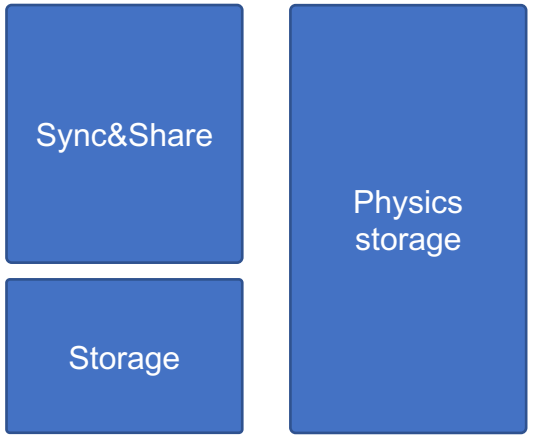
cern.ch/cernbox

CERNBox, the Scientific Cloud powered by EOS

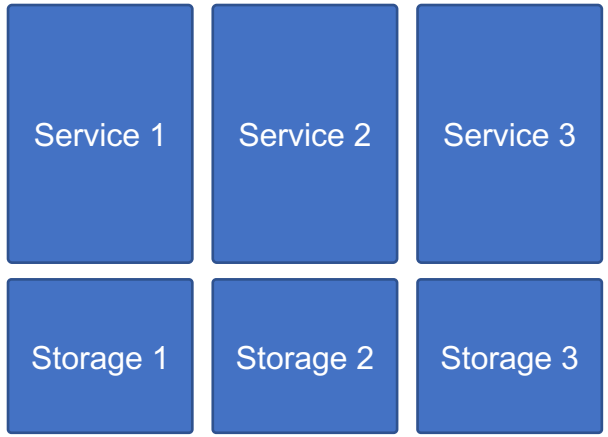
Diogo Castro
On behalf of the CERNBox team

27 April 2023
EOS 2023 Workshop

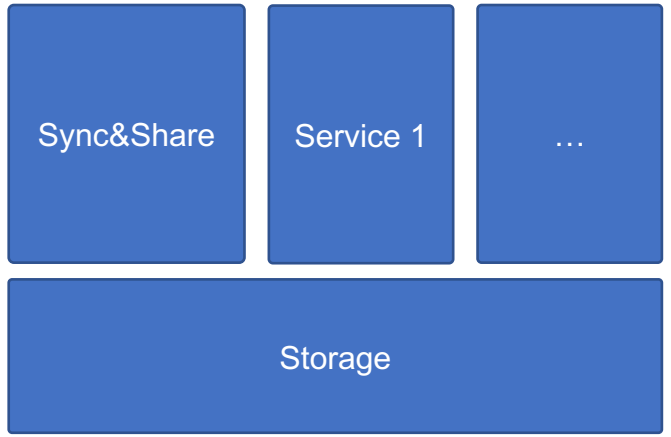
Introduction



Scenario 1



Scenario 2



Scenario 3

- In many sites, physics data and sync&share are split (or each service with its own storage backend)
- There is a trend and will to unify (and user feedback)

What is CERNBox?

Built with **open source** components

Turnkey access to data

+

+

Safely stored in CERN campus

CERN(Box) user community

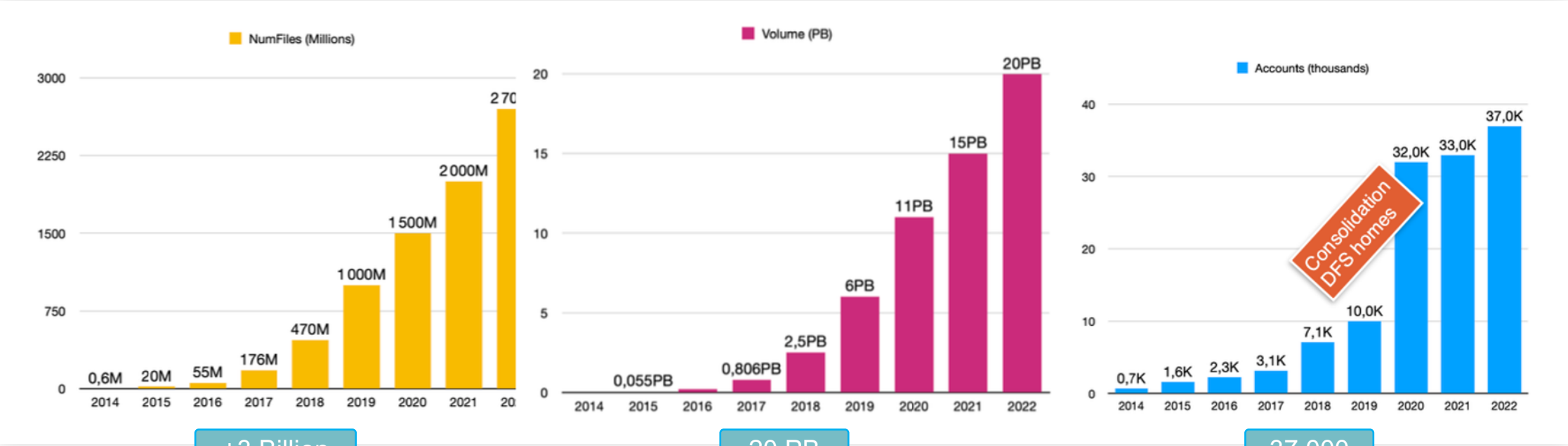
- Extremely diverse
  - Around **37K accounts**
  - Different work backgrounds
  - Different needs
- Different expectations on how the system should support their daily work

