Towards data sharing service for Physical Sciences Data Infrastructure

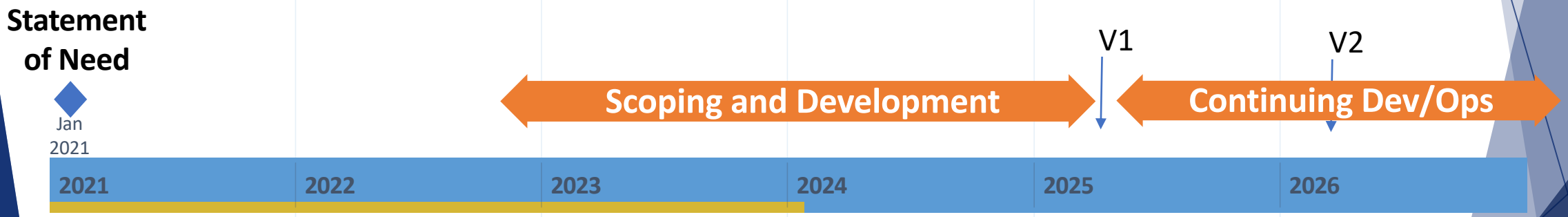
Jonathan Bathe, Vasily Bunakov
Science and Technology Facilities Council, UK Research and Innovation

CS3 2024 – Cloud Storage Synchronisation and Sharing
11–13 Mar 2024, CERN

PSDI: Timeline

Funding stages:

Project stages:



www.psdi.ac.uk

PSDI is funded by EPSRC through Digital Research Infrastructure programme (EP/X032701/1 and EP/X032663/1)

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Modalities of data sharing

- Peer-to-peer (irregular) sharing
  => EFSS can be useful but is not necessarily required
- Regular sharing within the team
  => having EFSS is beneficial
- Regular sharing between the teams
  => having EFSS is essential
- Regular sharing globally
  => it cannot be just any EFSS, and non-functional aspects are important

The same modality of data sharing can be supported by different deployment topologies and operational models

Choice of technology could be a driver for deployment topology and operational model, e.g. if incremental federated deployment is possible

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

We have not made up our mind (yet)

- We looked into three solutions for EFSS, and thought we might prefer one of them but…
- We consider adjacent / complementary technology beyond EFSS: for data migration, for backups, for virtual file systems
- PSDI is relatively small, and have no intention of putting effort in full data tech stack (but rather, capitalise on the existing foundational infrastructure available to us)
- We are here to listen and learn
- We are attending in person for three days of CS3 only (not for the rest of CERN's Data Tech Week)

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

What is needed?

- Open Source
- Federation
- Low thousands of users
- Role based user control
- Support of backends we have (CephFS or S3)

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

"Shortlisted" Solutions

https://owncloud.com/infinite-scale-4-0/

https://www.globus.org/

https://www.onedata.org/#/home

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

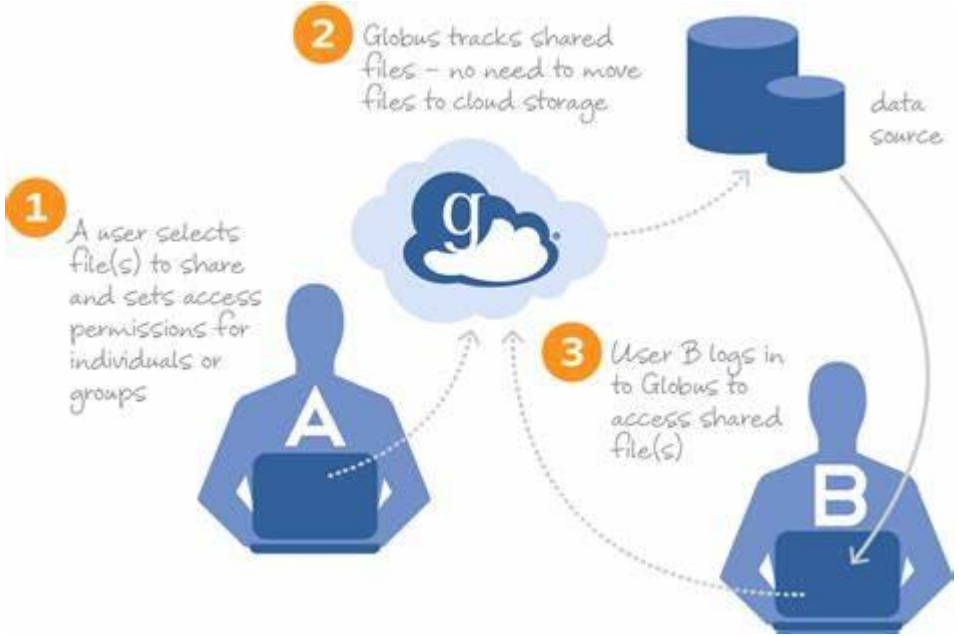
OwnCloud Infinite Scale (OCIS)

