CS3 MESH4 EOSC

Connecting European Data

Data Transfers in the Science Mesh

Ron Trompert

SciDataCon 2021

CS3MESH4EOSC has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 863353.

On-demand data transfers

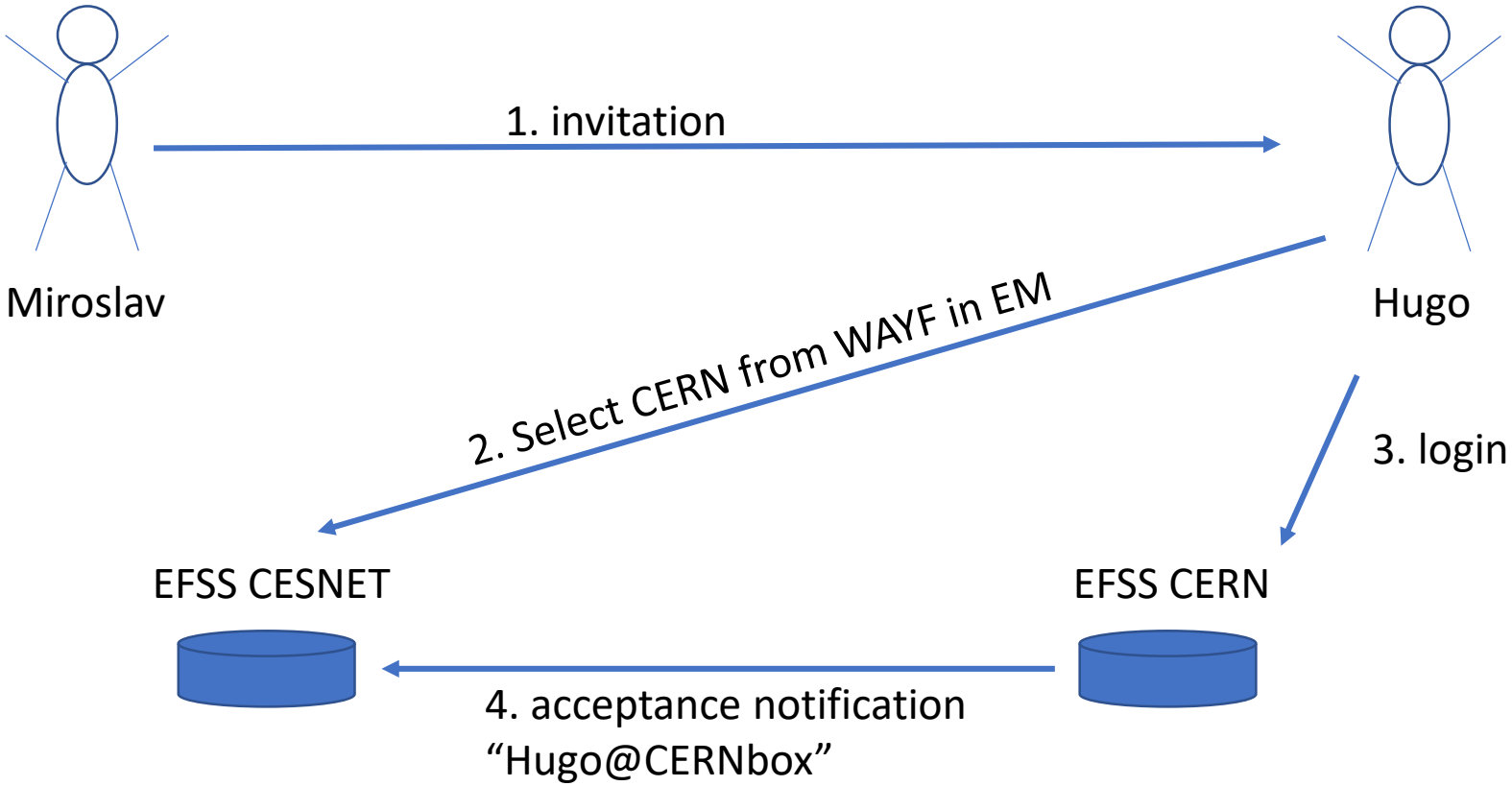
- Why on-demand data transfers?
  - Data locality may be important for some use-cases
    - Sharing of data is not good enough
      - Remote processing may cause overload on the source EFSS system
    - Larger data sets