Growing since 2013

- +1000 Project areas
- 31 LHC and non-LHC experiments working areas

13k unique users/month

Skyrocketing usage

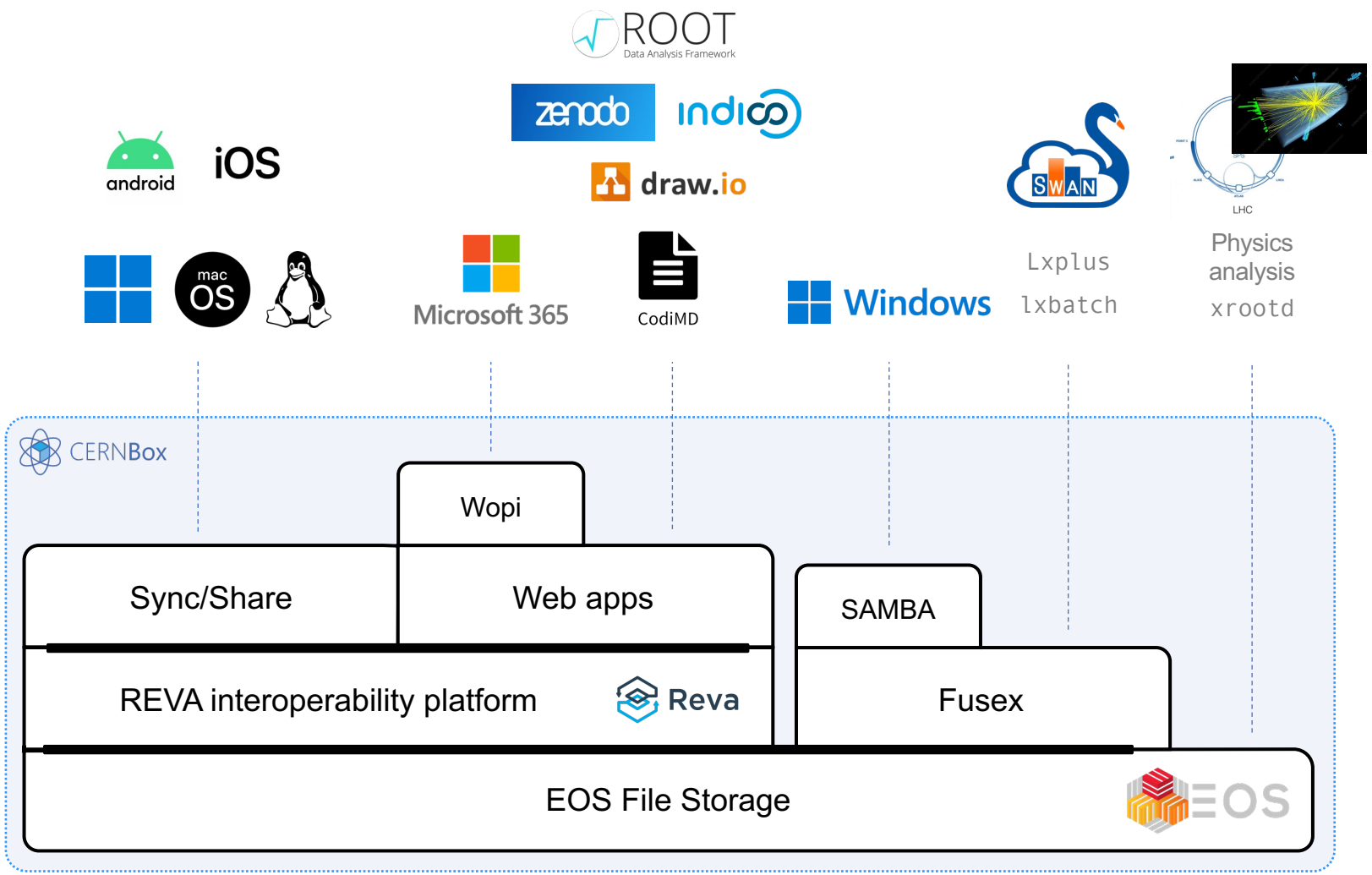


+3 Billion

20 PB

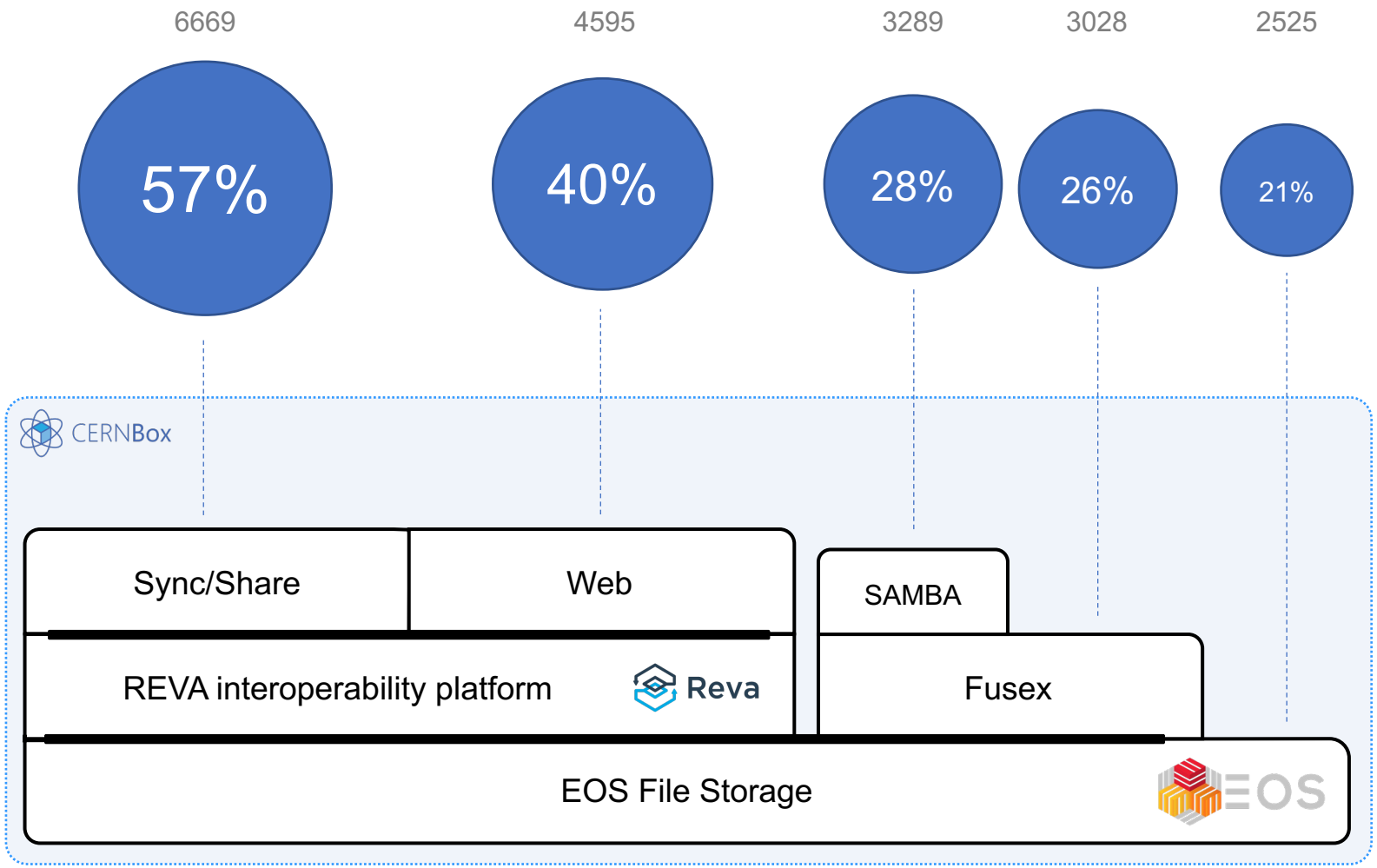
37 000

Global architecture



- 37k User accounts
- Sync clients for all major platforms (VFS support on Windows)
- 30k fuse clients
- 15 production clusters
- 18 PB general storage
- 400 PB Physics storage
- https://reva.link/

