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

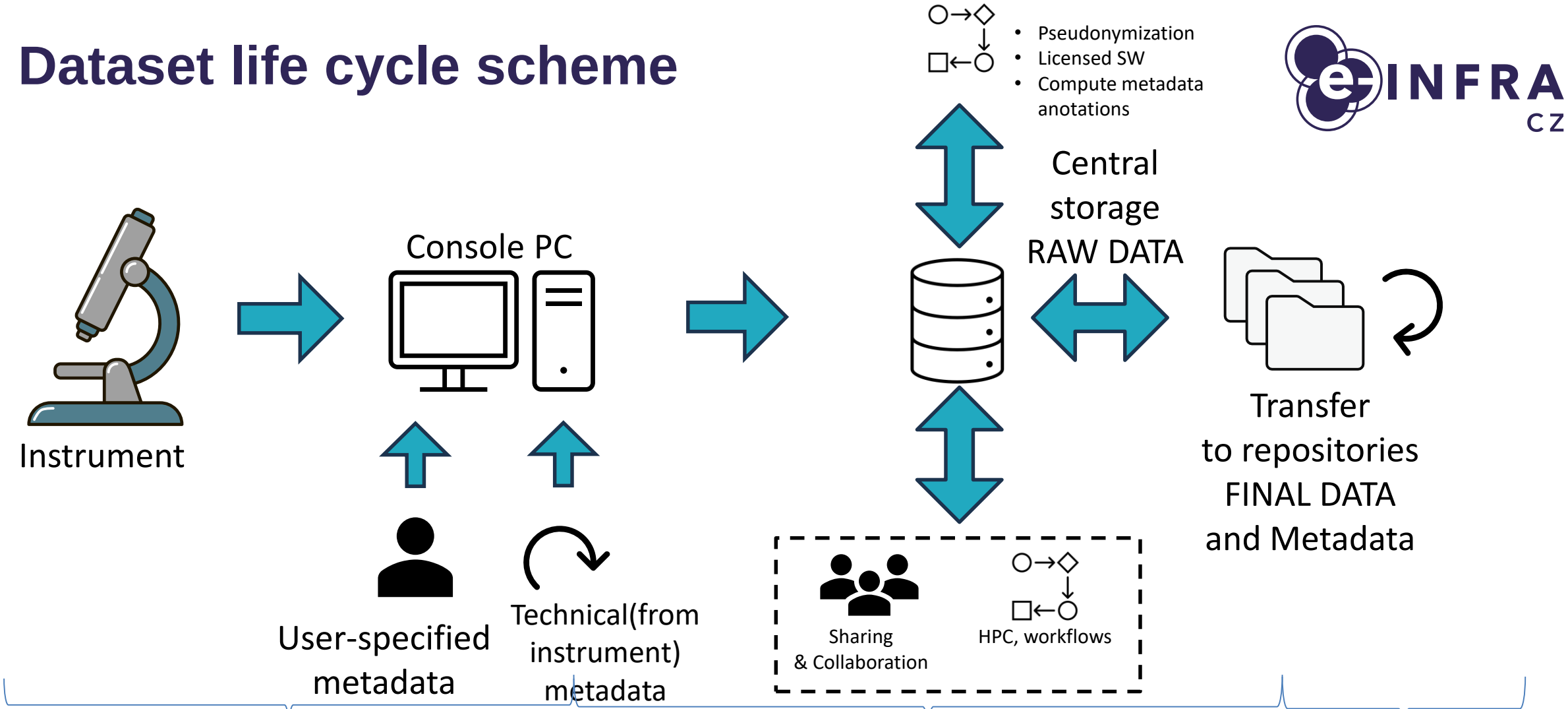
Motivation

- 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



Laboratory notebooks
Web interfaces

Data management systems

ONEDATA iRODS

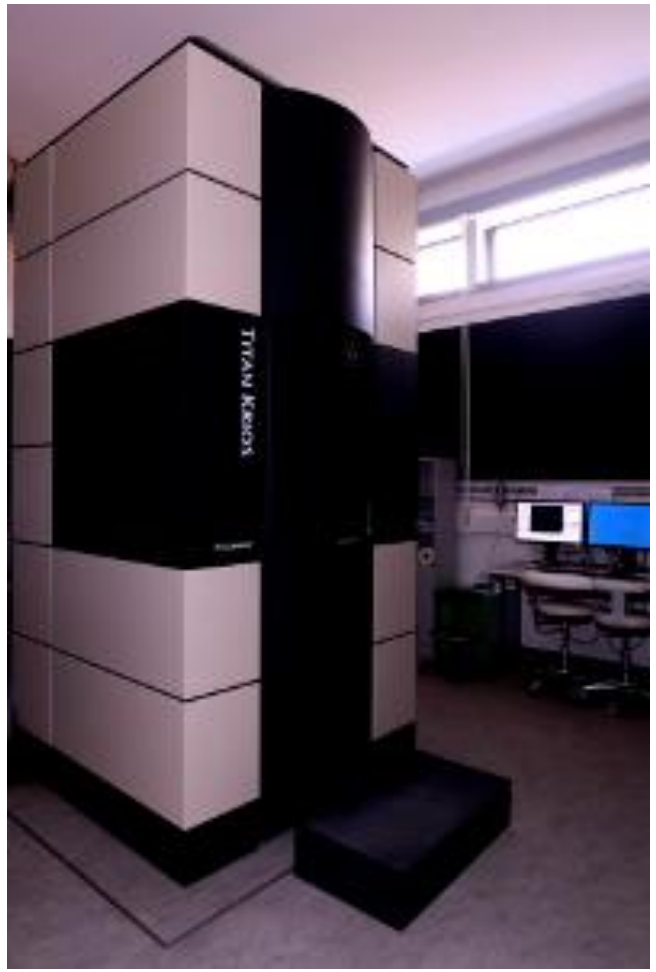
Repositories
Invenio
CLARIN Dspace
ARL

Onedata

Tomáš Svoboda

Motivation

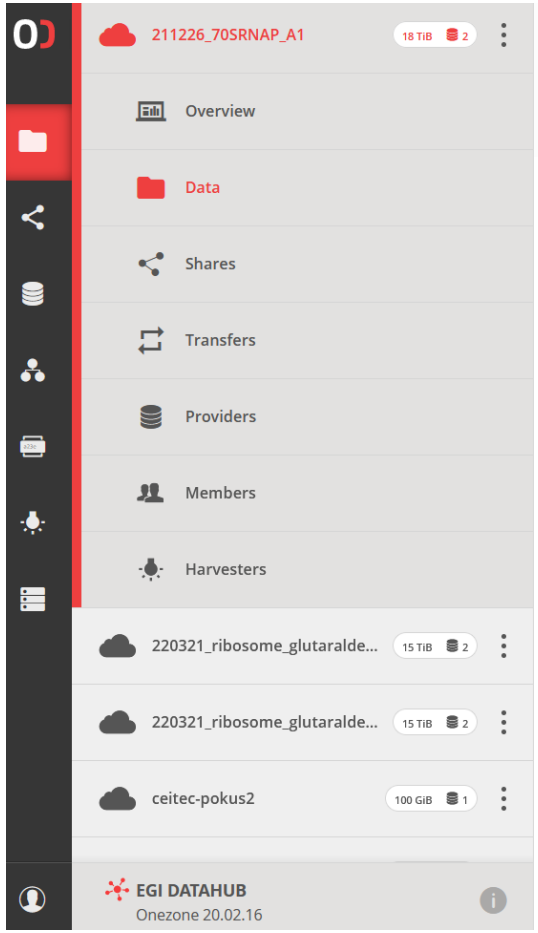
- Data management solution for two CryoEM facilities
  - CEITEC MU (Brno)
  - CNB CSIC (Madrid)
- Goals:
  - Handover data to users
  - Transfer data to HPC
  - Transfer data to archive storages
- Let’s try Onedata

