

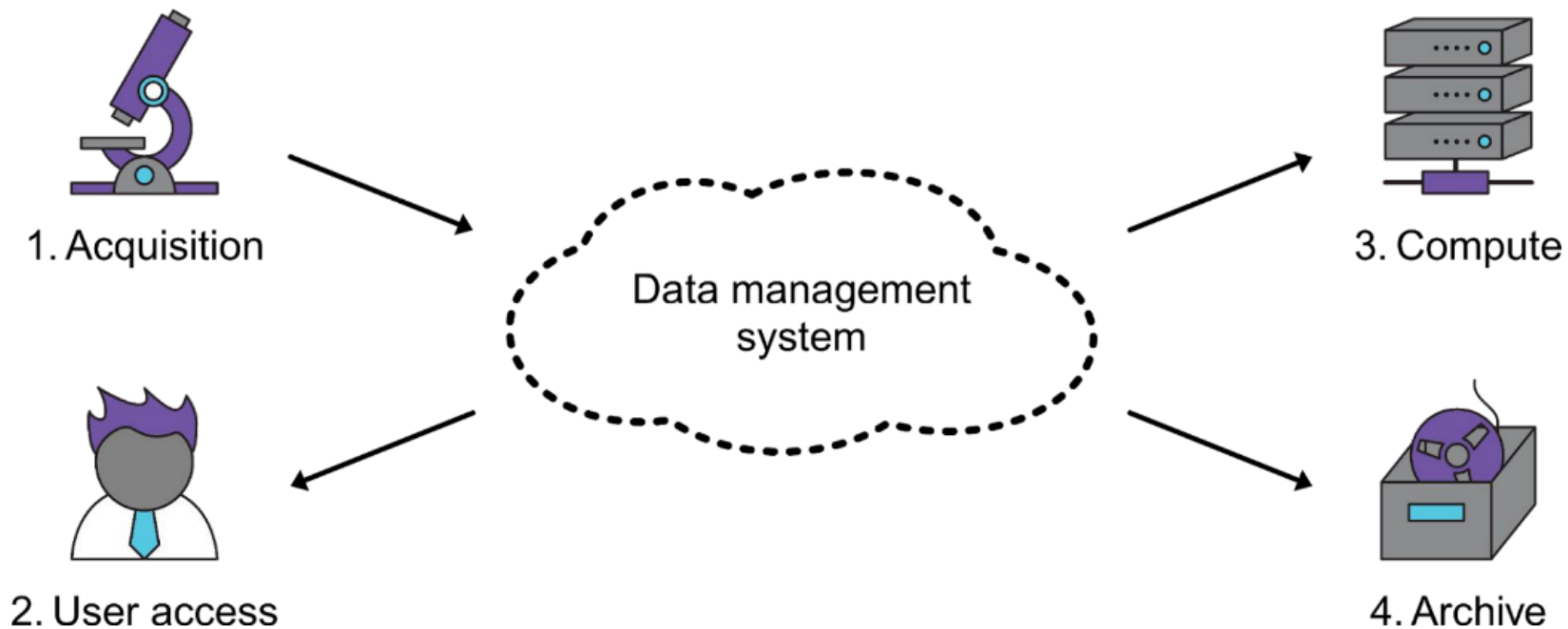
Hands-on experience with big data transfer via iRODS and Onedata

Ondřej Filip, Tomáš Svoboda, Adrián Rošinec

ondrej.filip@vsb.cz, tomas.svoboda@cesnet.cz, adrian@muni.cz

Motivation

- Efficient way for large data transfers on long distances
- Automatization of data management



Challenges

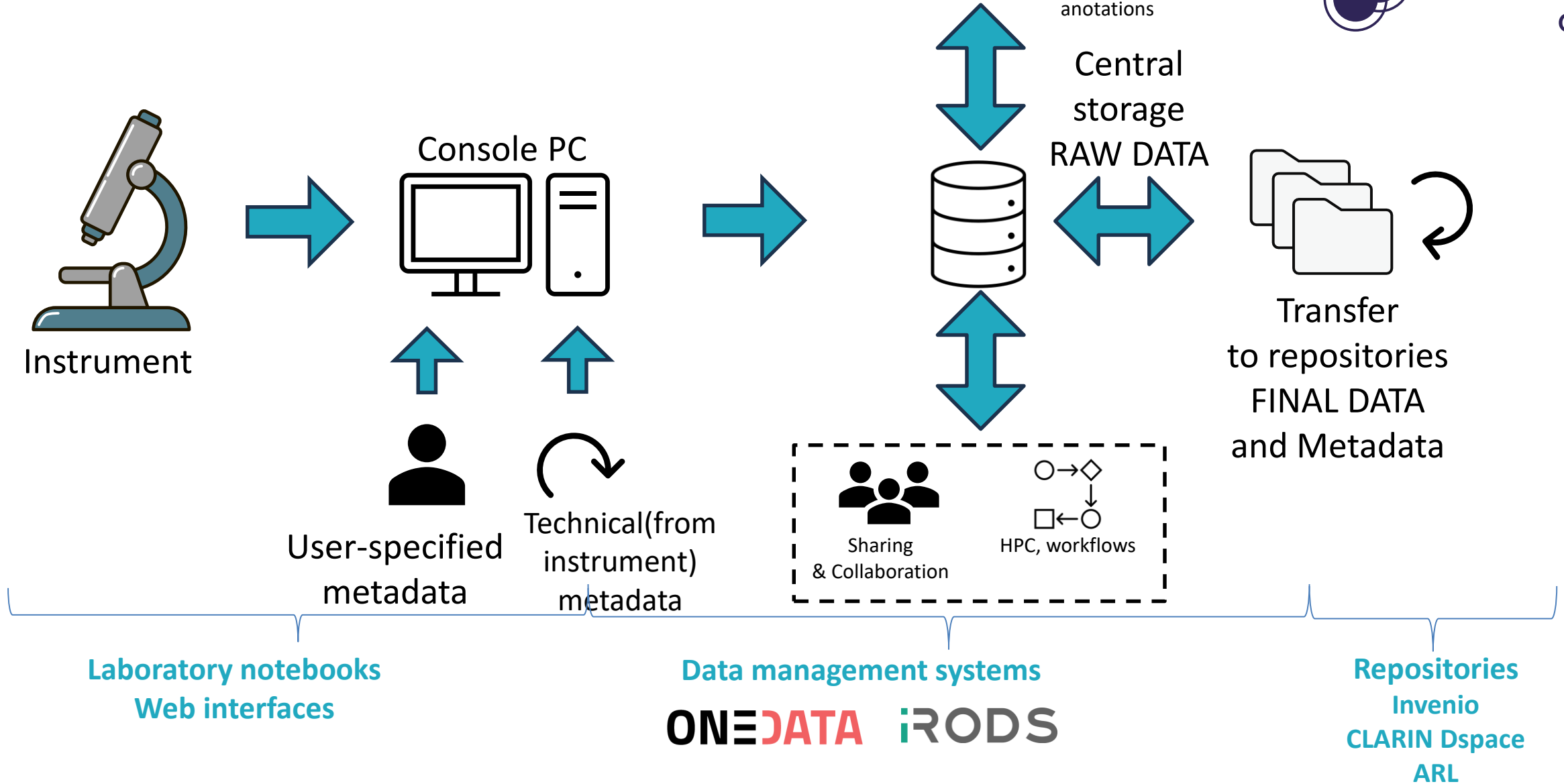
- Solution for data handover
 - How to get data from instrument to the user and his/her scientific publication
 - Avoid using USB/CDs, public clouds, ...
- Large data transfers over large distances
 - From CZ to European Supercomputer in Finland
 - Sharing large data between distributed communities (CZ – Spain, Denmark, ...)
- Metadata management (extractors, schemas)
- Publishing of data to corresponding community data repositories
- Preventing data loss
 - Obtaining data is expensive – instruments and samples
- Automatization of data “FAIRification”