Users per protocol

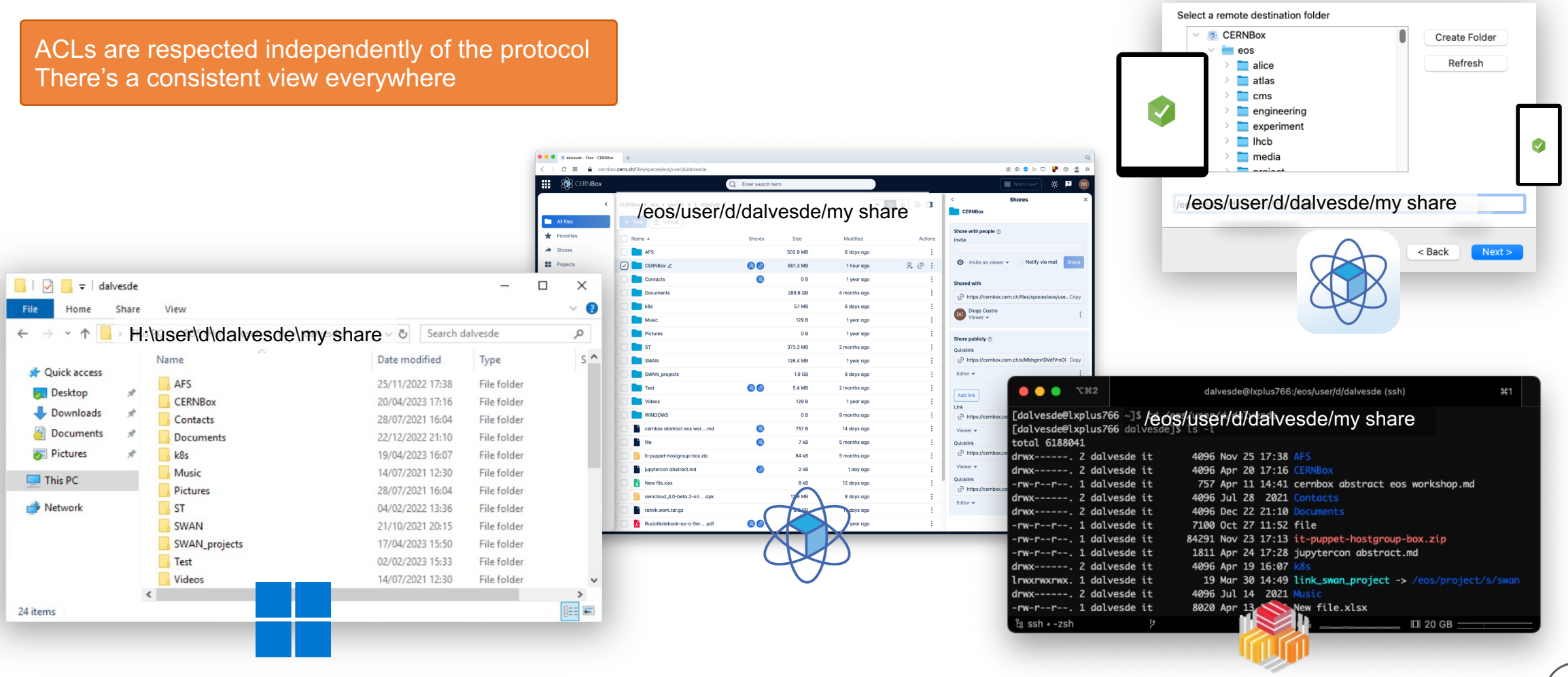


- Data from 1-14 April
- Unique users across all protocols: **11672**

- Although not the most representative, FUSE and XRoot have the most data transferred in volume

Sharing across protocol boundaries

ACLs are respected independently of the protocol
There's a consistent view everywhere

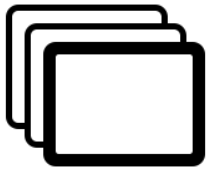


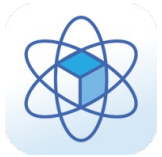
Features, apps and integrations

Features

Backups

Versioning





Sync client and
Mobile apps

Apps & Collaborative editing

Office 365

IFC.js

CodiMD

ROOT
Data Analysis Framework

jupyter

draw.io

Integrations

CERN

To infinite scale, and beyond

The new CERNBox version

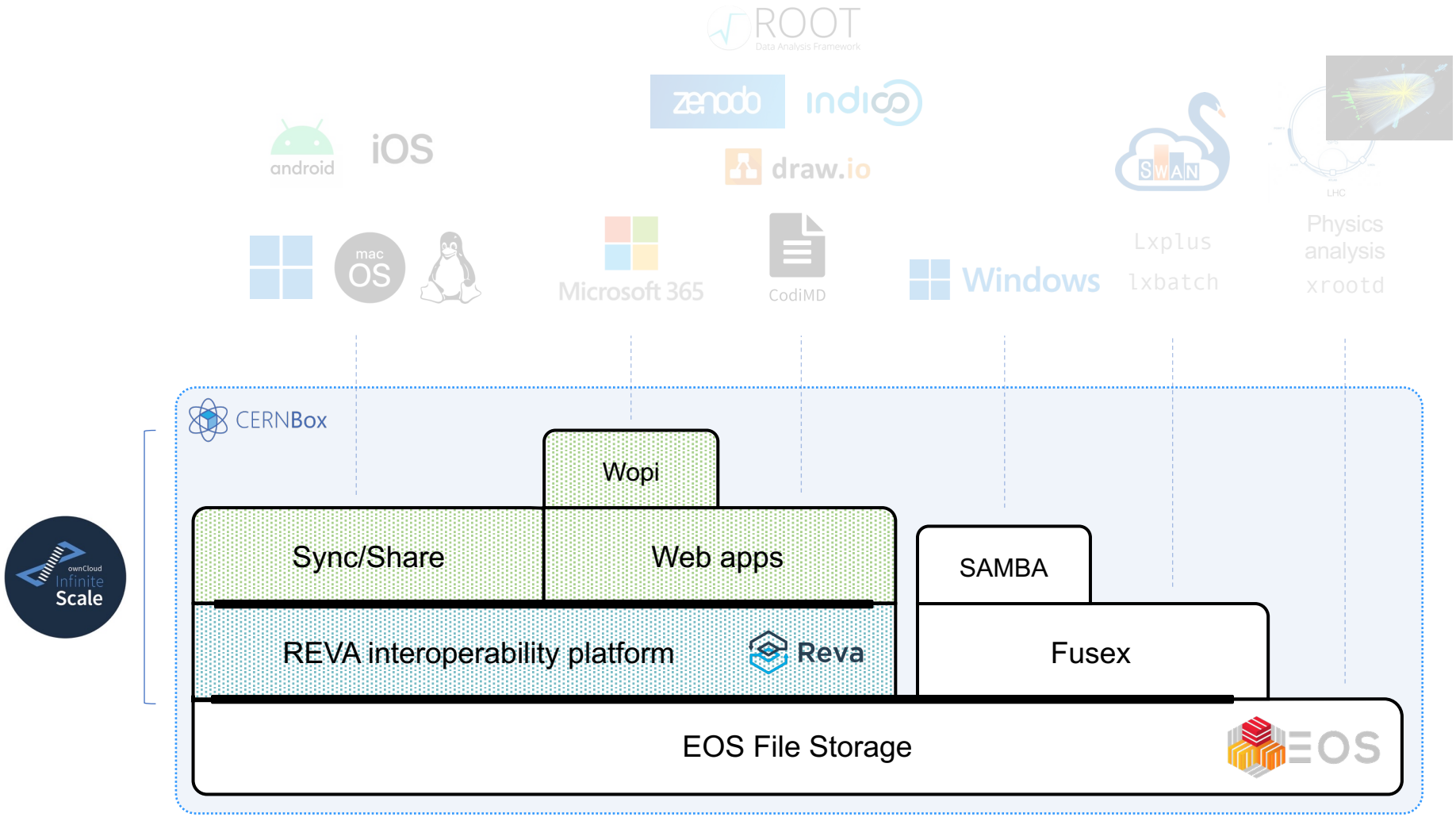
Challenges: from Sync&Share to Integrated Scientific Platform

- Direct Access to Underlying Storage
  - Assumption that OC is the single entry point
  - Caching might cause inconsistent views and conflicts
- Homogeneous Access to Shared Data
  - Need to sync ACLs to storage
- Data Ownership
  - Owner of the data is the user running the server (apache/www-data)
  - To increase security and give users full control of their data, the owner should be the real user
- Native File Versioning and native Recycle bin
  - Some storage backend provide these, so the OC layer duplicates the functionality
- Redundant Expensive Calls and inefficient design
  - Unique filenames for shares requires checking all existing shares before creating a new one

A new solution was required!

