Report da CS3
Cloud services for syncronisation and sharing

https://indico.cern.ch/event/1210538/

General info

9th edition, Barcelona, 6-8/3

120 participants on side + 30 remote, > 50 presentations

Jointly organized by CERN, Esade, ETH, SUNET, SURF, Trust-IT

INFN: Diego Ciangottini, Stefano Stalio and me

Industry Participation

Collabora Online
Comtrade Three-Sixty
ddn
Nextcloud
ONLYOFFICE
ownCloud
safespring
Seafile

This conference is supported by

CS3 MESH4 EOSC
Connecting European Data

This project has received funding from the European Union's Horizon 2020 research and innovation programme.

https://cs3mesh4eosc.eu/

General info

9th edition, Barcelona, 6-8/3

120 participants on side + 30 remote, > 50 presentations

Jointly organized by CERN, Esade, ETH, SUNET, SURF, Trust-IT

INFN: Diego Ciangottini, Stefano Stalio and me

Visit to Mare Nostrum supercomputer in The Chapel

“This is science, you know, science can wait”

Keynotes (life science)

Prof. Alfonso Valencia

Technical and scientific challenges in biomedicine; a European perspective

Monday, 6 March, 10:00 – 11:00

Prof. Barend Mons

Stop data sharing

Tuesday, 7 March, 09:30 – 10:30

Barend Mons (GO FAIR initiative) - FAIR principles recommend that data (and services around them) should be Findable, Accessible, Interoperable and (thus) Reuseable, first and foremost by machines.

Machine-actionable data (Fully AI-Ready) -> less time spent data munging, less ambiguity and non-reproducibility

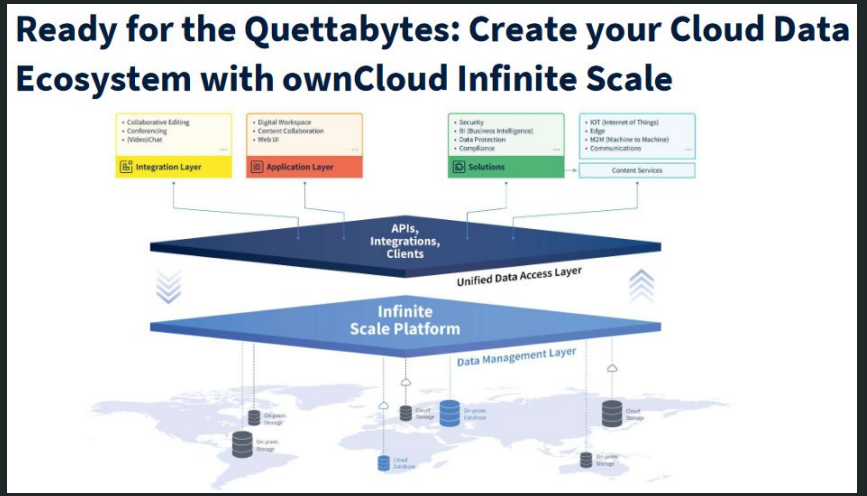
Many datasets are now either too large and/or too privacy-sensitive -> from data sharing to data visiting (distributed machine learning, swarm learning), from data management to data stewardship