Dataset life cycle scheme



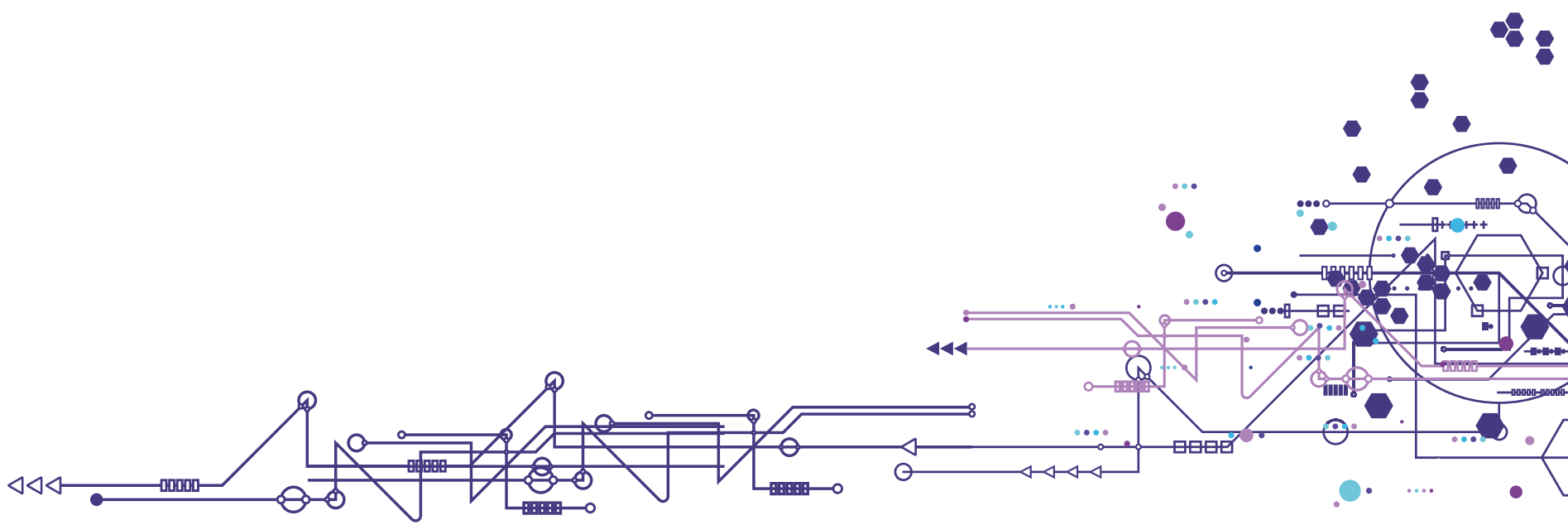
- ◇
 - ◇↓○
 - ←○
- Pseudonymization
 - Licensed SW
 - Compute metadata annotations





Onedata

Tomáš Svoboda



Motivation

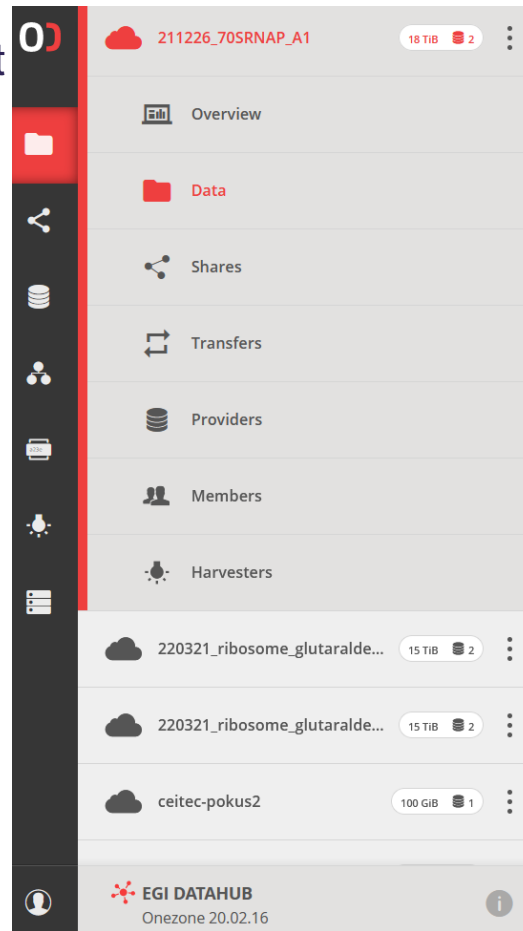
- Data management solution for two CryoEM facilities
 - CEITEC MU (Brno)
 - CNB CSIC (Madrid)
- Goals:
 - Handover data to users
 - Transfer data to HPC
 - Transfer data to archive storages
- Let's try Onedata




Onedata

Introduction

- Globally distributed data management system
- Open-source, on-premise
- Dropbox-like access
- Adapted for
 - HPC
 - Scientific data (FAIR)
- Data access
 - Web, desktop application, API, Python lib
 - Jupyter Ntbs., Kubernetes storage driver



DATA

View provided by  MUNI-ICS. [Choose other Oneprovider...](#)

211226_70SRNAP_A1

Selection (1)

Files	Size	Modification
.temp.Movies	—	4 Jan 2022 17:33
Gain	—	4 Jan 2022 17:34
Movies	—	31 Dec 2021 8:42
SPA.yml	3 KIB	27 Dec 2021 23:38