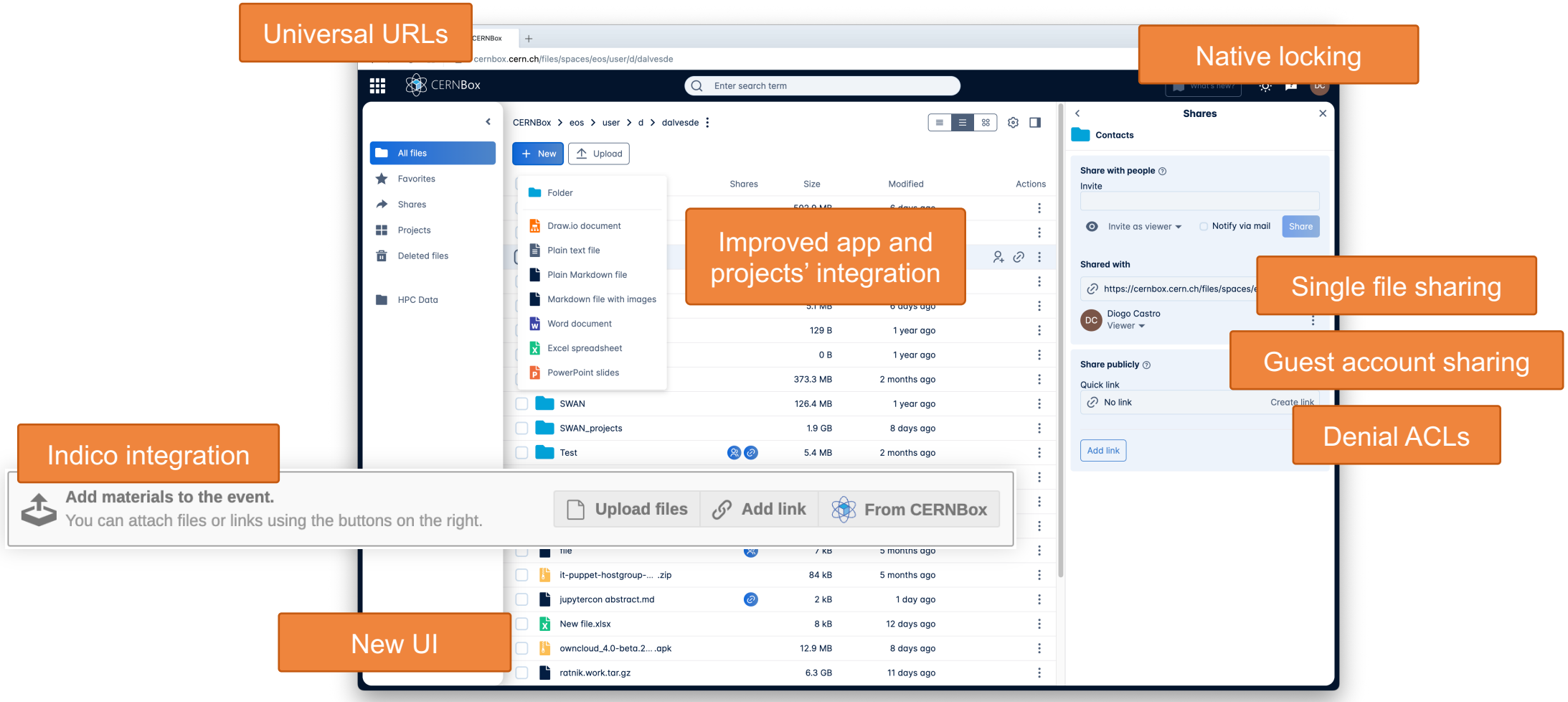
CERN

The new CERNBox



- New system was deployed to everyone in October 2022
- Based on REVA (https://reva.link/) and the new ownCloud Infinite Scale (https://owncloud.com/infinite-scale/)

New features



Ongoing and future work

New UI features

Self restore
Backup

Search

Notifications

Empowering CERNBox users with self-service restore functionality Gianmaria Del Monte
Modern notification system for Sync and Share using NATS Javier Ferrer, CS3 Conference 2023