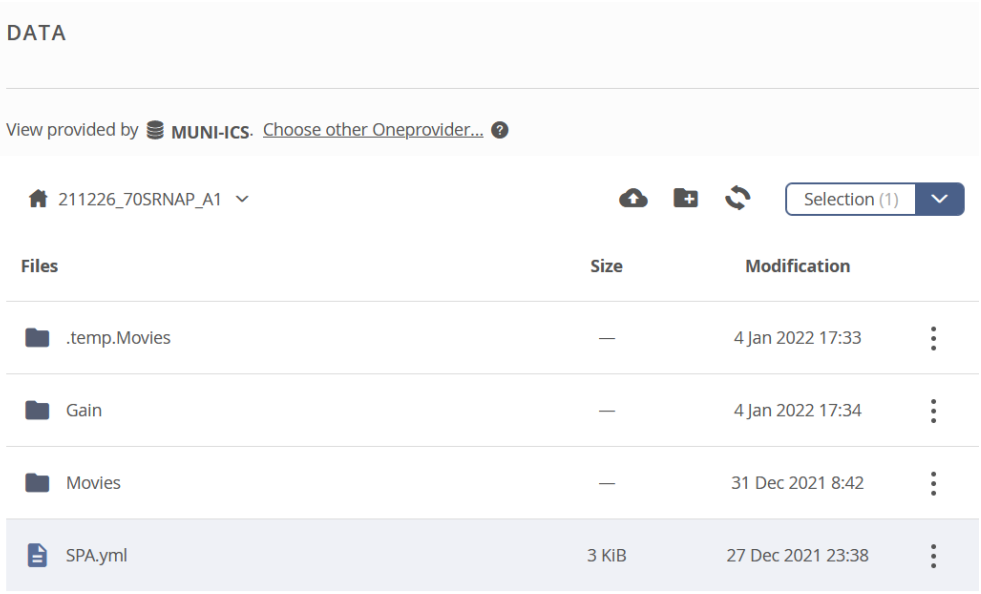


Onedata

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

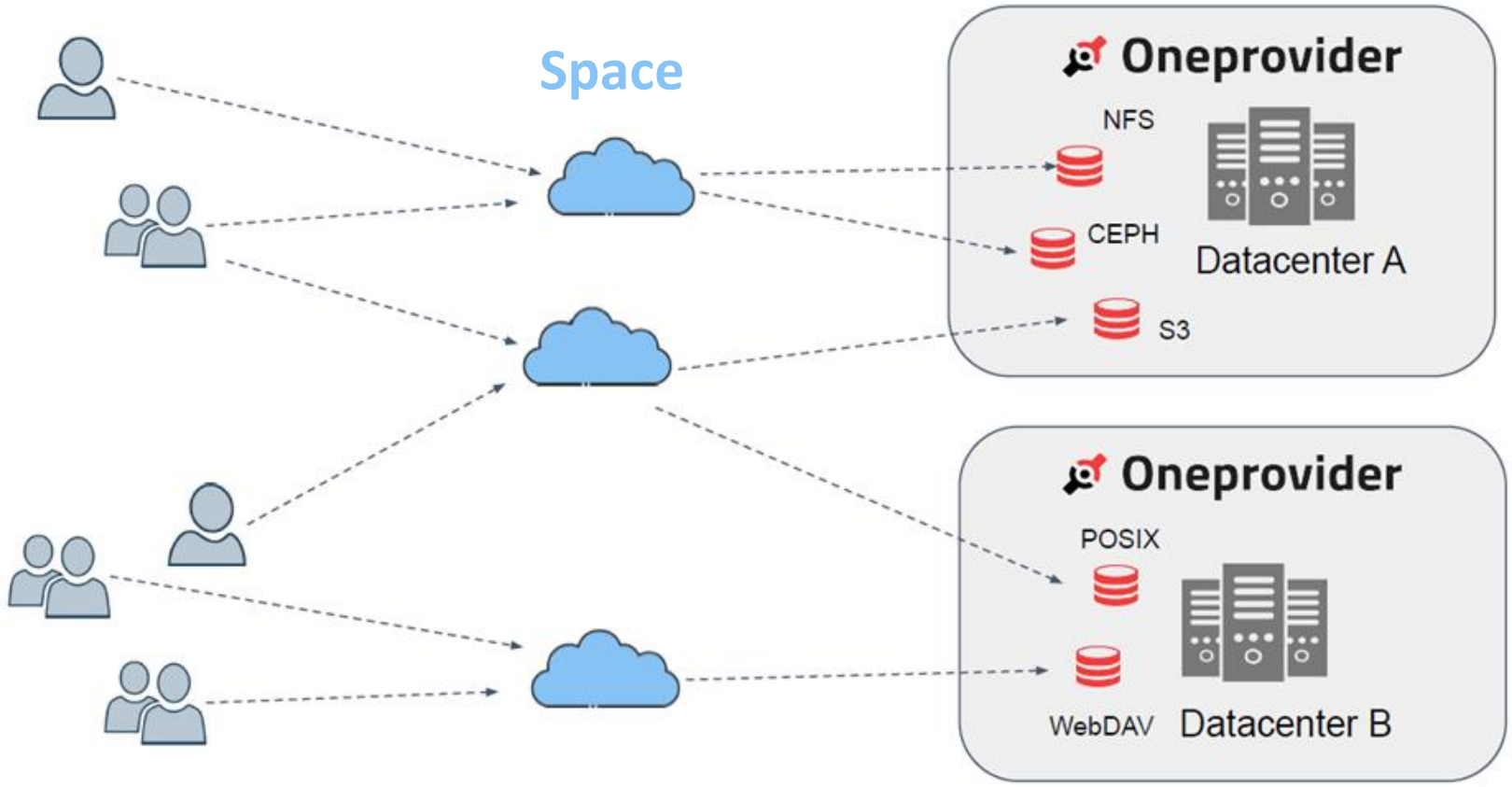




Onedata

The basic idea

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