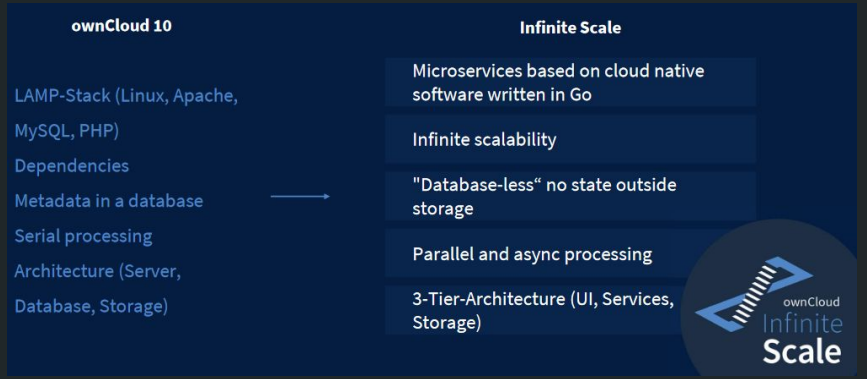
EFSS products

ownCloud Infinite Scale

Hoger Dyroff

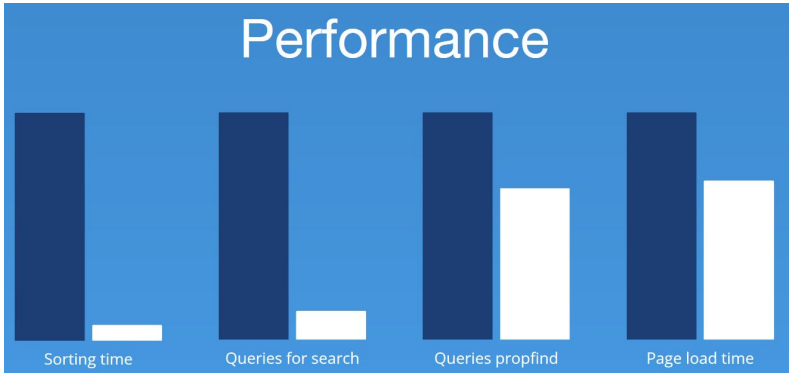
Cloud system testing toolbox
https://github.com/owncloud/cdperf

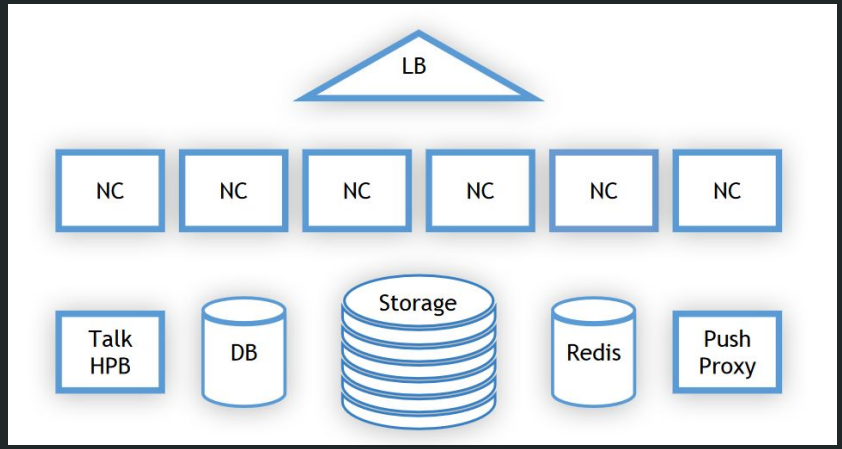




Nextcloud.
State of the nation

Frank Karlitschek





- T-systems: cloud migration (6 PB, 3 million users, 10k parallel users) took one year
- Transparent for the users

Seafile, what's new in the year 2022

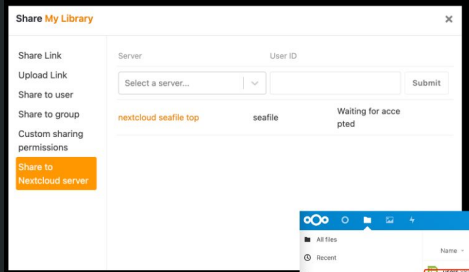
Jonathan Xu

https://wiki.geant.org/display/OCM/Open+Cloud+Mesh

Seafile 10.0 Features

- Notification server
- OCM: Access Seafile shares from NextCloud
- Watch libraries and receive notification on updates
- Multi-tenancy: configure SAML login for each tenant
- UI enhancements: colors, alignments, interactions
- Upgrade 3rd party components: NodeJS, libraries, WebDav, ElasticSearch

OCM: Connecting to NextCloud



- Folders and files in NextCloud can be shared to Seafile server, and vice versa
- NextCloud uses webdav protocol to access folders and files from Seafile
- WIP: Still some issues to be fixed when getting contents from Seafile