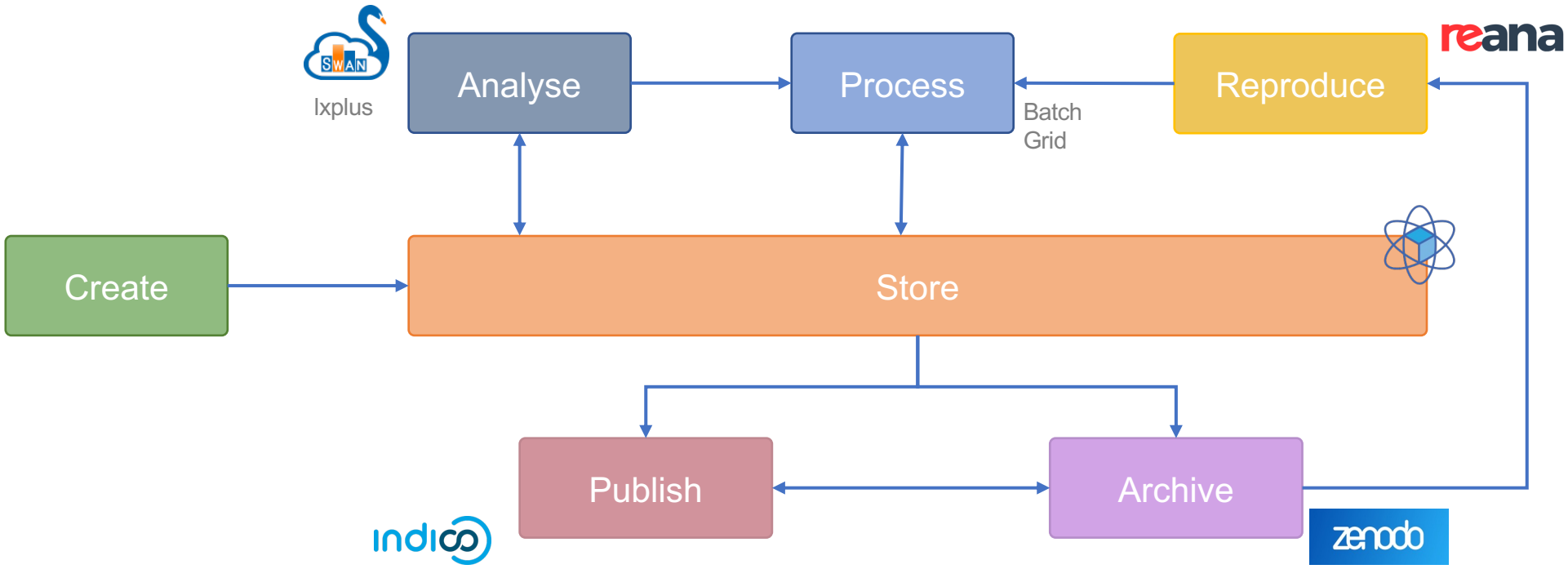
Improved external integrations

- Jupyter native integration with Reva (CS3APIs)
- Deployment in SWAN

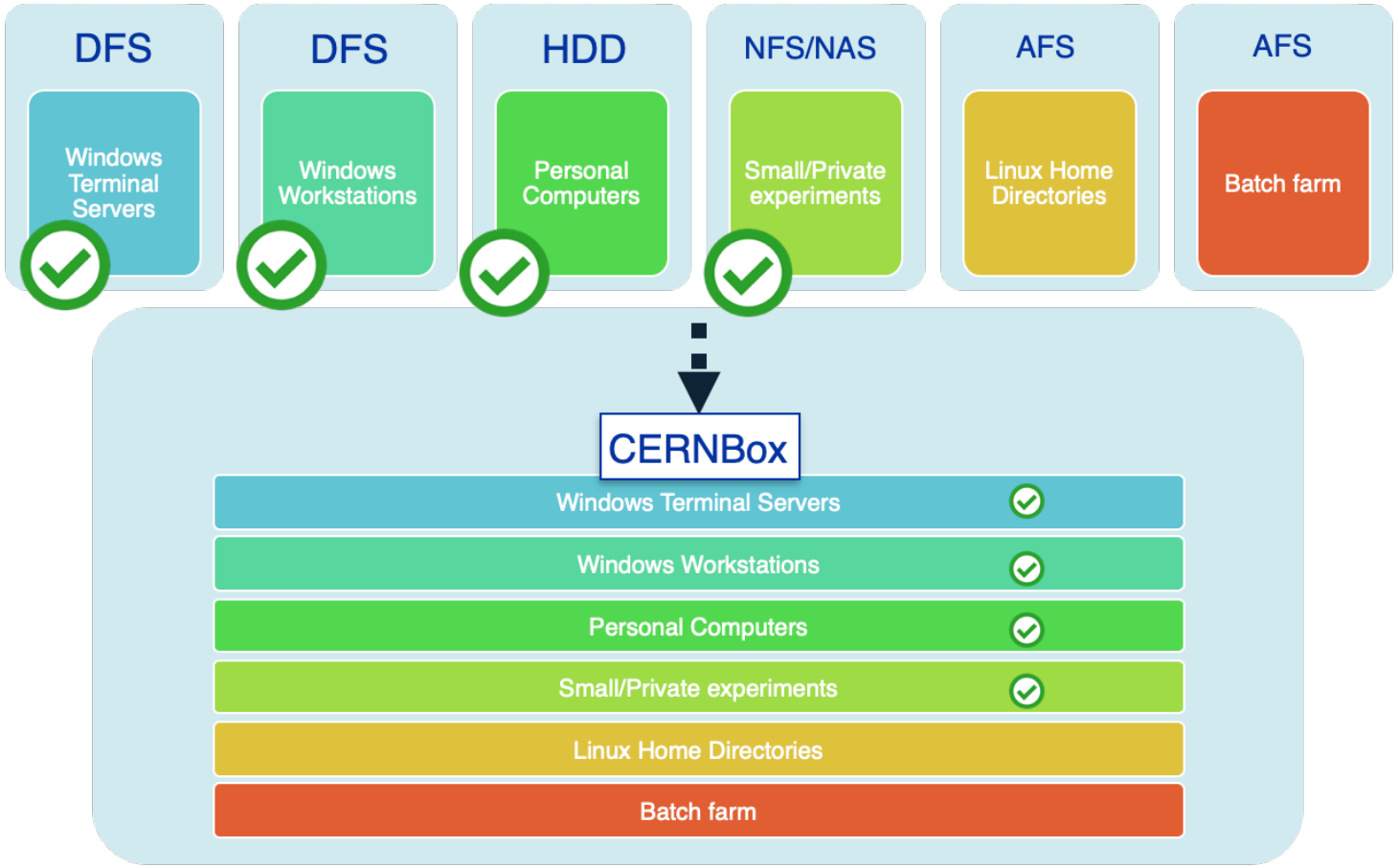
- Integration with FTS
- Native Support for HTTP Third-Party copy