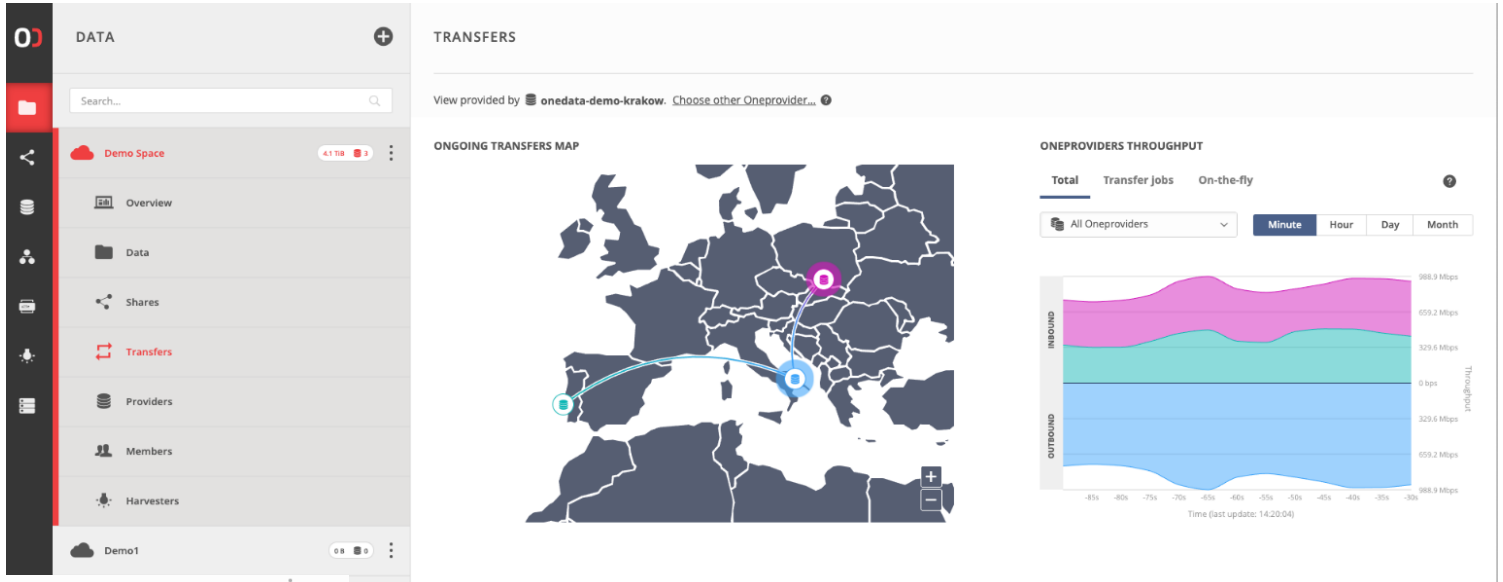


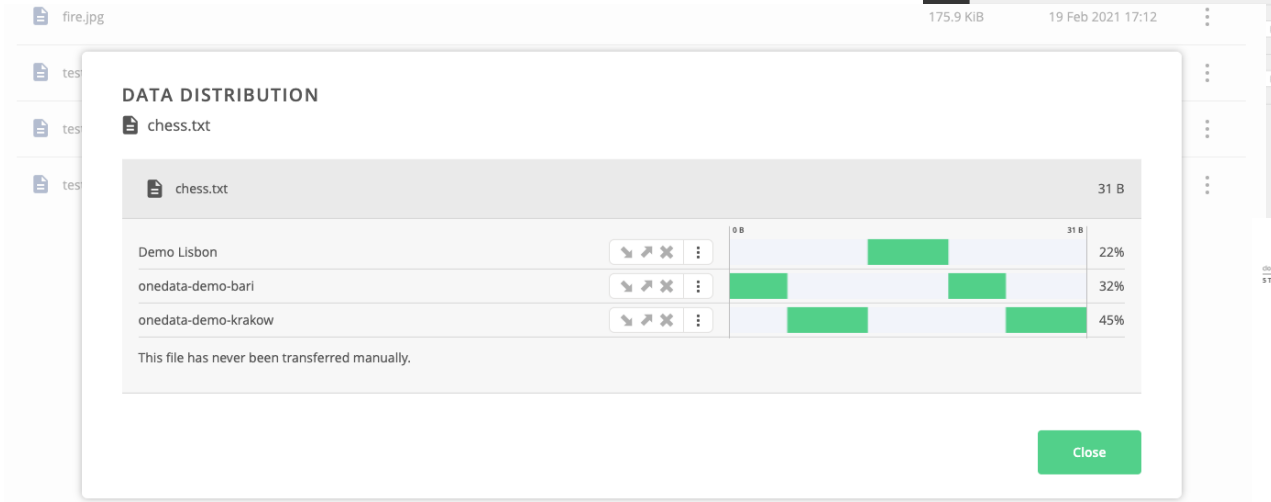
Source: OPIOLA, Lukasz. Onedata 101 presentation.

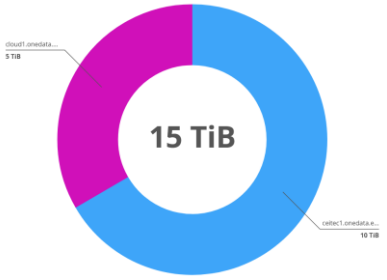
Onedata

Data distribution

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