- Support for data transfers between EFSS systems and between EFSS systems and other storage systems

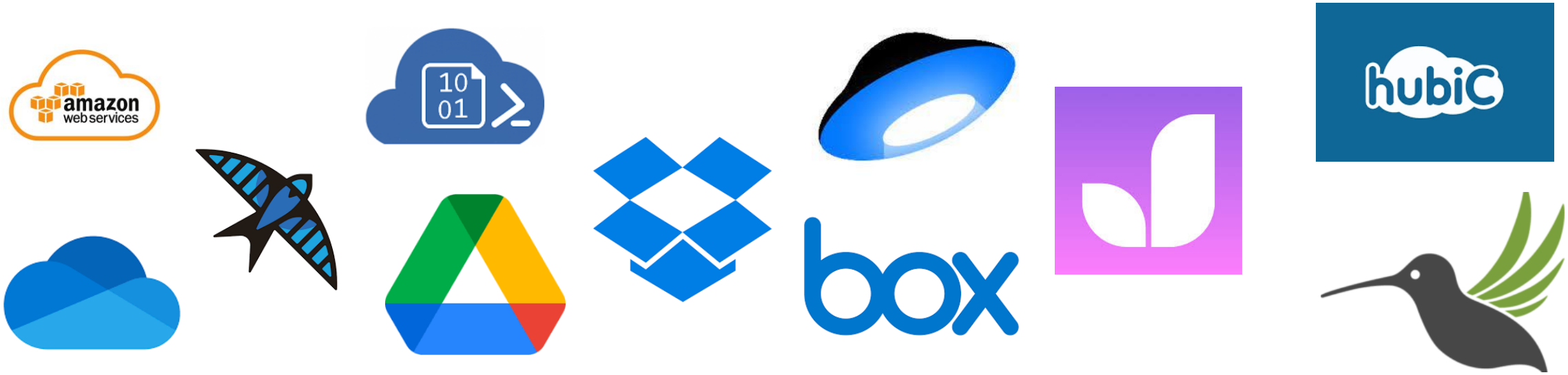
On-demand data transfers

- Different 3rd party data transfer mechanisms
  - Rclone
    - For small/medium sized data transfer needs
    - Individuals or individual research groups
  - Rucio/FTS
    - Connecting Big Science with sync-and-share (EFSS)
    - Communities