Site reports

Sync&share software

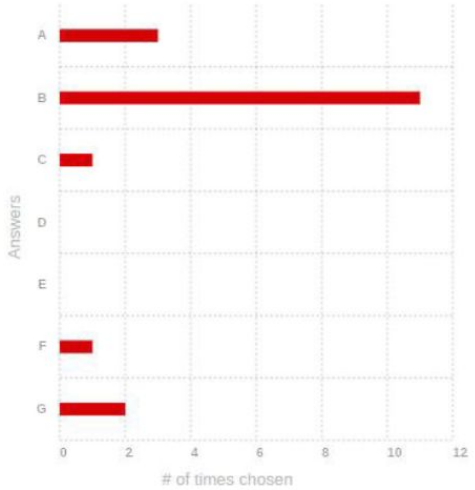
Answered: *16*

- **A. Nextcloud**: 3 (18.75%)
- **B. Owncloud**: 9 (56.25%)
- **C. Powerfolder**: 0 (0.00%)
- **D. Pydio**: 0 (0.00%)
- **E. Seafile**: 3 (18.75%)
- **F. other (please specify below in "Additional comments")**: 1 (6.25%)

Integrated software

Answered: *16*

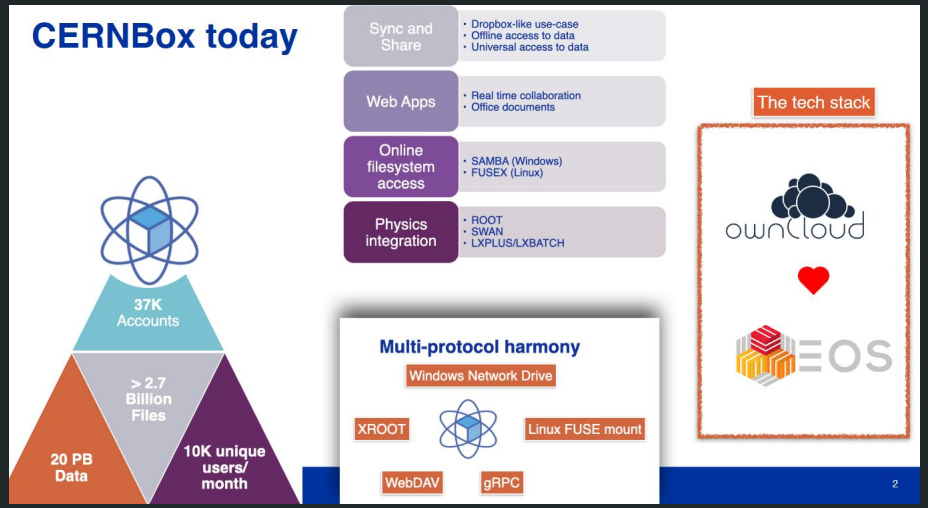
- **A. Collabora Online**: 3 (16.67%)
- **B. OnlyOffice**: 11 (61.11%)
- **C. SWAN Jupyter Notebooks** : 1 (5.56%)
- **D. Other Jupyter Notebooks** : 0 (0.00%)
- **E. Kopano (email)**: 0 (0.00%)
- **F. Microsoft Office Online / Sharepoint**: 1 (5.56%)
- **G. other (please specify in comment area)**: 2 (11.11%)



[Site report summary]

Standing innovation for 10 years: CERNBox

Hugo Gonzales Labrador - CERN

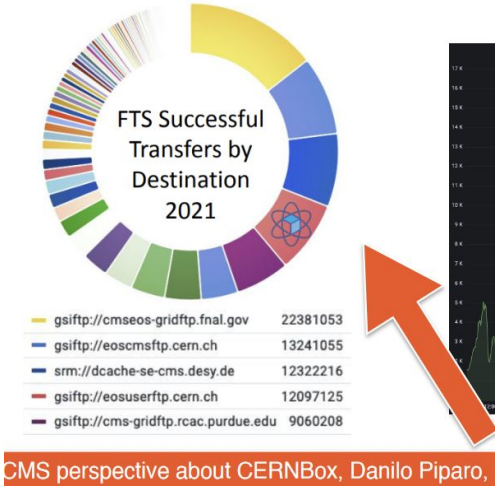


Gateway to the future (as seen in 2013)