Improved external integrations

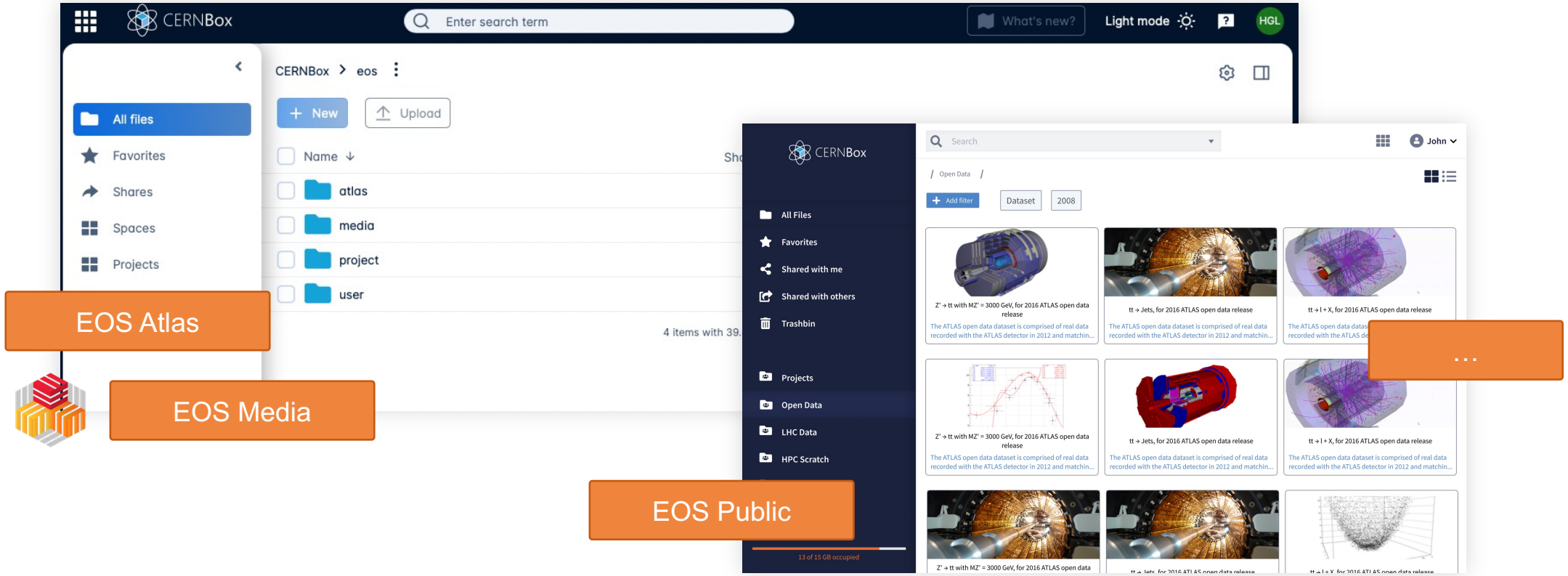
- Integration of ScieboRDS for preservation in a Digital Repository (Zenodo/OSF)
  - Supporting Research Data Lifecycle



Consolidation of services



CERNBox: federation of storages using REVA



ScienceMesh

- Federated + trusted network of Sync&Share services
  - 300K user base (CS3 community) across Europe (SURF, DESY, PIC, SUNET, ...)
  - Product of EC H2020 funded project CS3MESH4EOSC (ending in June 2023)
  - REVA provides access to the mesh