Establishing trust



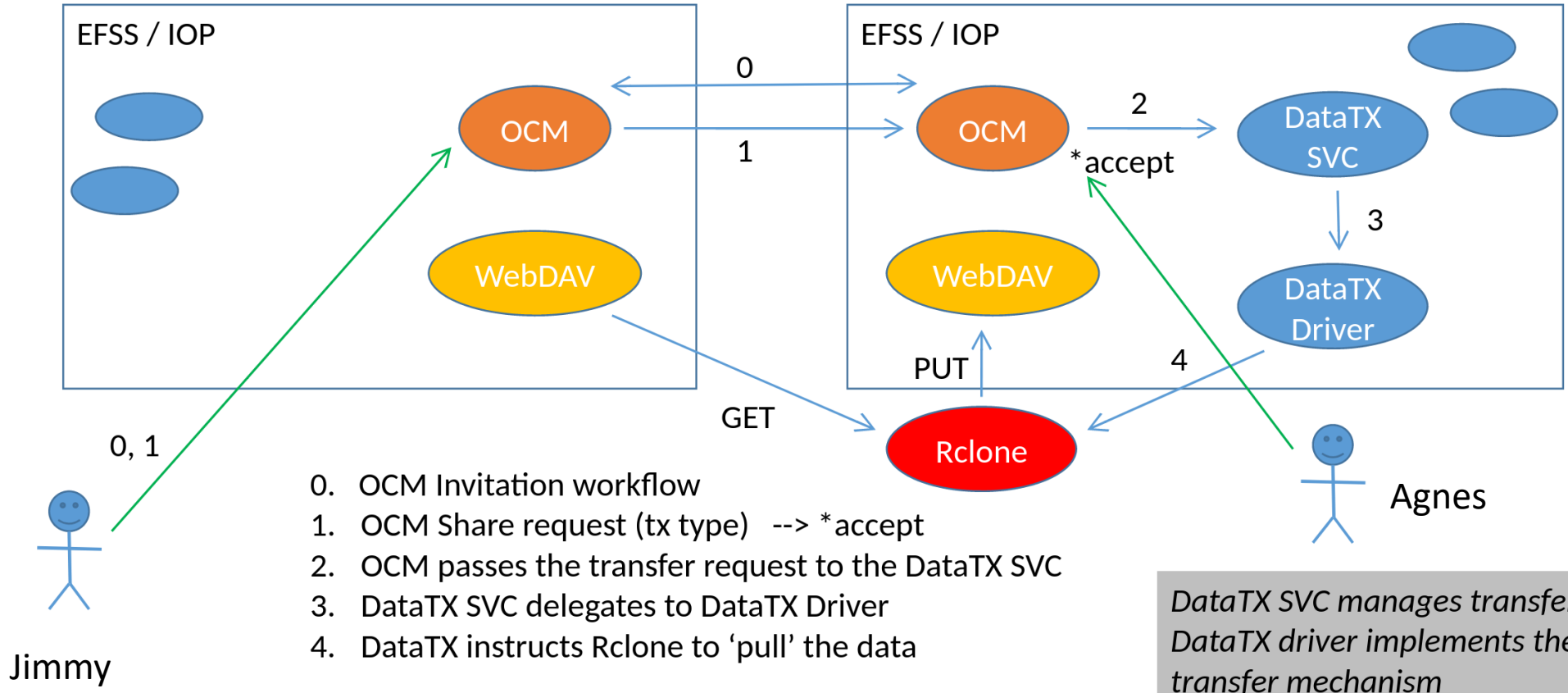
- Commandline programme to manage files on cloud storages
- It can interface with over 40 cloud storage products



- Implements various unix commands like rsync, cp, mv, mkdir, mount etc. for cloud storages