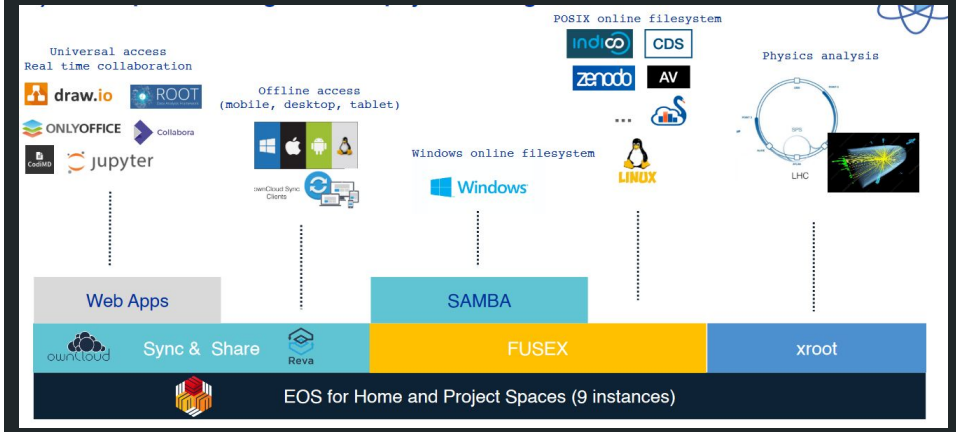
A) Classic DropBox use-case
B) Unified platform integrated with physics storage
C) Support scientific workflows
D) Delivering home directories
E) Federated dropbox service for HEP community

Standing innovation for 10 years: CERNBox

Hugo Gonzales Labrador - CERN



CMS perspective about CERNBox, Danilo Piparo,

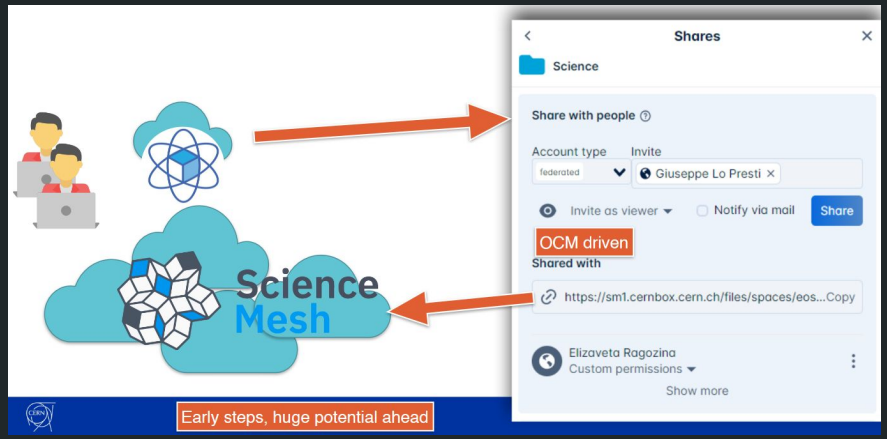


Gateway to the future (as seen in 2013)

A) Classic DropBox use-case
B) Unified platform integrated with physics storage
C) Support scientific workflows
D) Delivering home directories
E) Federated dropbox service for HEP community