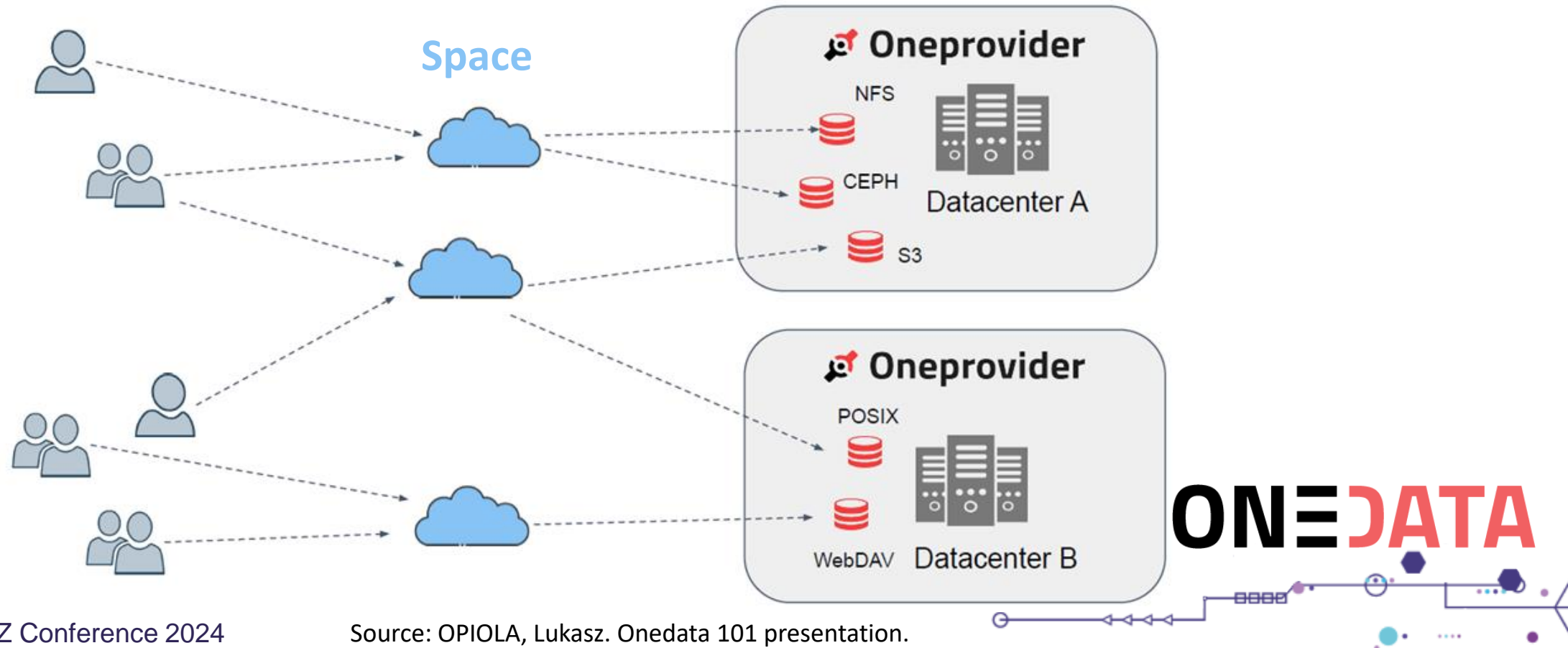
ONEDATA



Onedata

The basic idea

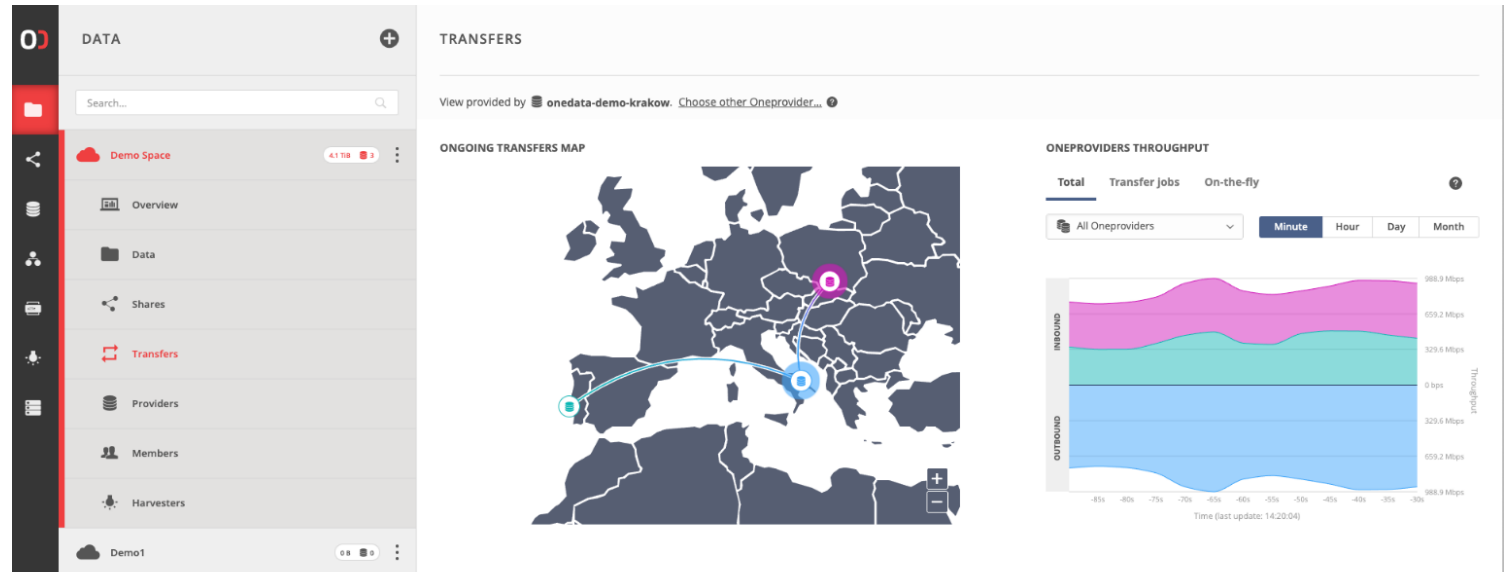
- Logical (virtual) data container
- Unified data access
- Emulating a POSIX-like filesystem
- Different locations and technologies