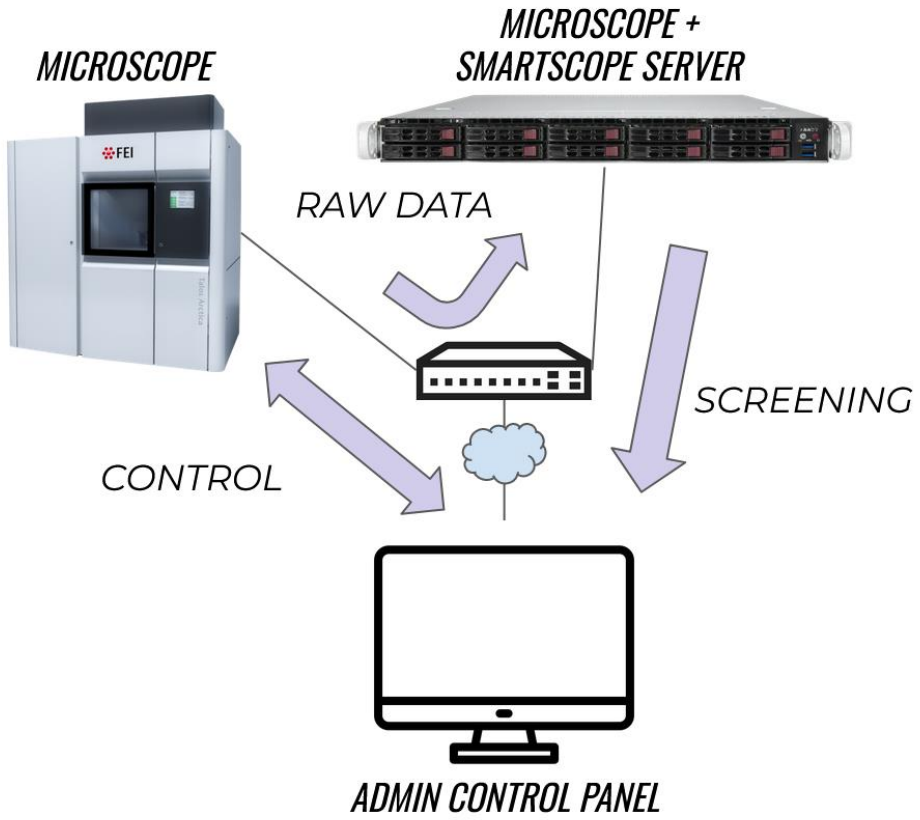


ONEDATA

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



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

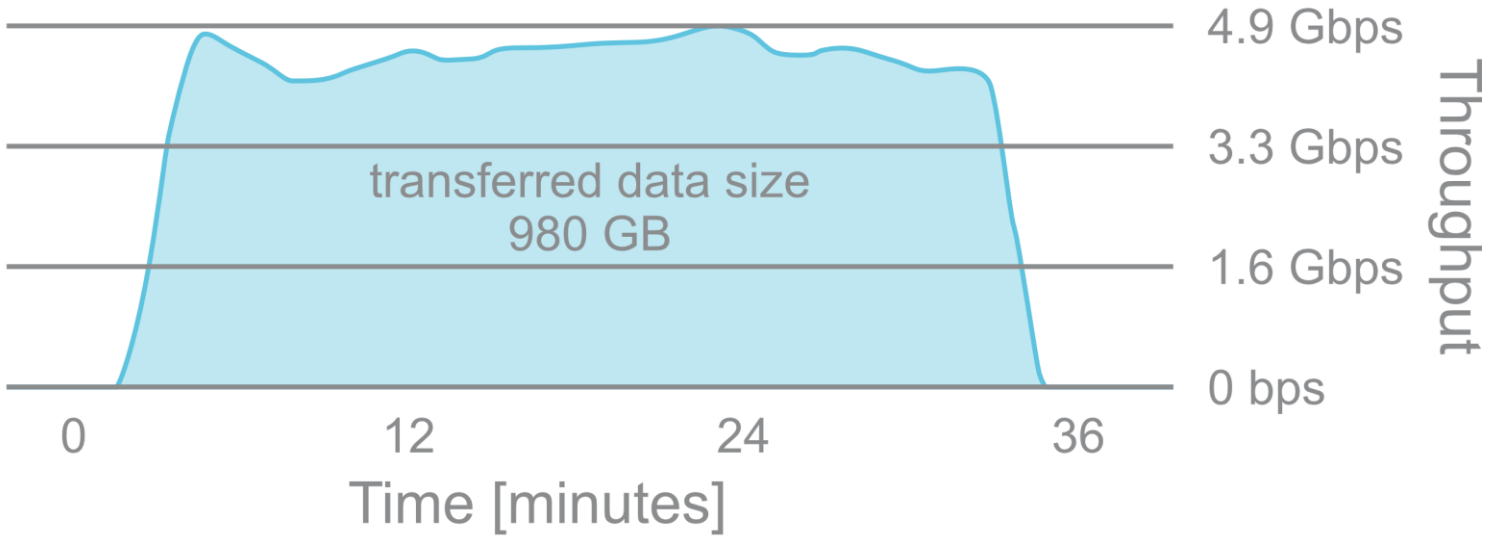
Onedata access information stored by dataset