Rclone-based data transfers

Data transfers between two EFSS' based on Rclone



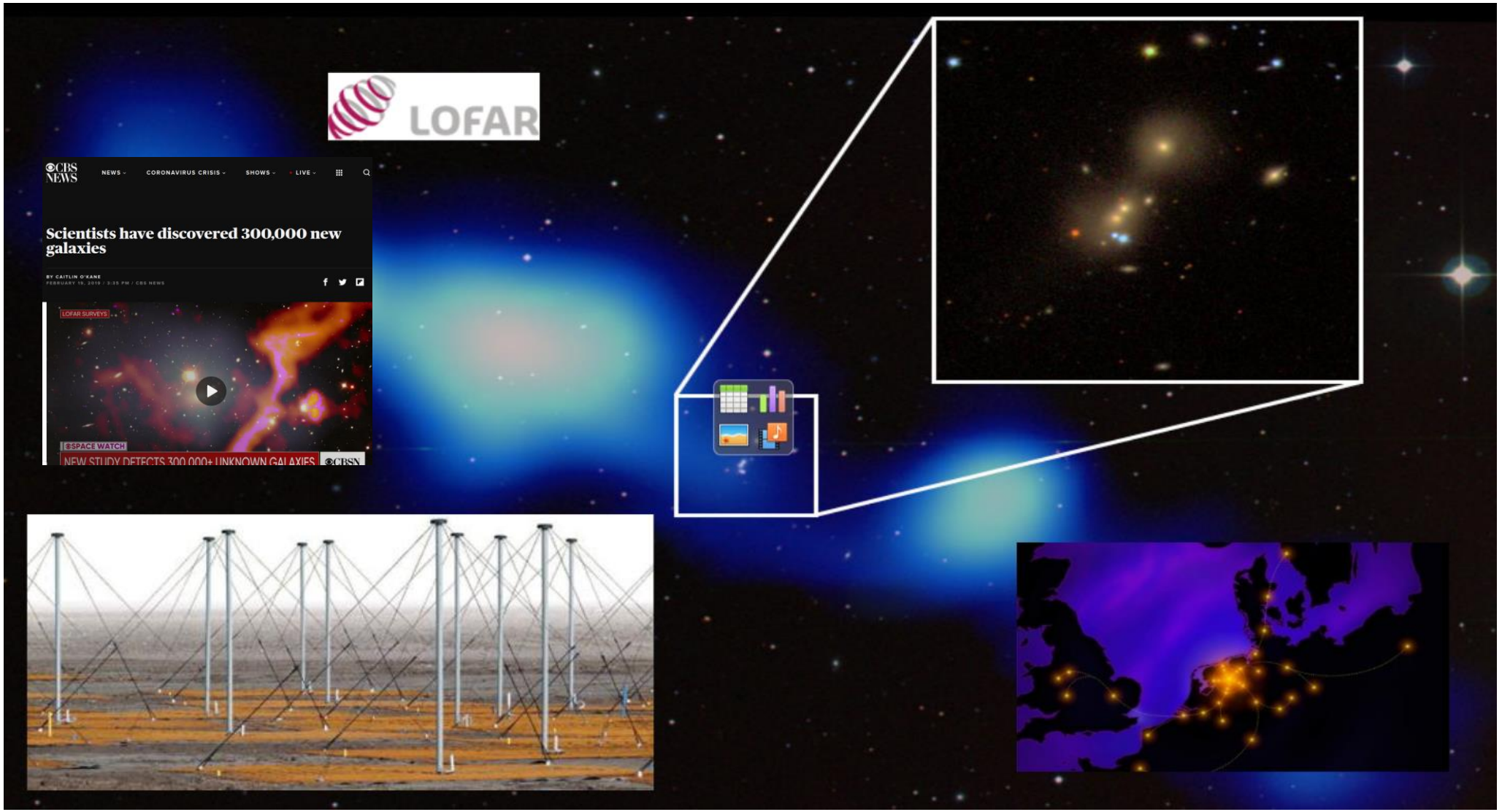
0. OCM Invitation workflow
1. OCM Share request (tx type) --> *accept
2. OCM passes the transfer request to the DataTX SVC
3. DataTX SVC delegates to DataTX Driver
4. DataTX instructs Rclone to 'pull' the data

DataTX SVC manages transfers
DataTX driver implements the transfer mechanism

Rclone-based data transfers

- https://youtu.be/m1PIzTrazoA

Use-case: LOFAR



Dataset transfer between research groups

Data stored at SURF and FZJ.
Initially processing (64x reduction).

LOFAR

LOFAR Surveys Key Science Project
Collaboration between researchers

- Leiden University and ASTRON (NL)
- Jagiellonian University, Kraków (PL)

On demand large dataset transfer

Science Mesh

RCLONE

FTS

RUCIO

Data shipped to Kraków
for creating science quality images

- Data transfer scheduler
- Developed by CERN to distribute LHC data
- Adopted by many other projects

- Data management
- Developed by ATLAS
- What does it do?
  - Managing data collections
  - Metadata management
  - Replicating data to registered storage elements
  - ……
- Uses the FTS for data transfers
- Not only used by ATLAS, but also by CMS, SKA and many others