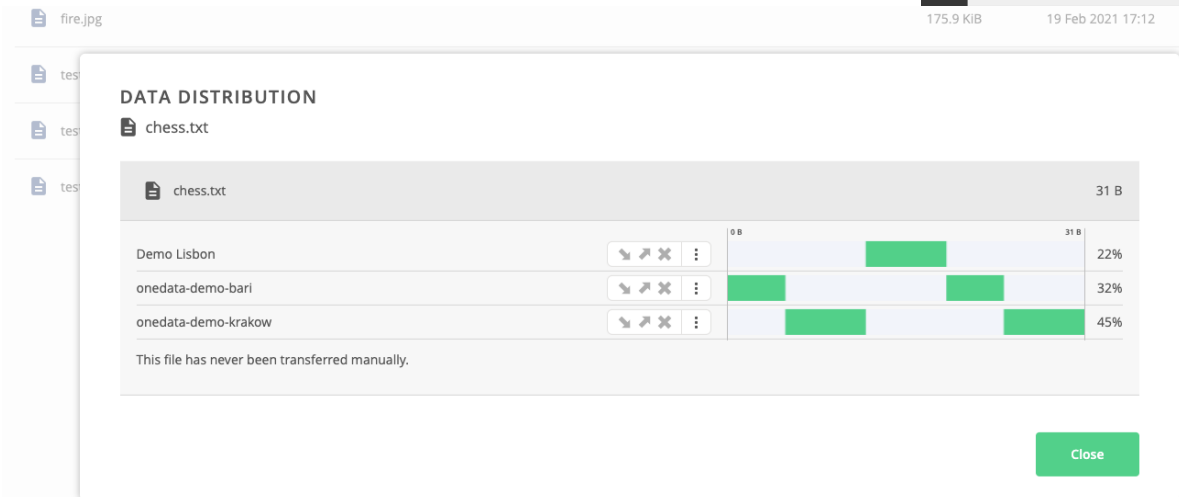
Onedata

Data distribution

- Transfers
 - On demand or on-the-fly
- Quality of service (QoS)
 - According defined policies
- Overview of data transfers and distribution



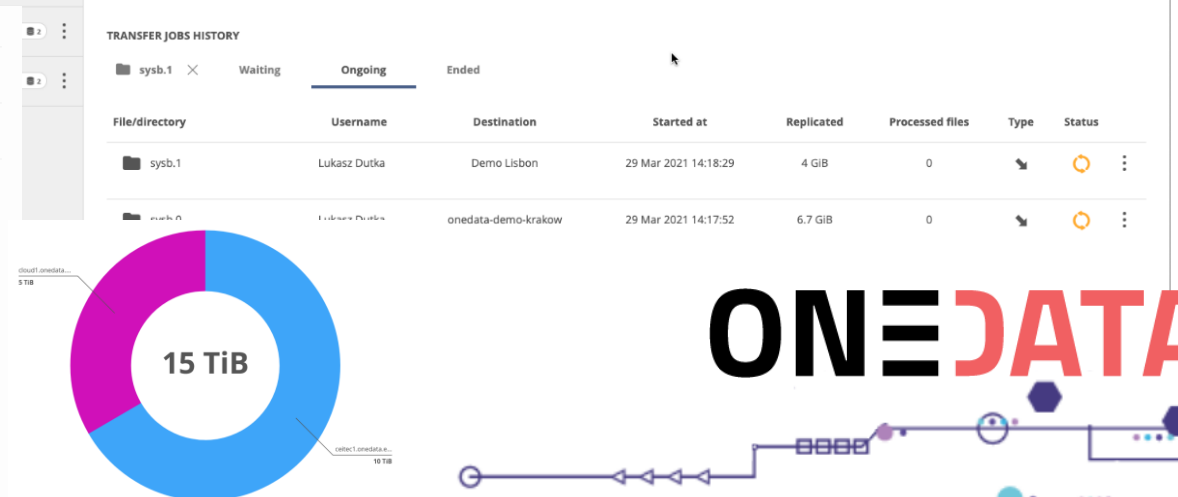
The screenshot shows the Onedata web interface. On the left, the 'DATA' section displays a sidebar with navigation options: Overview, Data, Shares, Transfers (highlighted), Providers, Members, and Harvesters. The main area shows a search bar and a 'Demo Space' with 4.1 TiB. On the right, the 'TRANSFERS' section includes an 'ONGOING TRANSFERS MAP' showing a map of Europe with data transfer paths between locations. Below the map is a 'TRANSFER JOBS HISTORY' table with columns for File/directory, Username, Destination, Started at, Replicated, Processed files, Type, and Status. A 'ONEPROVIDERS THROUGHPUT' chart shows Inbound and Outbound data flow over time.



The 'DATA DISTRIBUTION' dialog box for the file 'chess.txt' (31 B) shows the following distribution:

Destination	Percentage
Demo Lisbon	22%
onedata-demo-bari	32%
onedata-demo-krakow	45%

A note at the bottom states: "This file has never been transferred manually." A green 'Close' button is at the bottom right.



The 'TRANSFER JOBS HISTORY' table shows the following data:

File/directory	Username	Destination	Started at	Replicated	Processed files	Type	Status
sysb.1	Lukasz Dutka	Demo Lisbon	29 Mar 2021 14:18:29	4 GiB	0	👤	🟡
sysb.0	Lukasz Dutka	onedata-demo-krakow	29 Mar 2021 14:17:52	6.7 GiB	0	👤	🟡

Below the table is a donut chart showing a total of 15 TiB. The chart is divided into two segments: a blue segment representing 10 TiB and a pink segment representing 5 TiB.