- Automatically inserted to specified metadata file

```
1 onedata:
2   onezone: https://datahub.egi.eu
3   spaceId: c2956f8d21fffd7bcbb628b382f0c17bch2e97
4   inviteToken: MDAxY2xvY2F00aW9uIGRhdGFodWIuZWdpLmV1CjAwOT...
5   publicUrl: https://datahub.egi.eu/share/885c806b0c94730195fe65e...
```

Data transfers performance

- Two types
  - Between datacenters
    - Up to 4.9 Gbits
  - Datacenter <=> user
    - Depends
- Users can access by
  - Oneclient (FUSE) app (~1.2 Gbits)
  - REST API (~2.5 Gbits)
  - PythonFS (~1.7 Gbits)

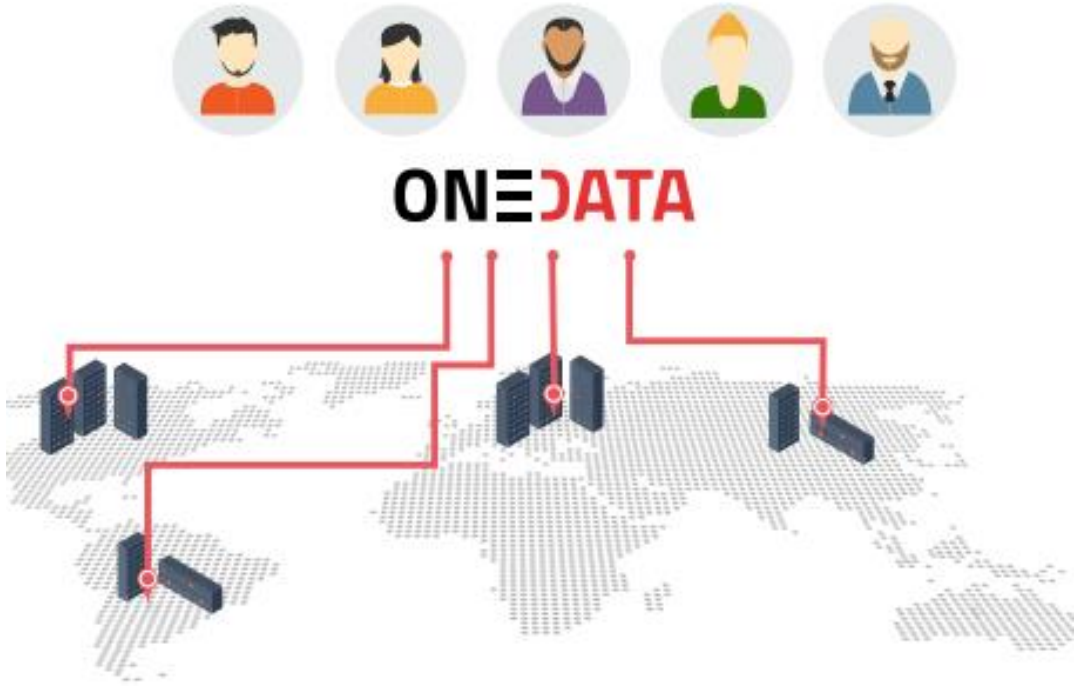