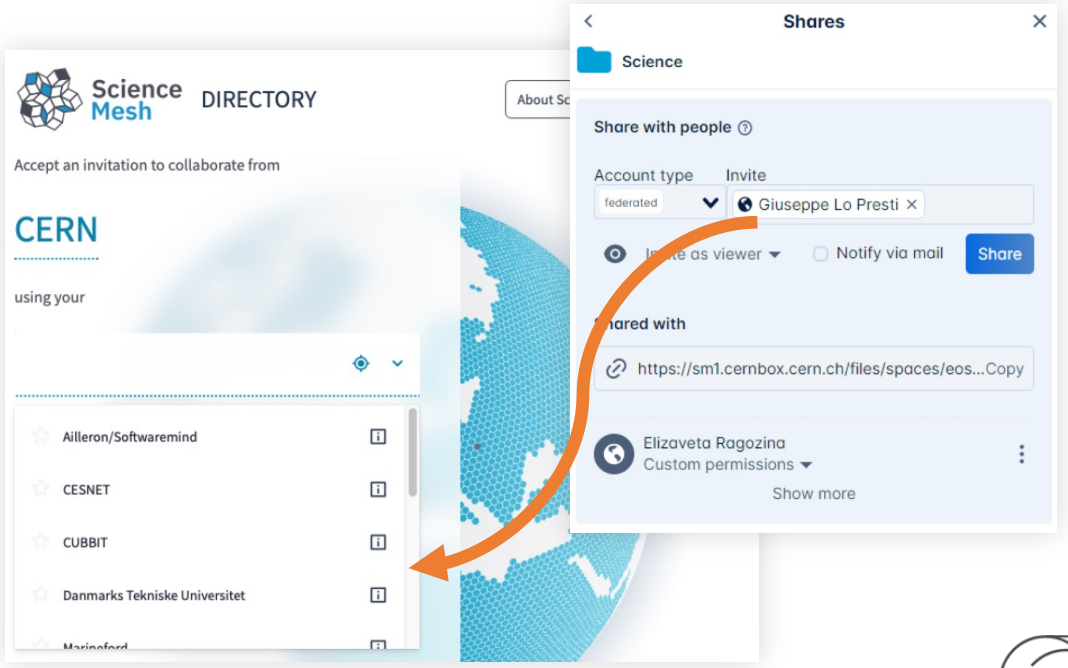
OPENCLOUDMESH

Federated sharing of data across different sync-and-share services.

Federated use of applications such as collaborative document editing, data archiving, and data publishing.

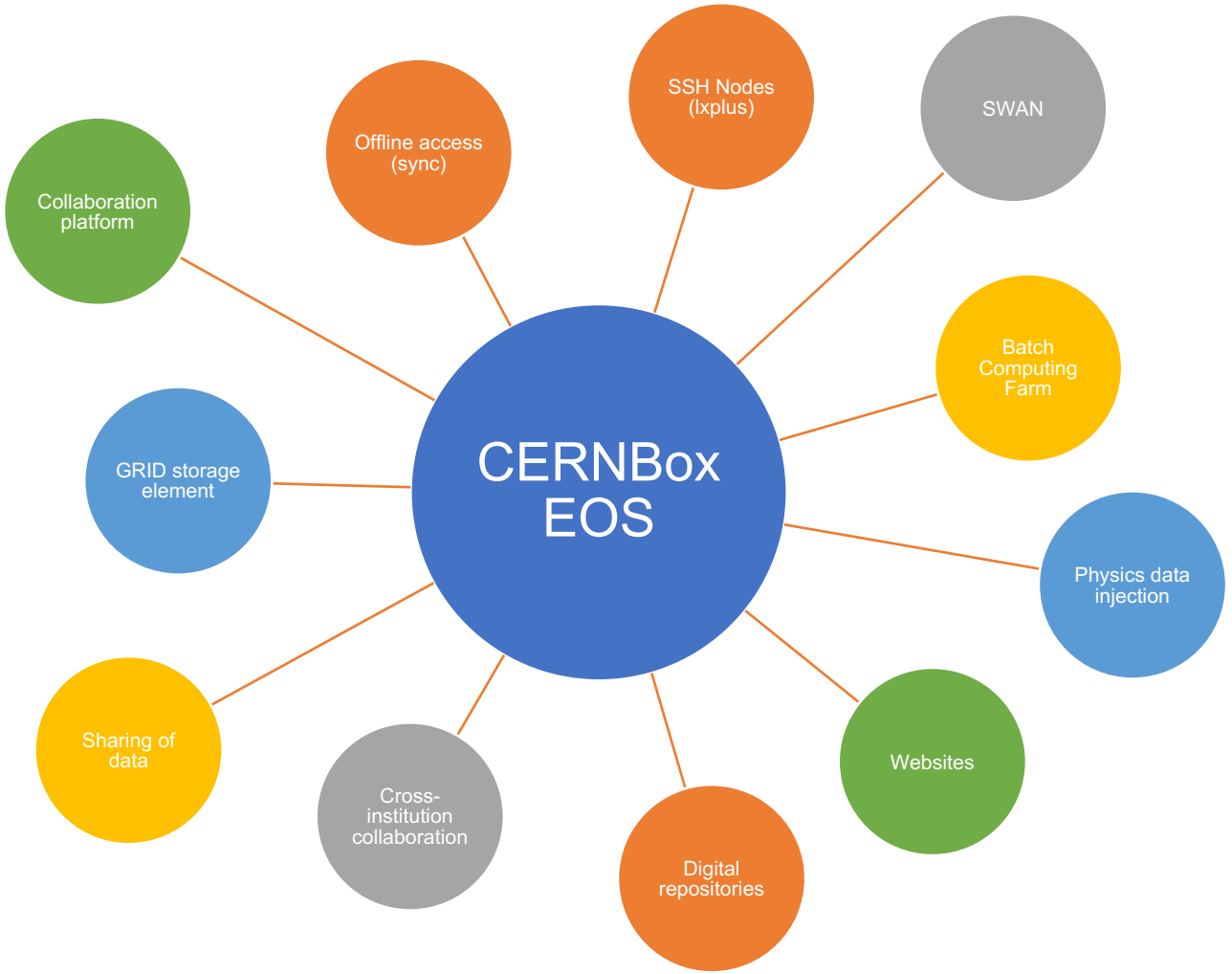
Fast transfer of large datasets from one site to another.

Remote data analysis through sync-and-share services.



Conclusion

CERNBox, the integrated layer for Scientific Data Lifecycle



Reminder:

BoF session this afternoon

on how to deploy a *CERNBox-like* system

CERNBox, the Scientific Cloud powered by EOS

Thank you

Diogo Castro
diogo.castro@cern.ch