National Federated Compute Services NetworkPlus

Spring Conference
26-27 February 2026

Federated data movement

Alastair Basden, George Beckett, Shaun de-Witt, Manu Antony, Jonathan Hollocombe, Kieran Leach, Mark Lovell, Chris Mountford, James Perry, Mark Wilkinson

Durham University

COSMA

DiRAC

epcc

UK Atomic Energy Authority

UKRI

Quote or key statement: User-driven data movement at scale is not a solved problem!

National Federated Compute Services NetworkPlus

Data movement tools

Rsync, Rclone, Globus, Rucio, UPS

Landscape questionnaire

Multiple key use cases considered:
SKA UKSRC
Cosmology
LSST/Vera C. Rubin Telescope
UKAEA

And from the community:
Tessera, CryoEM, National X-ray CT, EU Open Science Cloud, Royce, Rosaline Franklin, MRC

These will be summarised for the final report. Please fill it out!

Summary of responses:
Dataset sizes are >100TB
File sizes typically 10GB-1TB
Both data to user and user to data models
Transfers can be continuous or discrete files
Data can be cut out of larger files
RUCIO has been adopted into some workflows
Rsync is achievable for 100TB scales (at night)
Some access restrictions may be required
Mixed user community expertise
HDF5 streaming over http, with a Python client. Cutting out regions of a 2PB dataset with ~10GB files.

Project team

Project Leads:
Alastair, George, Shaun

Co-leads:

James, Jonathan, Mark

Project Team Members:
Manu, Chris, Kieran, Mark


Quote or key statement:
The Data Federation project seeks to identify and remove obstacles for large scale data movement

Actions taken

Test installations and PoCs

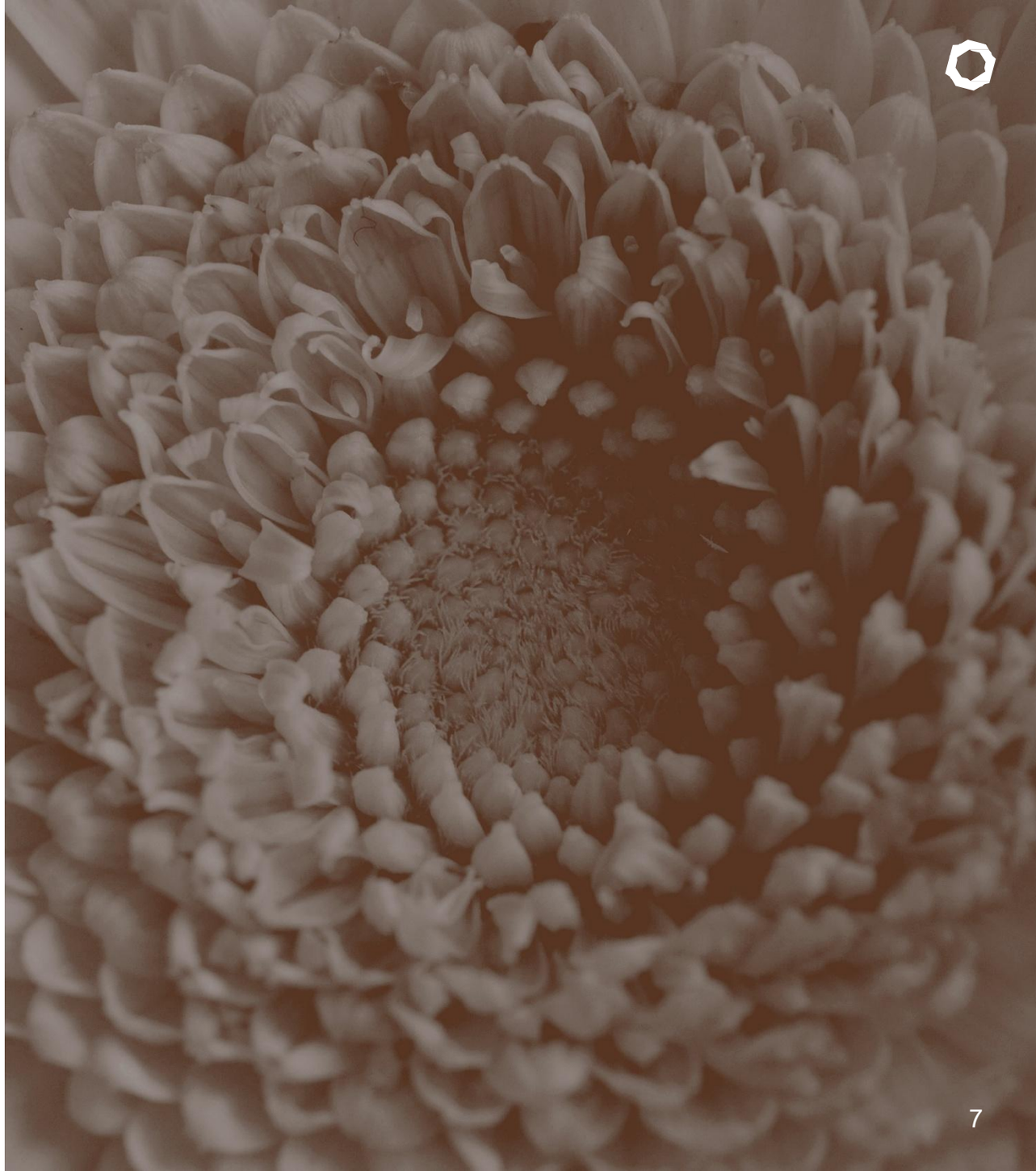
Globus: Investigation of transfers between sites – Cambridge, Durham, Edinburgh, JISC, IRI – exploring benefits of paid-for license

Rucio: Server installation and multiple clients

S3: StorJ on-premise cloud storage used with both Rucio and Rclone

Rsync: Transfer rates between sites

Step-by-step instructions given



The Dataset Generator

Dataset size affects performance

- A dataset generator has been created
- Simple Python code
- Give it a dataset name and it will generate the files
- Identical across sites
- Enabling repeatable benchmarking and testing
- Seeded random binary files of predefined size and count
  - e.g. 10x 50GB files, or 1000000x 1kB and 10x 1TB files

The TODO list

Complete Rucio installation
Automated workflow studies
Finish data transfer benchmarking
Collate completed questionnaires

Summarise in final report
Go on holiday!

Conclusions

Improved user accessibility
Enhanced data access
Expertise and experience sharing
RTP training and skill development
Data governance issues
Access legitimacy

Recommendations:
Sites should put effort into installing data movement tools
Globus is well regarded within the community