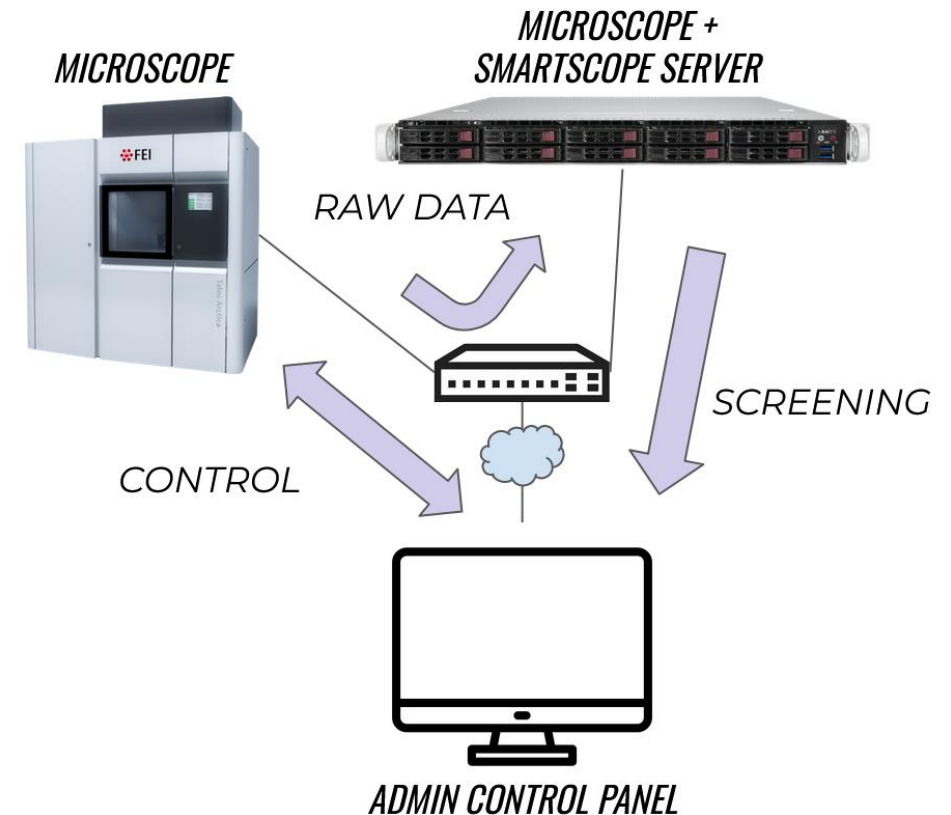
ONEDATA



Back to the CryoEM use-case

Need to implement some automation

- Integration with LIMS (Laboratory Notebook)
- Automatic data collection from instruments
 - Dataset discovery service
- Load metadata
- Data handover to user
- Transfer to central/archive storage and HPC
- Evict data from facility of data origin



Onedata access information stored by dataset

- Automatically inserted to specified metadata file

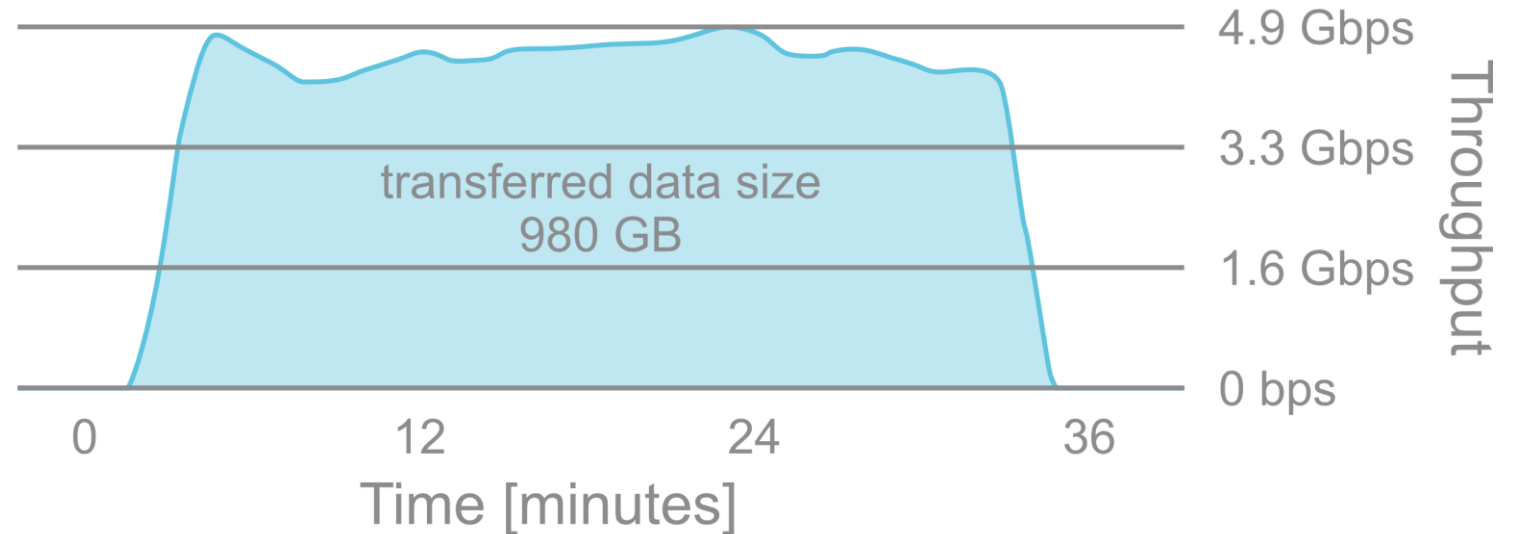
```
1 onedata:  
2 | onezone: https://datahub.egi.eu  
3 | spaceId: c2956f8d21fffd7bcbb628b382f0c17bch2e97  
4 | inviteToken: MDAxY2xvY2F0aW9uIGRhGFodWIuZWdpLmV1CjAwOT...  
5 | publicUrl: https://datahub.egi.eu/share/885c806b0c94730195fe65e...
```

Source: TENA I. Mikel, SVOBODA T. et al, IBERGRID 2023



Data transfers performance

- Two types
 - Between datacenters
 - Up to 4.9 Gbps
 - Datacenter \Leftrightarrow user
 - Depends
- Users can access by
 - Oneclient (FUSE) app (~1.2 Gbps)
 - REST API (~2.5 Gbps)
 - PythonFS (~1.7 Gbps)
- Larger files \Rightarrow larger throughput



Onedata at CESNET

Infrastructure service

- Managed Onedata service
- Optional access interface to CESNET object storage (S3)
- Setup Onedata component in a lab