- Used by CERN in CERNBox
- Built on Microservice architecture
- Programmed in GO
- Free to use on own servers
- Large community across various platforms
  - OwnCloud forum
  - Tech help forums (e.g. Stack Overflow)

https://owncloud.com/infinite-scale-4-0/

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Globus



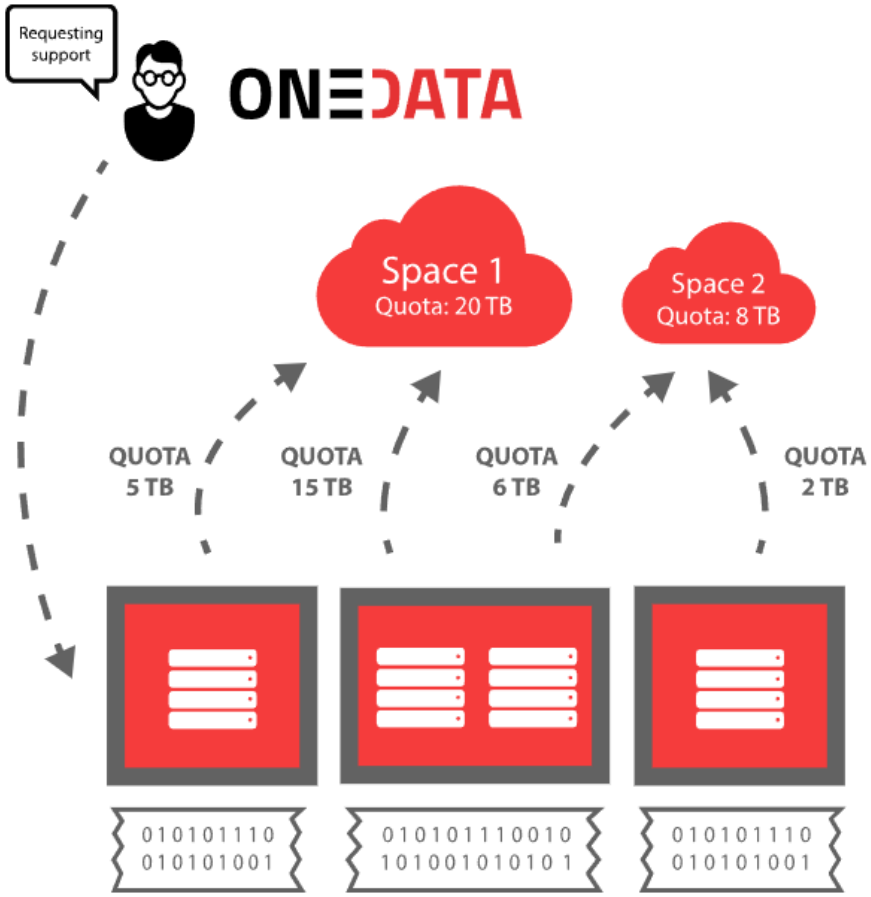
- Yale, Cornell University, HudsonAlpha
- Built Upon GridFTP
- Suitable for High performance file transfer
- Basic functionality is free
  - Premium features can be purchased on a subscription model
  - Premium connectors
  - Priority support
  - Metadata indexing

https://www.globus.org/

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

OneData

- Solution behind European Grid Infostructure https://www.egi.eu/service/datahub/
- Spaces
  - Allows for easy tracking and access management of data
- Providers
  - Allows for easy federation of storage
- Zones
  - Allows for easy access and transfer of data



