the source and destination storage endpoints
3. FTS maintaining tokens via refresh tokens
4. FTS running the actual Transfer

1.

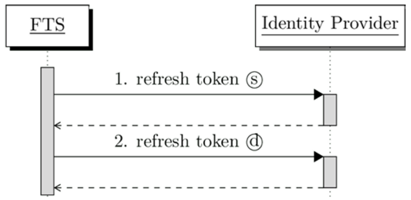


2.

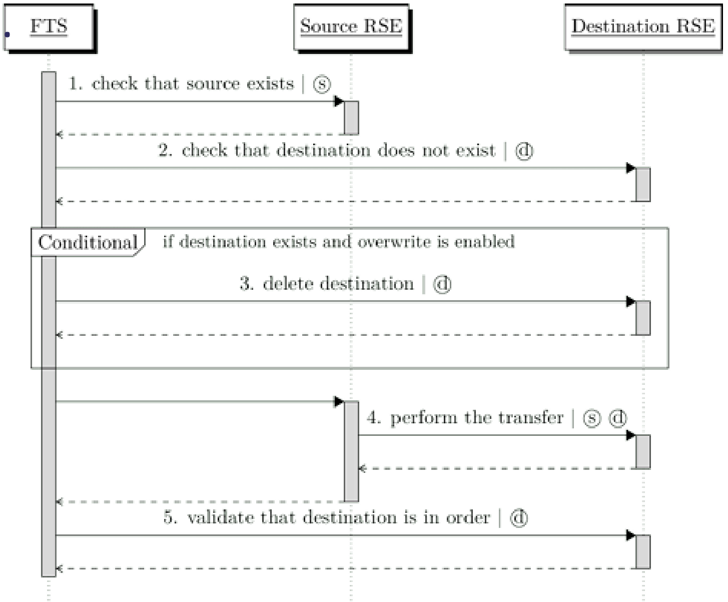


Rucio
Identity Provider
FTS
Conditional
if no eligible token in cache for the source RSE
1. request token | ⓡ
ⓢ
Conditional
if no eligible token in cache for the destination RSE
2. request token | ⓡ
ⓓ
3. submit transfer request | ⓕ ⓢ ⓓ

3.



4.



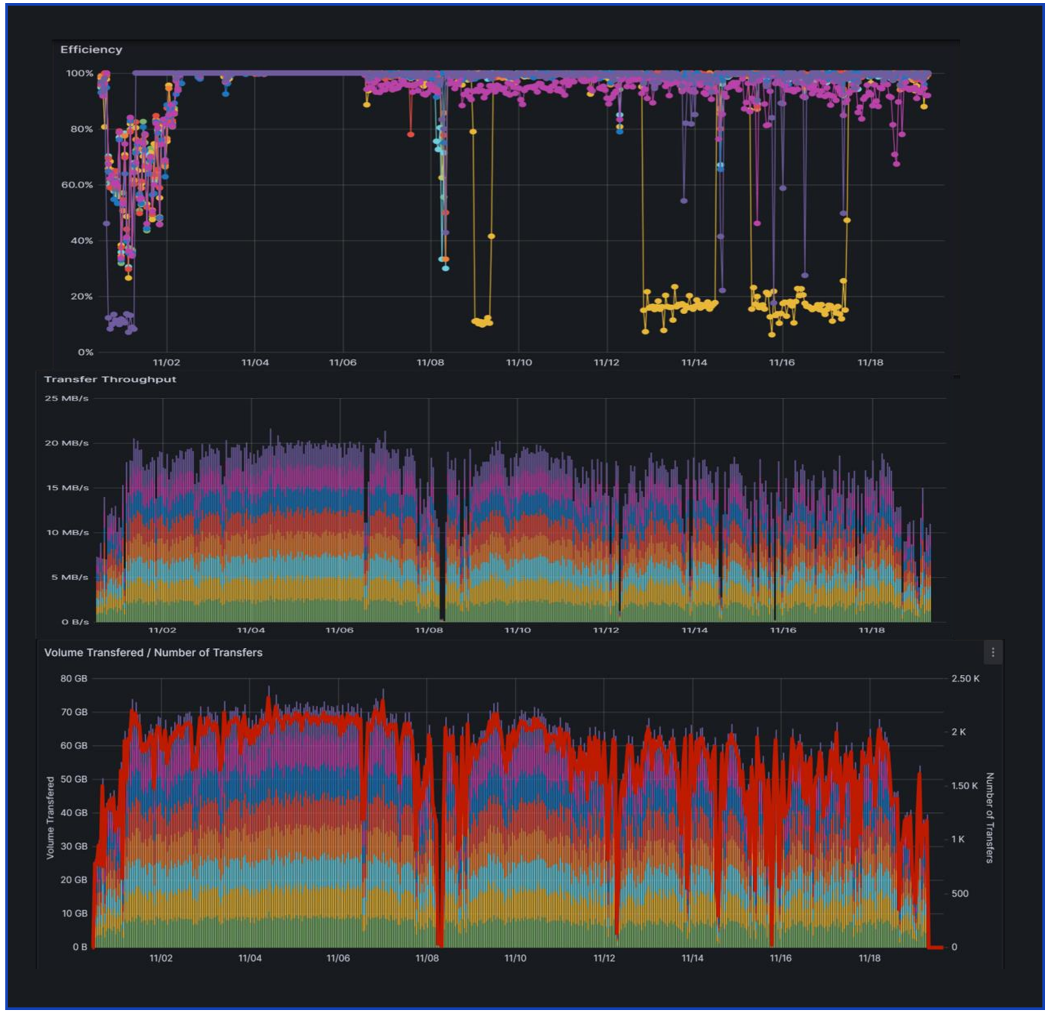
FTS
Source RSE
Destination RSE
1. check that source exists | ⓢ
2. check that destination does not exist | ⓓ
Conditional
if destination exists and overwrite is enabled
3. delete destination | ⓓ
4. perform the transfer | ⓢ ⓓ
5. validate that destination is in order | ⓓ