- Larger files => larger throughput

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

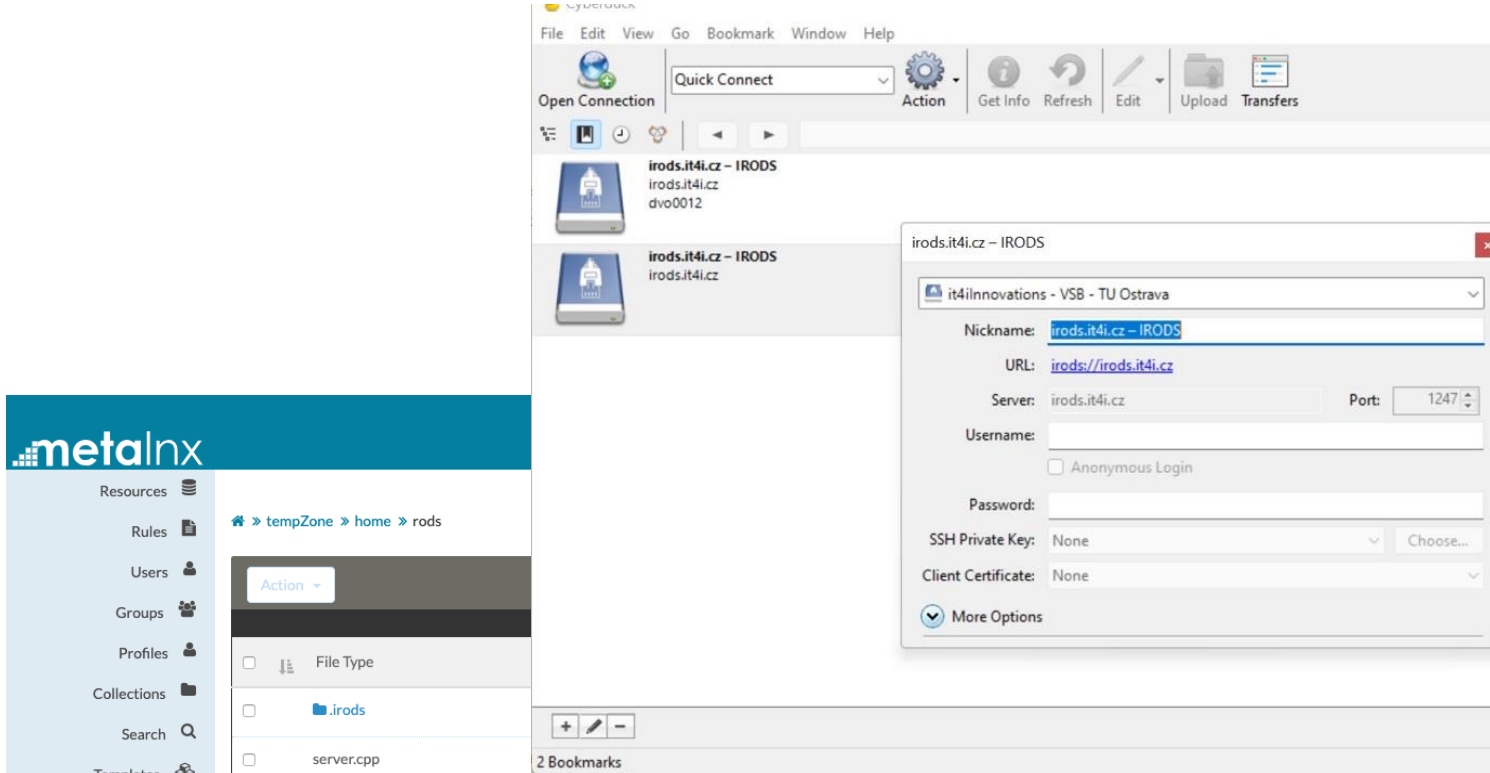
iRODS

- Data managment 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

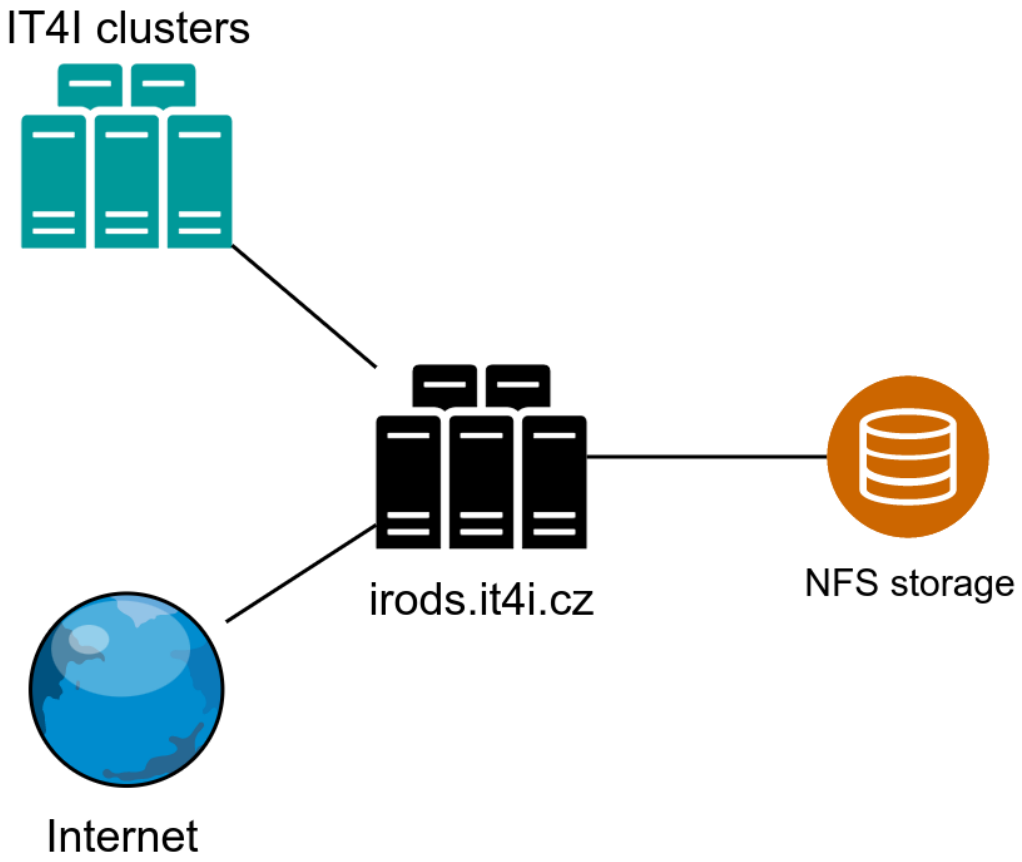
- Workflow Automation
  - extensive set of policies and rules

- Data discovery
- Collaboration and data sharing
- Resource federation among organisations

- Access
  - Desktop **Cyberduck** (Mac,Win)
  - CLI **icommands, irodsfs**
  - Web **metalnx**
  - APIs **S3, REST, python**



