Using Globus for data management - the Monash experience

Monash eResearch Centre

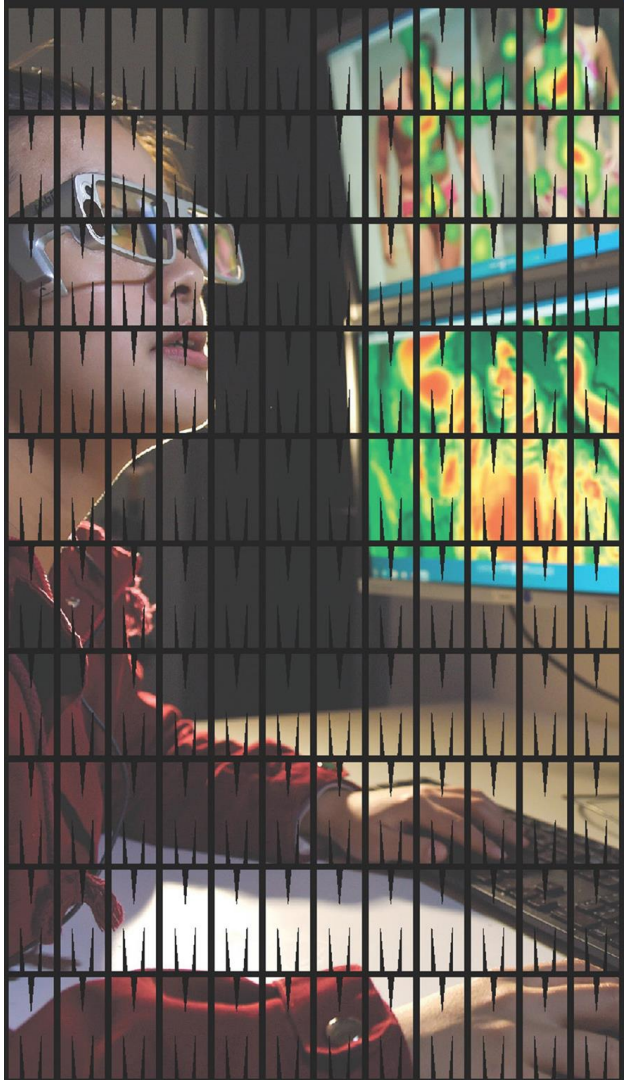
Date: 30/10/2024 | Presenters: Mitchell Hargreaves and Geoff Duniam

Globus at Monash

- Globus is a large scale data movement ecosystem
- Designed for the TB and PB scale
- In use with many educational and research institutions worldwide
- Globus is being implemented for large scale data transfer at Monash
- Working with various research groups (MMI, CryoEM etc) to securely transfer instrument data to long term and compute storage

In this presentation, we will share our experience implementing an automated data management workflow using Globus to transfer experimental results from two electron microscope data servers into our long-term storage platform.

MONASH University



Why Globus?

Benefits of using Globus

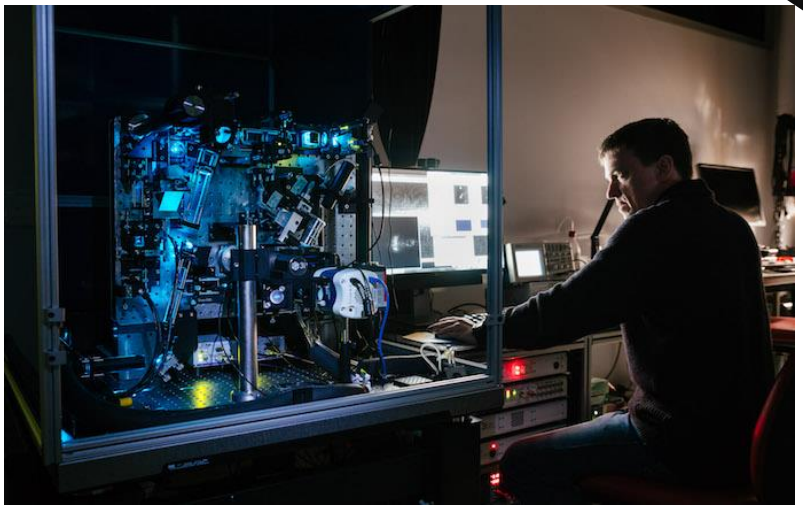
- File transfer integrity is supported by checksum matches before and after the transfer
- Automated transfer recovery and restart in the case of network or machine failure
- Data encryption for sensitive data
- Multi-threading and bandwidth control to optimise transfer speeds

Use Case- MMI

Monash Micro Imaging (MMI)

Why start with MMI?

- Creating data products >1TB for some experiments
- Need automation to clear data from instruments to allow research to continue
- Existing movement tools Monash has used break down at this scale



https://www.monash.edu/researchinfrastructure/mmi

1. https://www.monash.edu/researchinfrastructure/mmi

MONASH University

Use case requirements

Simple data movement?