https://www.onedata.org/#/home/documentation/21.02/intro.html

https://www.onedata.org/#/home

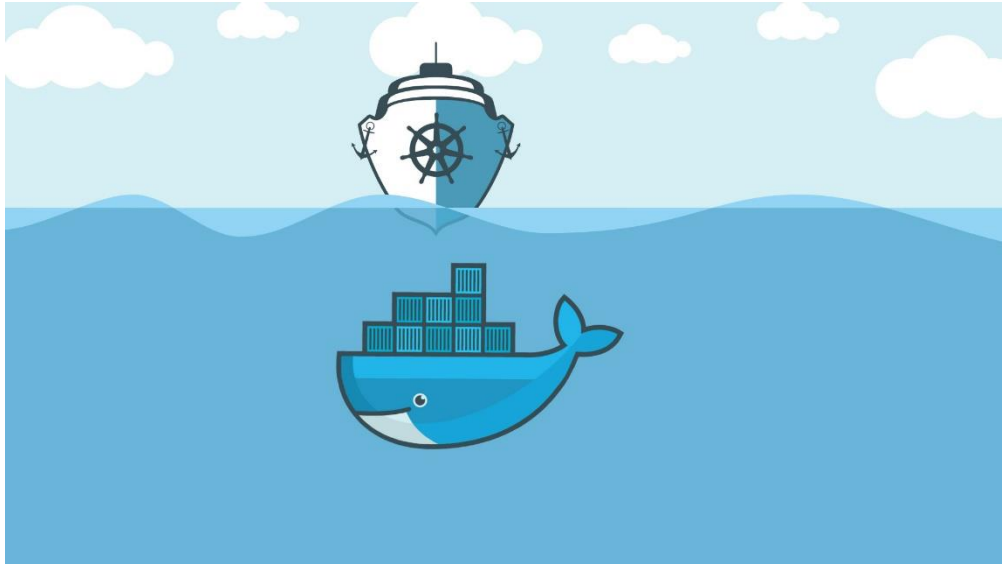
PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Evaluation Matrix

Requirements	OneData	Globus	OwnCloud Infinite Scale
The solution must be elastic by allowing for scaling up or down based on user demands	3 OneData has no restrictions on data transfer, and providers control the amount of storage space available.	3 Globus is somewhat scalable as it can have no restrictions on the amount of data that can be stored or the rate of data transfer, and each user can transfer more than a sufficient amount of files per task	5 OwnCloud is built in a microservice architecture, thus allowing for each microservice to be scaled based on its users' needs
The solution must be open-source.	5 OneData is open-source and free	3 Globus is open source, but only the basic functionality of Globus is free to use any special connectors that are needed required to be part of the subscription	5 OwnCloud is open-source and free to use
The solution must have a community that can provide support.	2 OneData has a small community presence, mainly on GitHub, where contributors post updates	5 Globus has a large and often active community on Reddit and social media like Facebook, and it occasionally holds conferences	3 OwnCloud has a significant community presence on its website, but only a small part uses Infinite Scale.
The solution should provide good documentation, support and training.	3 OneData has good documentation and can provide some support; however, this is essential support with setup provided on a case-by-case basis.	5 Globus has expansive documentation; they provide a useful welcome pack for user training and options for purchasing additional support.	2 OwnCloud has expansive documentation; however, it does not provide support of infinite scale.
The solution must have some form of access control, restricting particular	3 OneData has accounts that can have different privileges, allowing them to do different actions within Spaces, groups, OneZone and Handling services.	5 Globus has Globus accounts and can be used to create teams. The teams themselves can have restricted access; within the teams, sub-teams can be created that have their restrictions, and individual accounts can have restricted access	3 oCIS can put people in groups to share files only within the group and allow only certain users to do certain things within the file like some users can have read-only rites.
The solution must be accessible in multiple ways (e.g. Console and Web interface)	3 OneData has a web interface, Console interface	4 Globus has a web and command line interface both with expansive documentation and videos to help with use	3 oCIS has a web interface and a console
The solution should allow administrators to define access rights to shared files.	3 OneData allows administrators and owners of files to give access rights to the files	4 Globus allows administrators and owners of files to give access rights to the files and additional rights to different parts of the system.	5 OwnCloud allows administrators and owners of files to give access rights to the files and additional rights to different parts of the system.
Overall	**47**	**52**	**51**

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Experimentation



- Deployment
  - Kubernetes
  - Docker
- User Interface
- Pathfinder (use case) Testing

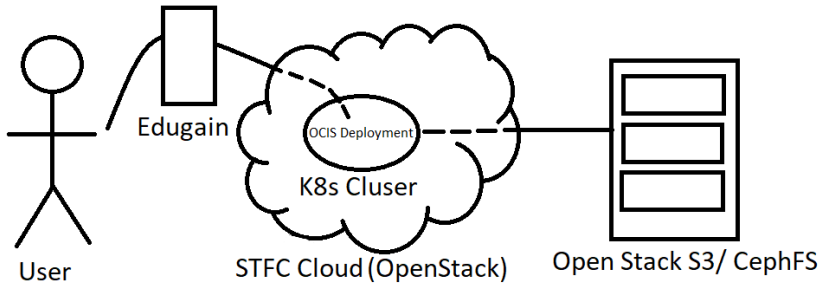
PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Current Position

- Looking at deployment of ownCloud Infinite Scale
- Looking into deployment options
  - Kubernetes
  - Docker
- Opening to select pathfinders for testing

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Planned Deployment



- Kubernetes in STFC Cloud (OpenStack)
- Backend (Open Stack S3) (CephFS)
- eduGAIN
- Limited availability → Open user registration

PSDI
PHYSICAL SCIENCES
DATA INFRASTRUCTURE

Thank you!