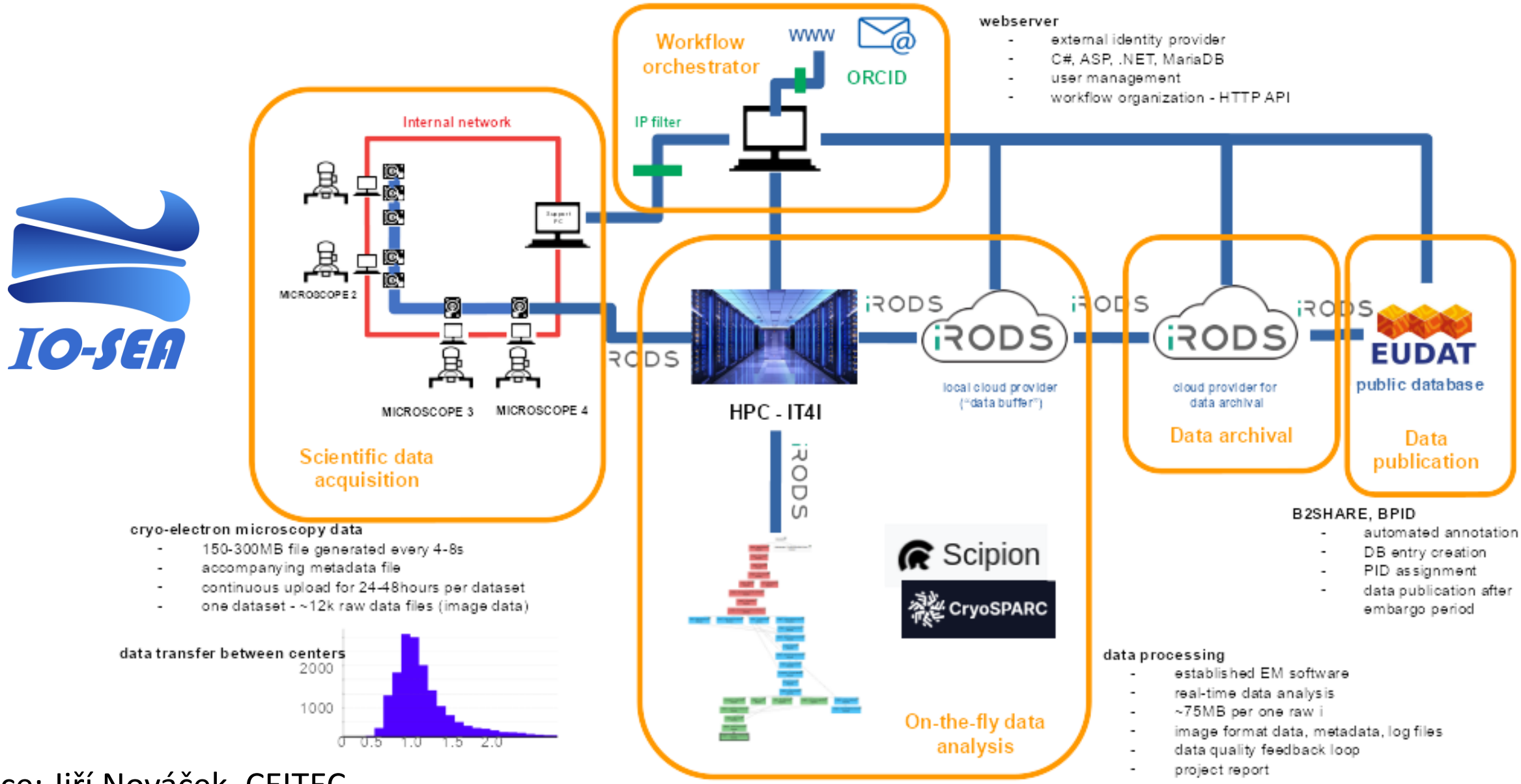
iRODS @ IT4Innovations



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

DICE
Data Infrastructure Capacity for EOSC

iRODS @ IT4Innovations

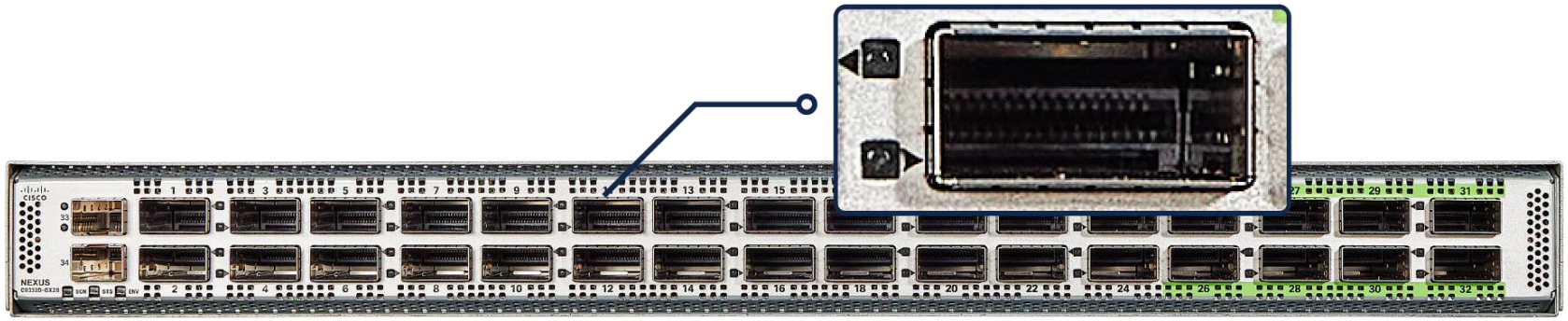


Source: Jiří Nováček, CEITEC

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 limitted 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