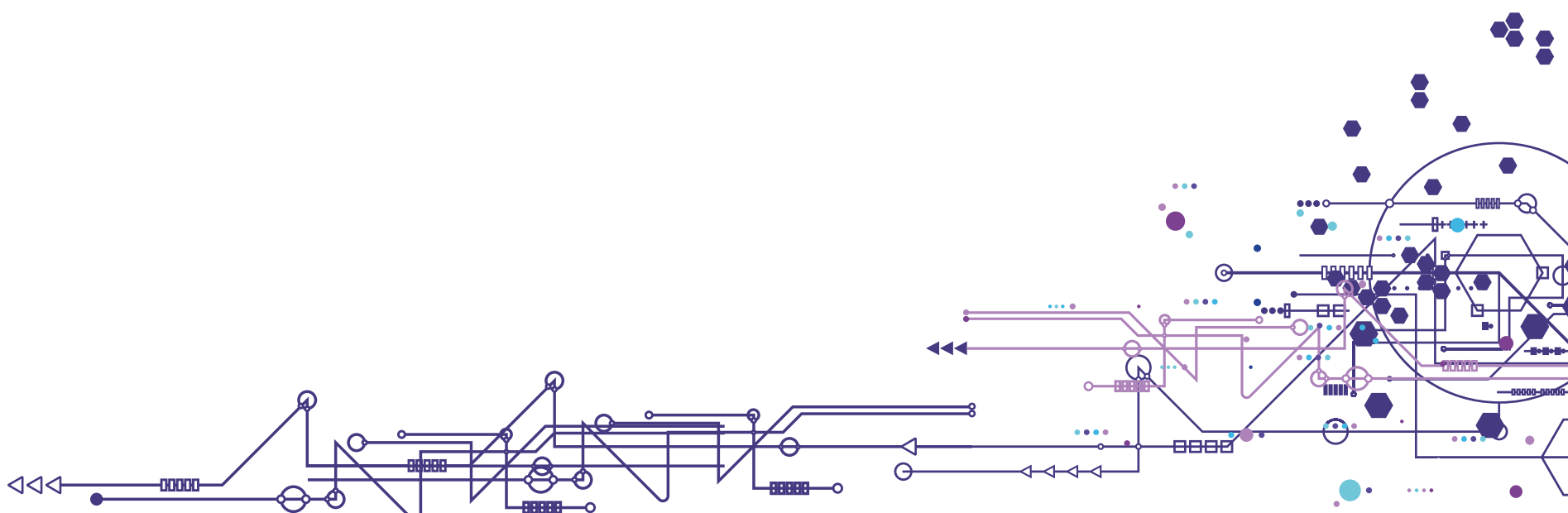
onedata.e-infra.cz





iRODS

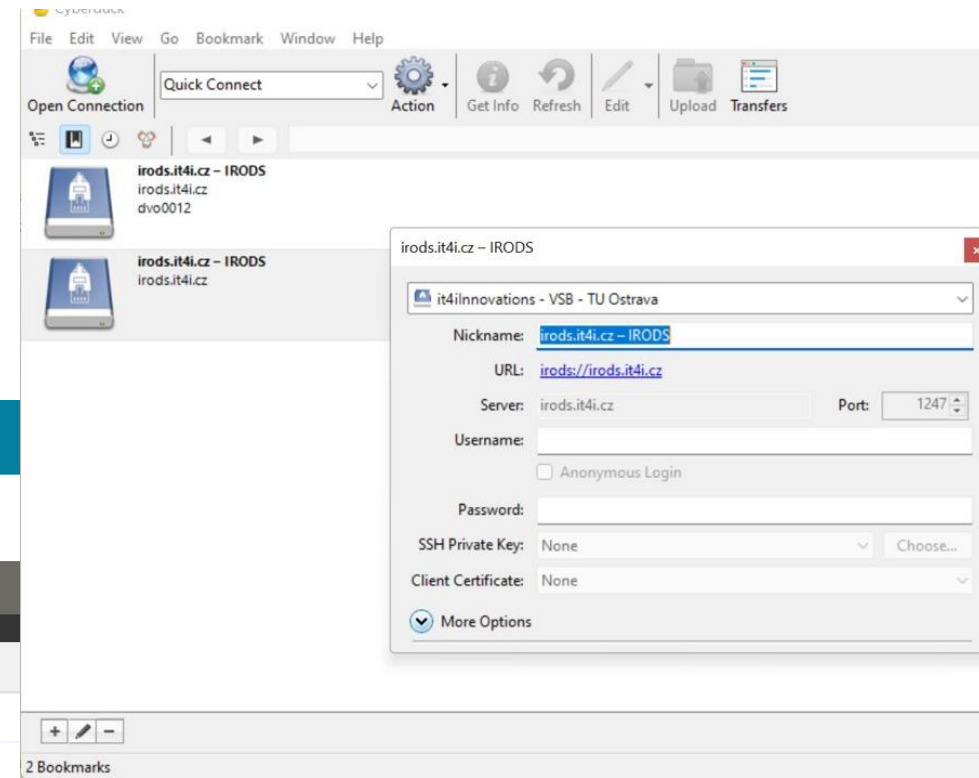
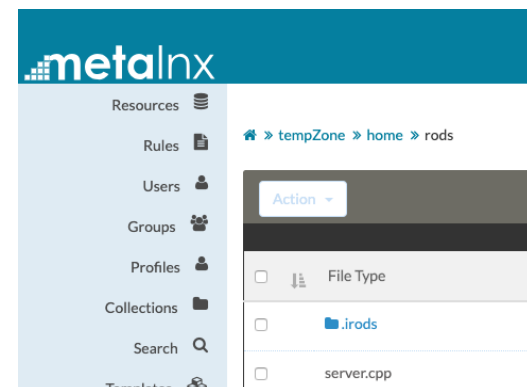
Ondřej Filip



iRODS

- Data management software again?
 - BSD 3-Clause License
 - governed by iRODS consortium
 - IT4I is iRODS member since Feb 2024
- VFS with single namespace in front of existing storage solutions
- Workflow Automation
 - extensive set of policies and rules
- Data discovery
- Collaboration and data sharing
- Resource federation among organisations
- Access
 - Desktop **Cyberduck** (Mac, Win)
 - CLI **icommands, irodsfs**
 - Web **metalnx**
 - APIs **S3, REST, python**

