Multi-VO Rucio for EGI

Timothy Noble
Service owner and developer for Multi-VO Rucio
What is data management

• Data management is the practice of keeping and using data securely, efficiently, and cost-effectively.

• A robust data management solution becomes more necessary as the number of people accessing, generating, and sharing data increases across a number of sites.
Rucio

Overview

• Data management tool
  ▪ Integrates with many storage solutions
  ▪ Data can be stored across multiple sites, with different setups and protocols
  ▪ Data can be anything, images, text....
  ▪ Designed with more than a decade of operational experience in very large-scale data management (Several billion files, approaching an Exabyte of data)

• Developed by ATLAS, now used and developed by other communities

• Open source
  ▪ Community-driven development
Rucio

Terms

• Rule – relates to and protects a file and allow for replication
• Dataset – A collection of files
• Container – A collection of datasets
• RSEs – Rucio Storage Endpoint
• Daemons – A program that performs a specific task, there are many kinds
Rucio

Files, Datasets and Containers

• Single files can be replicated using rules
• Files are grouped together in datasets
  ▪ Can belong to multiple datasets
• Containers are collections of datasets
• Containers and datasets can have properties to protect datasets
  ▪ E.g. Open/closed – can have data added
Moving Data

• Replication Rules
  ▪ A file from a data source may be uploaded to Rucio to an RSE
    $ rucio upload -- rse storageendpoint scope:first_file scope:second_file
  ▪ This may not be where it is needed
    o Add a replication rule
      $ rucio add-rule scope:first_dataset scope:second_dataset 2 'country=uk'
    o This would add the two datasets to the RSEs in the UK, and have 2 copies of this dataset within the UK
      $ rucio add-rule scope:first_dataset scope:second_dataset 2 'country=uk\site=RAL-LCG2'
    o This would add the same two copies across the UK but excludes RAL as one of the potential sites to keep this data
  
• RSE expressions
Each account is given a quota of space

Rules that account has count towards their quota, but this does not mean that space is taken up at an end point

If two users have a rule on the same file, each has that space allocated to their quota but only one physical copy is present

If there is a rule on a file, the file will not be deleted

Primary and secondary data
Rucio

Architecture
Multi-VO Rucio at RAL

Setup and maintenance

- Rucio was initially setup in April 2018 at RAL to support SKA development
  - Evolved to supporting Multi-VO
  - Developers have worked with core dev team to develop Multi-VO functionality
- Contact point for new users and provide tutorials
  - Timothy.Noble@stfc.ac.uk
- Test transfers between sites to ensure endpoint functionality
- Working on documentation to help User and VO-Admin use:
  - https://www.gridpp.ac.uk/wiki/Rucio
  - To be made available to EGI when complete
- Rucio-Support@stfc365.onmicrosoft.com for support
Multi-VO Rucio at RAL

Advantages

• Running a Rucio instance that supports multiple VOs is beneficial for small experiments:
  ▪ Maintained by RAL not by smaller experiments
  ▪ Low levels of load from smaller experiments.
  ▪ One instance to support and maintain.
  ▪ Shared RSE configuration
  ▪ New VOs are quick to add – work with VO admin to setup their environment
• More contact with Developers and larger communities using Rucio to know how best to utilise
Accessing Rucio

WebUI and EGI Check-in / IAM

• Work in progress
• Next 3-6 months
Accessing Rucio

Clients

• Containerised Client
  ▪ Installed using docker
  ▪ Easy to install & setup

• Lightweight Client
  ▪ Easy to install & setup
  ▪ Fewer dependencies
  ▪ Run without a config file

• Rucio Desktop
  ▪ Graphical User interface for Rucio
  ▪ Developed as a summer of code project
  ▪ Work in progress

  ▪ [https://github.com/rucio/desktop](https://github.com/rucio/desktop)
Multi-VO Rucio at RAL

RAL Plans

• EGI-ACE Developments
  ▪ WebUI for Multi-VO
  ▪ Integration with IAM and EGI check-in
  ▪ Invite new experiments to use Multi-VO Rucio
  ▪ Improve documentation

• Internal Developments
  ▪ Multi-VO selection of certificates for daemons
  ▪ Containerisation
  ▪ Improve monitoring
  ▪ Gain experience with on boarding a variety of communities and requirements
Multi-VO Rucio at RAL

Conclusions

Multi-VO Rucio at RAL is available to anyone who wishes to use it

- RAL Multi-VO Rucio is working with EGI to offer an opportunity for communities to use the powerful data management software of Rucio without needing to invest in the setup and maintenance.
- RAL Multi-VO Rucio at RAL is being viewed as a long term service to support any community that wishes to use it.
- RAL Multi-VO Rucio is being deployed and developed with new users in mind to make it easier for communities who have no experience with Rucio or the grid environment.
Thank you for your attention.

Questions?