Science and Technology Facilities Council

Monitoring of Functional Tests

Functional tests across the UK data lake setup now using Rucio components (Automatix and Transmogrifier)

- Movement of around 60-70 GB an hour in tests, using 32MB files
    - Simulating LSST data flow
- Scalable functional testing tools to enable higher data rates and variable data sizes

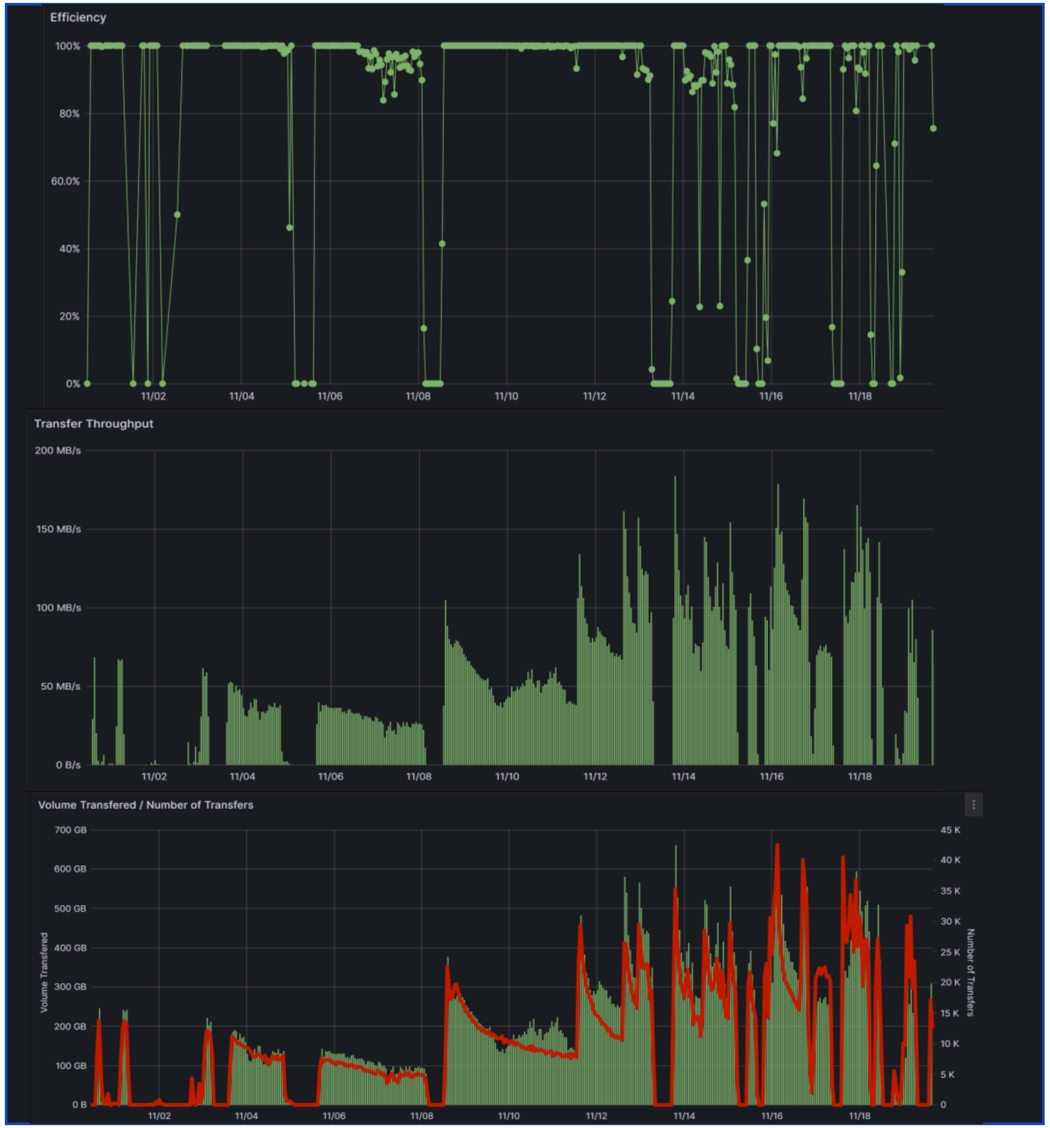


Science and Technology Facilities Council

Monitoring of Functional Tests

LSST data movement from the US to UK

- Movement of 80TB of data from LSST, peaks and trails off an interesting behaviour from many small files
- Many connections (280-480 active connections recorded in FTS)



Science and Technology Facilities Council

WP5 Analysis Facility

Rucio is capable of integrating directly with the analysis facility JupyterHub

- Work from REANA, Rucio can be integrated with JupyterHub
- Allows more flexibility for users
- Currently uses X509 and VOMS proxies for authentication
  - But from the documentation should be easy to swap to tokens

Plan for extension to project

- SSD pools and testing at sites
  - Setting one up at RAL
  - Lancaster has stated they are also going to set one up
- Development work on Rucio QoS to include and prioritise SSDs
  - more than just the tagging possible right now
- Work with LSST to develop better monitoring that will also benefit and give more information on the Rucio instance and data movement to better optimise the data movement