

Federated data movement

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

Alastair Basden
DiRAC / Durham University

 **Durham**
University
Institute for Computational
Cosmology

DiRAC
High Performance
Computing Facility

Data movement at scale

- A solved problem?
 - Possibly for IT professionals
 - Definitely not for users!
- UK DRI is linked by the state-of-the-art JANET network
- UK institutions have cutting edge data-storage and data-analysis services
- <https://durham.rtf.d.io/en/latest/datafed>

The user view

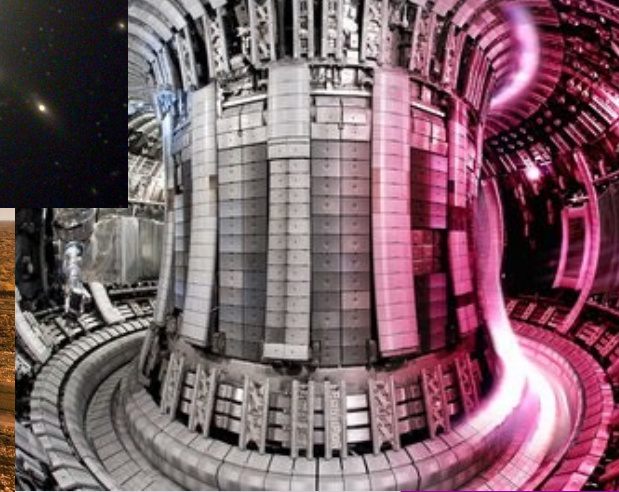
- Scp? Rsync?
- Possibly Globus?
- Rclone (depending on pedagogy)
- Can I send you a disk in the post?
- It's too hard: I won't bother

The data federation project

- Reduce these bottlenecks
- Provide clear guidance
 - Document what tools are available where
- Perform testing and benchmarking to set performance expectations
- Enable and facilitate researcher-led large scale distributed data movement

Use cases

- LSST
 - Transfer of telescope data and events for UK-led processing and scientific analysis
- UKAEA
 - Transatlantic transfer of simulated fusion data for UK-based analysis and processing
 - Engagement with US IRI
- SKA UKSRC
 - Data movement into and around the UK of multi-TB scale datasets
- DiRAC / Virgo
 - Transfer of large cosmology simulation datasets from production to analysis sites



Technologies

- FTS3
 - Xrootd and https – COSMA and Somerville
- RUCIO
 - COSMA
- Globus
 - DiaL, RAL, UKAEA, Somerville
- Rclone
 - COSMA, EPCC, DiaL
- S3, including DiRAC StorJ
 - UKAEA

Actions

- Deploy test installations of data transfer tools at multiple sites
- Implement PoC for accessible bulk data movement
- Provide guidance and repeatable benchmarks
- Compare 4 transfer technologies
- Investigate workflow integration with batch schedulers and object storage
- Consider sustainability aspects
 - Vendor lock-in, up-front and lifetime costs, standards compliance

Training

- Materials to be made available online
- Data transfer workshop at HPC Days 2026 conference

Governance

- Users connecting to accounts at multiple sites
 - Transferring data between them
- Ensure legitimate access
- Ensure secure data transfer as appropriate
 - Ensure correct access permissions
- Lightweight management board within the project

Business model

- Requirements for continued sustainable operation
- Determine minimum subscription levels to paid services
 - e.g. Globus
-

Key benefits

- Improved user accessibility
- Enhanced data access
- Expertise and experience sharing
- Upskilling of technical teams and users
- Federated knowledge in data sharing