The (generic) problem:

- Limited storage on instrument servers
- Movement of data to secure storage uncontrolled

The solution:

- Automated movement of experimental data into long term secure storage
- Frees up storage on instrument servers for further experimental data

Easy, no?

“And the user replied with a snarl and a taunt - It’s just what I asked for but not what I want!!”[1]

1. Anonymous systems engineer, circa mid 1980s?

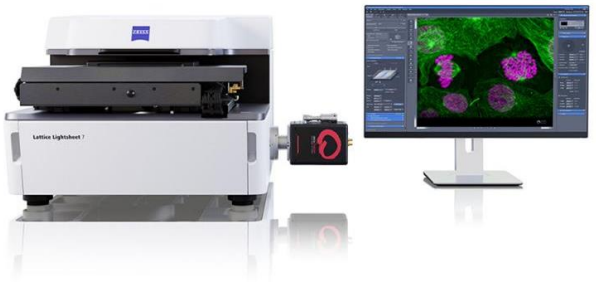
Use case details

The LLS7 Instrument

- Check that the target directory exists
- Copy data to target
- On success:
  - Delete raw source data

The Ultramicroscope 2 Instrument

- **Tar source files and create manifest**
- **Check that the target directory exists**
- **Copy data to target**
- **On success:**



https://www.zeiss.com/microscopy/en/products/light-microscopes/light-sheet-microscopes/lattice-lightsheet-7.html

MONASH University

Endpoints

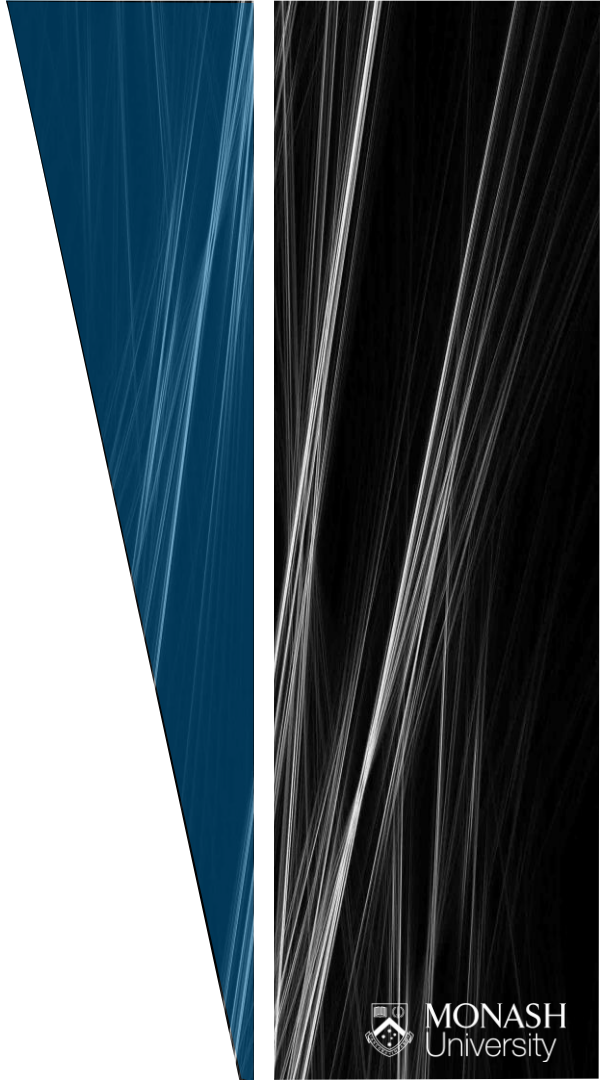
The Heart of the Matter!

Endpoints on Linux

- Globus Connect Server
- Single node and multinode endpoints
- Storage connectors - Posix, Ceph, S3, etc
    - We are exclusively using Posix

Endpoints on Windows instruments

- Globus Connect Personal - single user endpoint
- Guest Collections for extended exposure



Automated Process flow design

The general Globus move and delete process flow

- Written in Python
- Globus SDK calls incorporated into a Python class
- Seperate class for archiving, manifesting archives and the hash encryption of user data if required
- Collection, source and target directories, and endpoint contained in a .json configuration files passed in at program execution
- Processes initiated from an independent Linux VM via cron