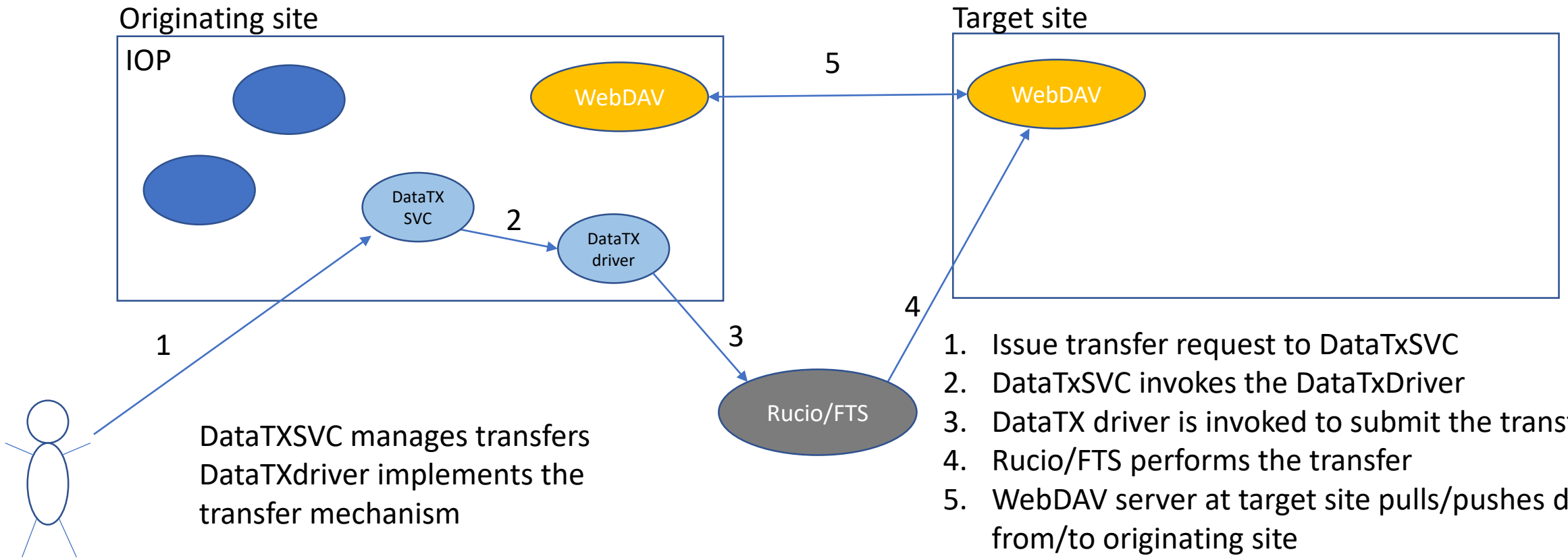
RUCIO
SCIENTIFIC DATA MANAGEMENT

Rucio/FTS-based data transfers

- Work in progress
- Google summer of code project (Rahul Cauhan)
  - https://hepsoftwarefoundation.org/gsoc/2021/proposal_rucio_cernbox.html
  - WebDAV third-party transfers for EFSS
- EFSS site as a:
  - FTS WebDAV endpoint
  - Rucio RSE
- Federated Identity Management
  - IAM (Indigo Datacloud)
  - EFSS', FTS and Rucio use the same IAM

Rucio/FTS-based data transfers

- Data transfers between an EFSS site and a regular or EFSS site using Rucio/FTS



DataTXSVC manages transfers
DataTXdriver implements the transfer mechanism

1. Issue transfer request to DataTxSVC
2. DataTxSVC invokes the DataTxDriver
3. DataTX driver is invoked to submit the transfer
4. Rucio/FTS performs the transfer
5. WebDAV server at target site pulls/pushes data from/to originating site

CS3 MESH4 EOSC

Connecting European Data

Thank you!

Discover more on...

cs3mesh4eosc.eu

company/cs3mesh4eosc

CS3org

CS3MESH4EOSC Project

https://www.youtube.com/channel/UCHKcZEkMqXjCvc3MLFjFxbw

CS3MESH4EOSC has received funding from the European Union's Horizon 2020 Research and Innovation programme under **Grant Agreement No. 863353**.