iRODS



iRODS @ IT4Innovations

IT4I clusters



Internet



irods.it4i.cz

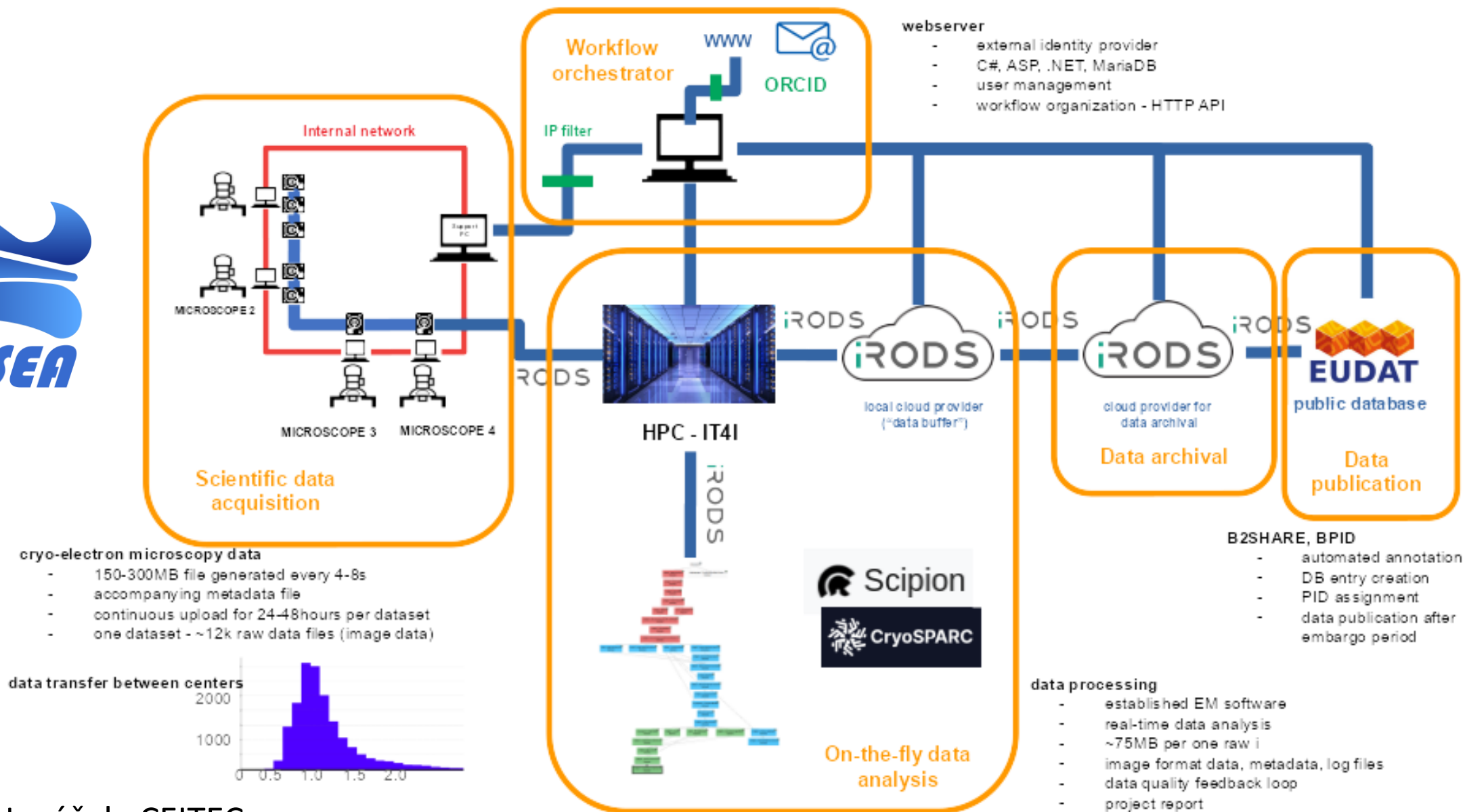


NFS storage

- Initially deployed for B2SAFE
 - data archiving service
 - software by EUDAT
- VM based deployment
 - HAProxied front-end
 - Vendor storage back-end
- Up to 2TB



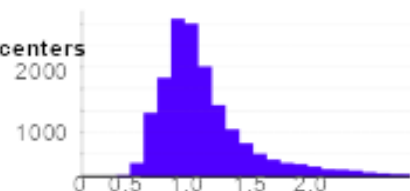
iRODS @ IT4Innovations



cryo-electron microscopy data

- 150-300MB file generated every 4-8s
- accompanying metadata file
- continuous upload for 24-48hours per dataset
- one dataset - ~12k raw data files (image data)

data transfer between centers



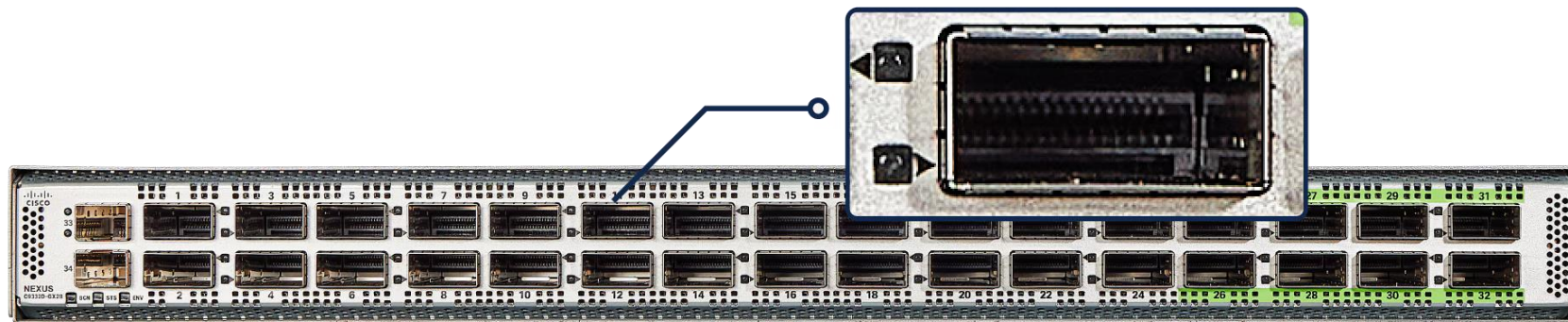
data processing

- established EM software
- real-time data analysis
- ~75MB per one raw i
- image format data, metadata, log files
- data quality feedback loop
- project report

IT4I data transfer

Problem introduction

- We suggested to our users to use LUMI cluster because our queues were full and wait times were long
- Complaint in RT -> SSH transfer to LUMI cluster is slow!
- But, but... We have nice and shiny 100G+ links and devices that surely cannot be?
- Let's investigate



IT4I data transfer

User perspective

- Users are not racing packets
- Perception of completion time
- Speed of light in fiber cable is given
- Yet with 8–20 MB/s ~ 64–160Mbit/s there is room for improvement



IT4I data transfer

Traditional way

- SCP/SFTP single connection
- TCP based
 - How much data can be in transit before acknowledgment is controlled by congestion window
 - Maximum congestion window is limited by buffer
 - Larger network delay requires larger buffers
- Hard to tune in remote HPC center or company



IT4I data transfer

Buffers first

- Linux sysctl madness
 - net.core.*mem
 - tcp.ipv*mem
- We adjusted our TCP buffers according to various docu sources
- Maxed out local iperf3 tests and tests towards CESNET DU in Brno
- We got 32MB/s ~ 256 Mbit/s
- Some improvement, less fluctuation