Automated Process flow design

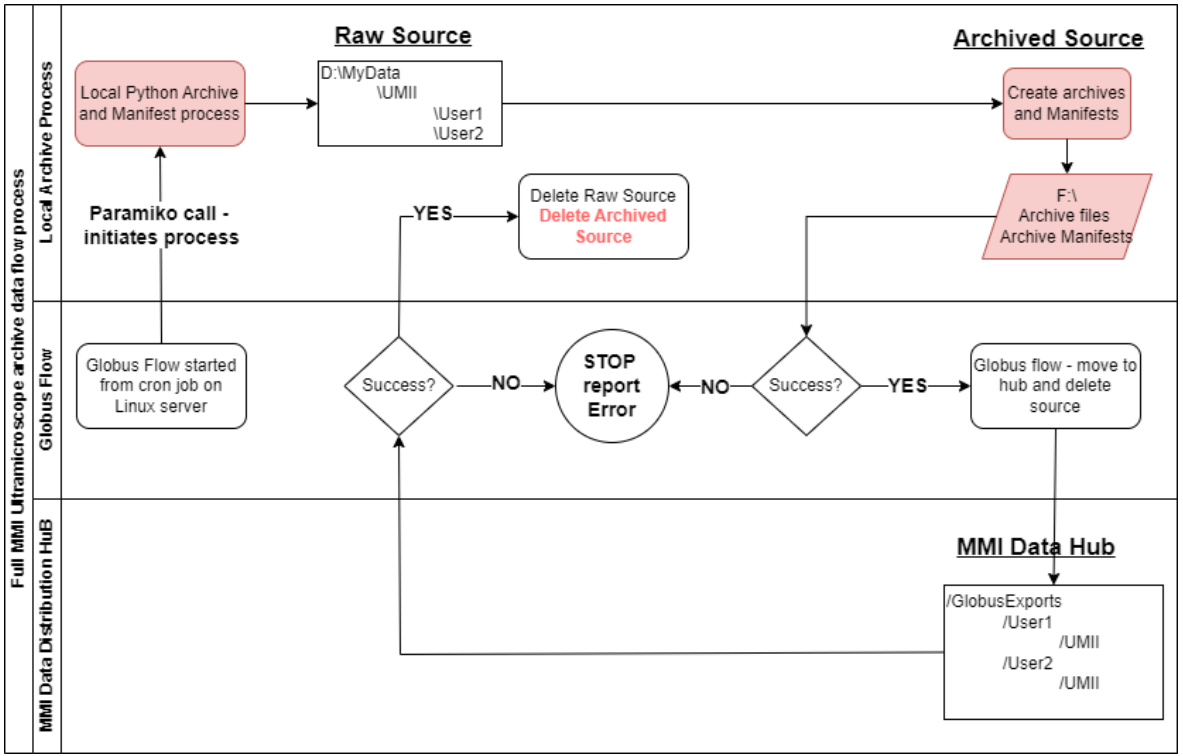
(continued)

Archiving source data

Functionality for archiving source material into .tar files, and creating manifest files is not native to the Globus API, hence the development of a separate class.

- Some experimental data can comprise tens of directories and thousands of small files
- Globus is more efficient at transferring a single large file
- Long term tape storage systems do not perform well with thousands of small files
- However - there are trade offs that need to be considered.

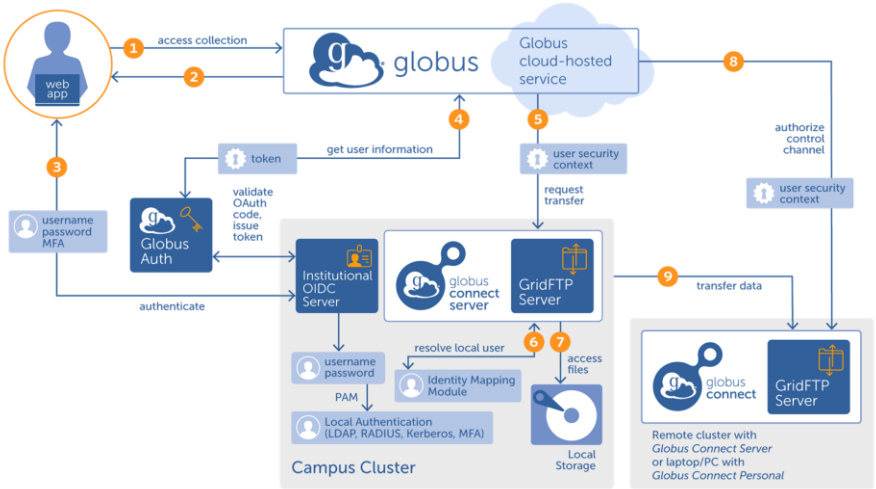
The MMI FlowProcesses



Authentication

Globus Posix endpoints map globus users back to linux users

- This allows you to expose the storage to the end users
  - Control user access from the file system
- Globus will bounce the user to SSO
- Mapping options:
  - Expression
  - Mapping Application



https://docs.globus.org/guides/overviews/security/reference-architecture-existing-oidc/

MONASH University

Authentication

Things we found

- Globus doesn’t respect sticky bits on folders
- Group permissions don’t scale well with NFS
- Expression based mapping isn’t robust
- Mapping applications are more complicated, but ultimately worth it

Conclusion

- Globus has proved to be an effective tool for migrating data
- Using it gives you access to the Globus ecosystem
  - Allowing transfer between other facilities with Globus
- Scheduling transfers with the Python SDK allows for building extra functionality
- Globus identity management allows for transfer between endpoints though this comes with complexities

THANK YOU

Questions?

Monash eResearch Centre
15 Innovation Walk, Building 75, Clayton Campus
Wellington Road, VIC 3800
Australia
monash.edu/eResearch