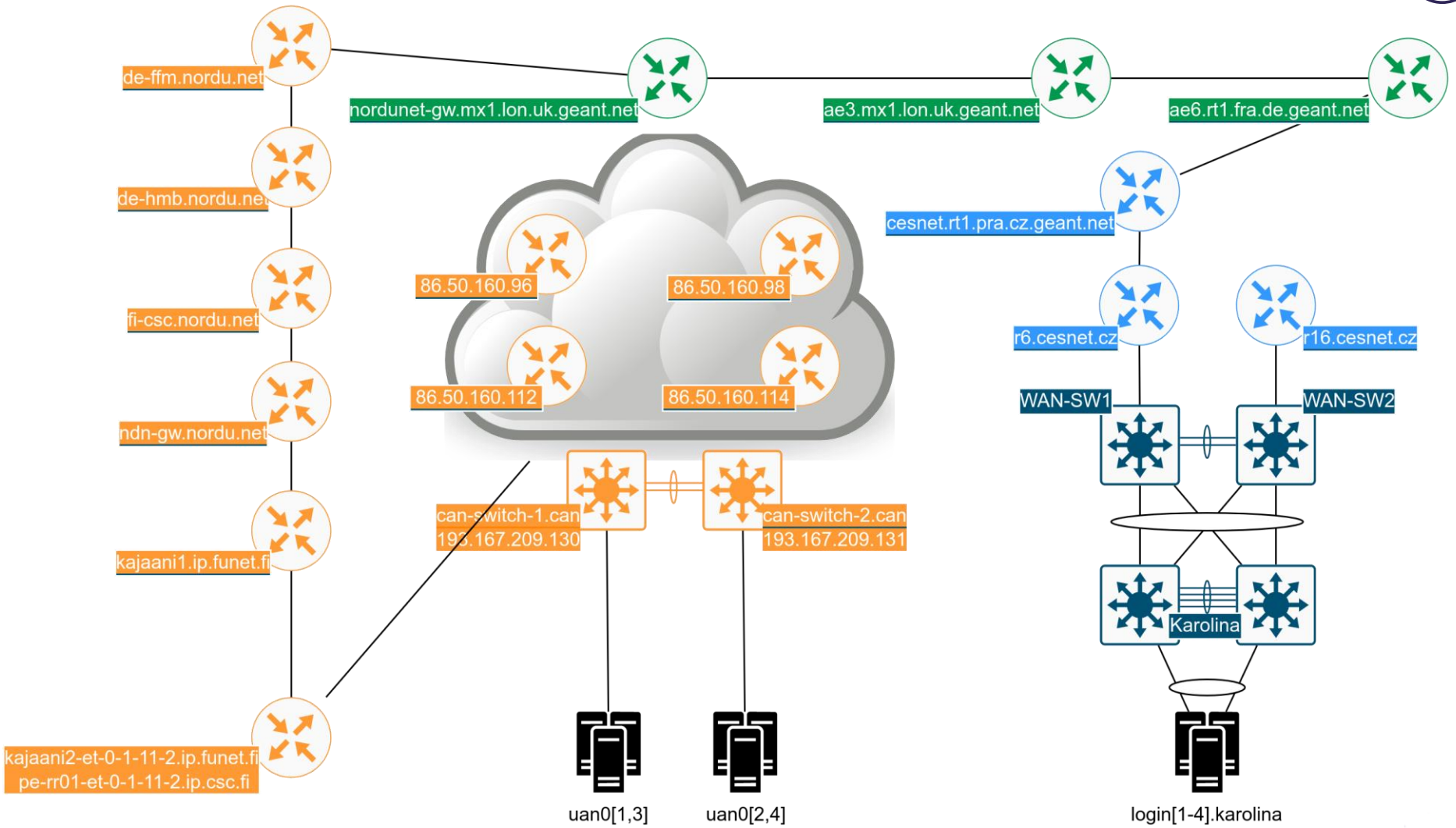
IT4I data transfer

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

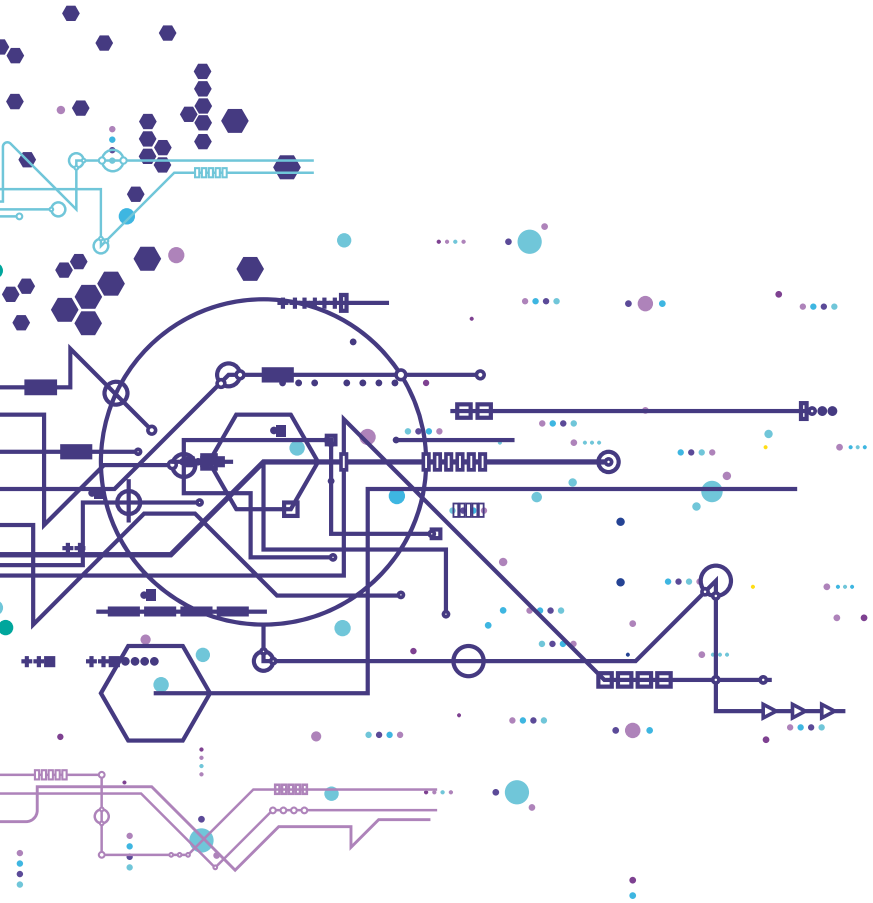
iRODS @ IT4Innovations roadmap

- 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

e-infra.cz