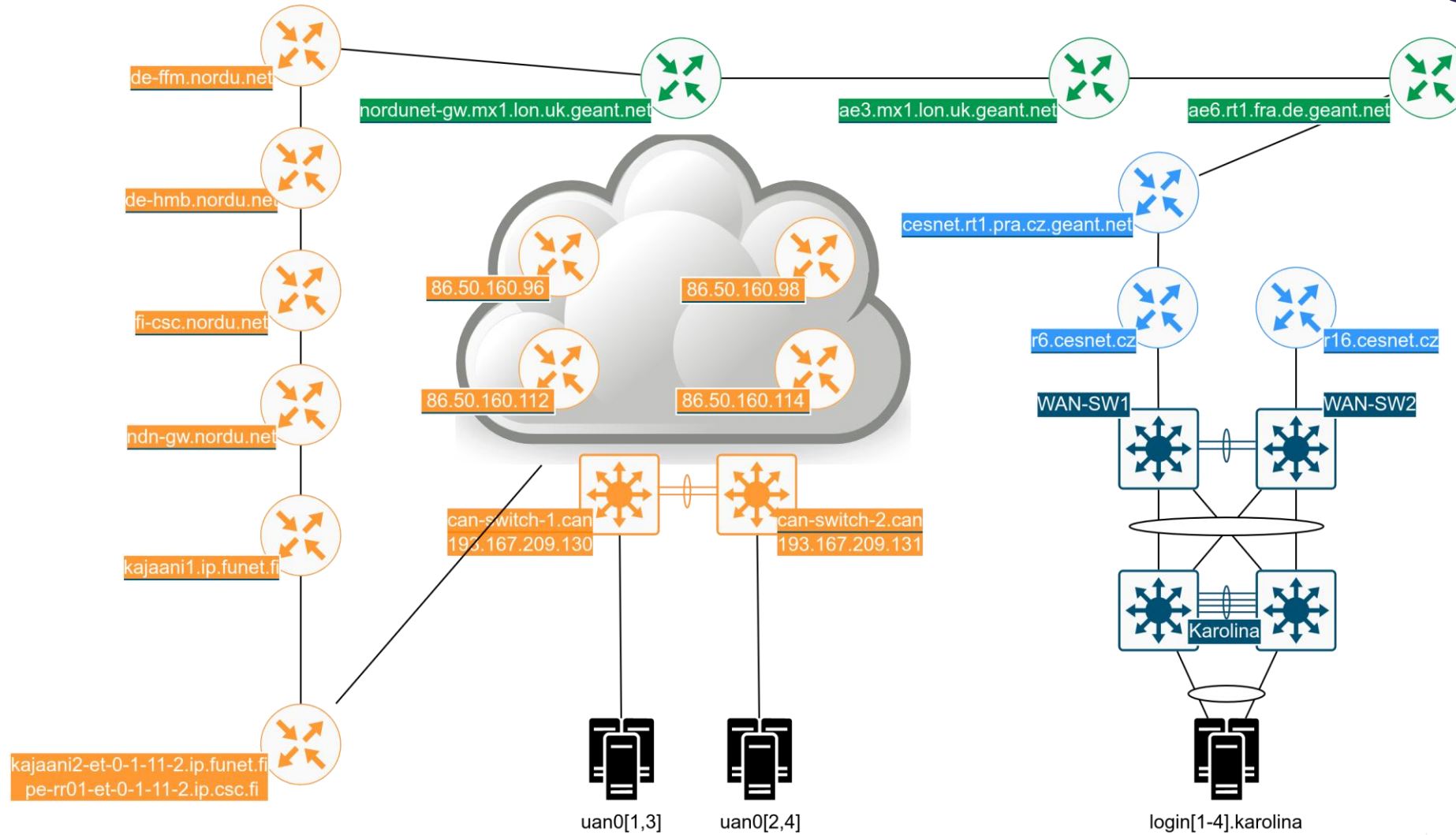
IT4I data transfer

Path next or which way to the cluster?

- Let's use venerable network tool
- login3.karolina # traceroute -I lumi.csc.fi
traceroute to lumi.csc.fi (193.167.209.163), 30 hops max, 60 byte packets
1 gw.karolina.it4i.cz (195.113.175.65) 0.436 ms 0.604 ms 0.760 ms
...
5 cesnet.rt1.pra.cz.geant.net (62.40.124.29) 5.675 ms 5.682 ms 5.771 ms
6 ae6.rt1.fra.de.geant.net (62.40.98.158) 12.684 ms 12.508 ms 12.497 ms
7 ae3.mx1.lon.uk.geant.net (62.40.98.179) 23.562 ms 23.350 ms 23.331 ms
8 nordunet-gw.mx1.lon.uk.geant.net (62.40.124.130) 23.444 ms 23.458 ms 23.455 ms
9 de-ffm.nordu.net (109.105.97.79) 33.923 ms 33.842 ms 33.835 ms
10 de-hmb.nordu.net (109.105.97.104) 38.515 ms 38.496 ms 38.497 ms
...
13 kajaani1.ip.funet.fi (86.50.255.191) 62.821 ms 62.515 ms 62.520 ms
14 pe-rr01-et-0-1-11-2.ip.csc.fi (193.167.244.185) 62.619 ms 62.553 ms 62.605 ms
...
18 lumi-uan01.csc.fi (193.167.209.163) 66.318 ms 66.291 ms 66.282 ms



IT4I data transfer



IT4I data transfer

Long way to Kajaani

- Traffic routed via London and then back to Frankfurt
- RTT of approx. 60 ms
- Not good – compare to well known finnish mirror site of FUNET **nic.funet.fi** at 35 ms
- Contacted NORDUNET/CSC representatives
- Created ticket in CESNET RT and asked to contact GEANT to sort it out
- Wait...



IT4I data transfer

Finish line

- Ticket was created in September 2023, GEANT fixed the routing from IT4I in maintenance window in December, NORDUNET fixed path towards IT4I earlier
- 48 MB/s ~ 384 Mbit single connection transfer rate
- RTT +/- 43 ms
- We made it but at what cost?



IT4I data transfer

Key takeaways #3

- Client wants to transfer data in easy way so he needs tool which is ready to perform from the start
- Bandwidth is not enough for effective data transfers, check other variables
- Physical distance has its impact
- Many parties involved means overhead when dealing with *simple* issues



IT4I data transfer

Key takeaways #2



- TCP
 - is **not dead yet** for large data transfers
 - **single TCP stream** will not utilize your 100G link, **multiply**
 - alternatives using UDP are coming closer eg. **QUIC** protocol (SMB over QUIC and others), not aware of fully functioning client/server file server with any kind of maturity or usability
 - evaluate TCP **congestion control algorithm** (BBR vs CUBIC)
 - adjust your buffers
- Pure SSH transfers are still **enough in local** environment but not on WAN, be vary of algo selected as chacha20-poly1305 is outperformed by AES-GCM



IT4I data transfer

Key takeaways #3 aka dead-ends

- **hpn-ssh** or am I really going to depend on ssh fork with one maintainer
- **UDT** software based on the protocol is not maintained for +10 years; *Sector* app, *UDR* rsync wrapper
- **sftp** tuning message size and number of outstanding messages



iRODS @ IT4Innovations roadmap

- Implement as data ingest to IT4I HPC
 - Bare-metal deployment
 - Direct connection to storage
- Tests show iRODS will utilize with automatic parallelization
 - 100G link locally
 - 25G link Ostrava–Brno
 - We will be performing tests against LUMI with iRODS delivered through Apptainer
 - LUMI and Karolina under maintenance
- B2SAFE continuation



EUDAT Collaborative
Data Infrastructure



Thank you!

- Audience for being awesome!
- **Cesnet DU** granted IT4I physical server for testing
- IT4I
 - **Radek Janáček** network perf tests, sysctl tuning
 - **Ondřej Dvořák** implementing iRODS





info@e-infra.cz

A circular logo for e-infra.cz. It consists of a large, dark blue circle with a smaller, slightly offset circle inside it, creating a ring effect. The text 'e-infra.cz' is centered within the inner circle.

e-infra.cz