Standing innovation for 10 years: CERNBox

Hugo Gonzales Labrador - CERN

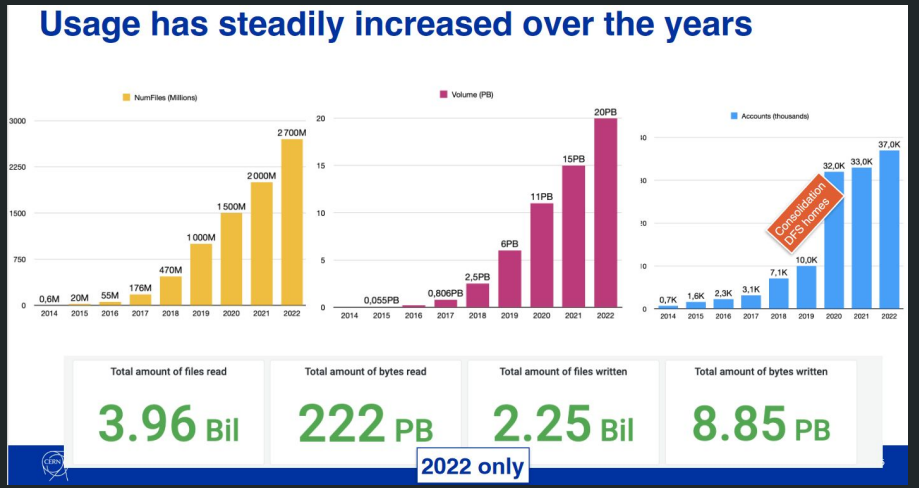


Gateway to the future (as seen in 2013)

A) Classic DropBox use-case
B) Unified platform integrated with physics storage
C) Support scientific workflows
D) Delivering home directories
E) Federated dropbox service for HEP community

Standing innovation for 10 years: CERNBox

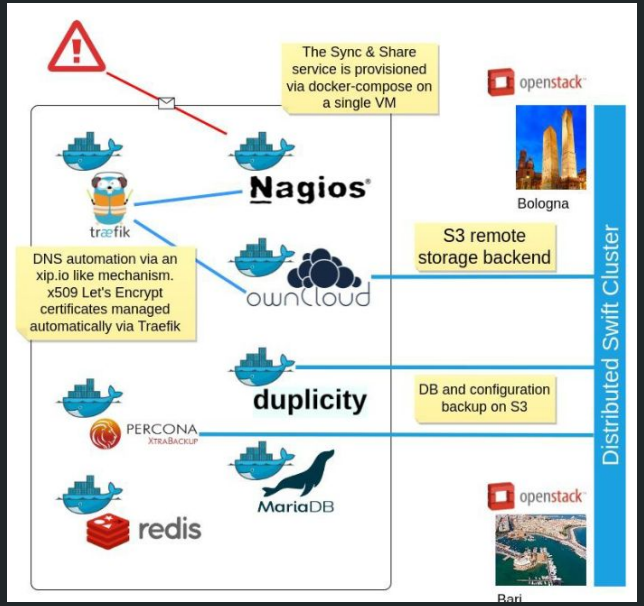
Hugo Gonzales Labrador - CERN



Extend beyond single storage technology (EOS, CEPH)
Expand beyond local borders (ScienceMesh)

INFN Cloud Sync & Share "aaS"

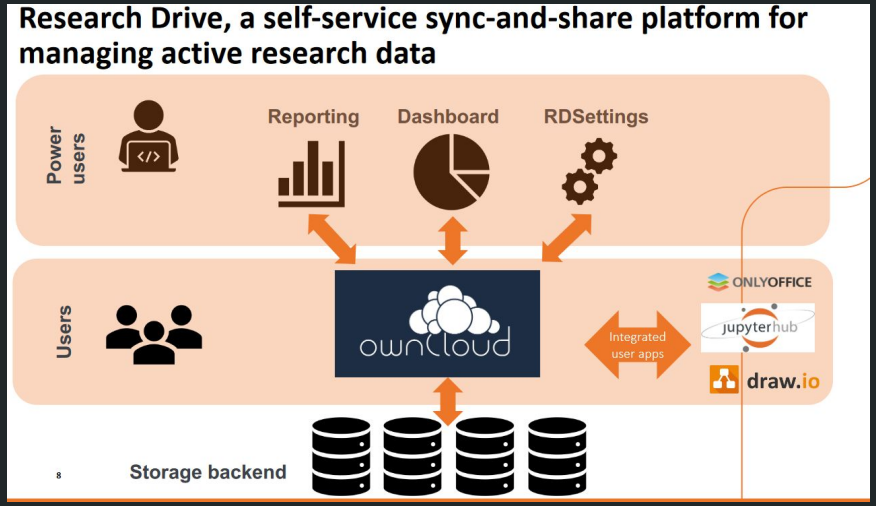
Stefano Stalio - INFN



- Small to medium communities
- Specific purpose
- Self-managed, sometimes only for limited period of time
- Built via docker-compose inside a single VM with external S3 backend
- Support for OwnCloud AND NextCloud

Research Drive - a platform for active data management

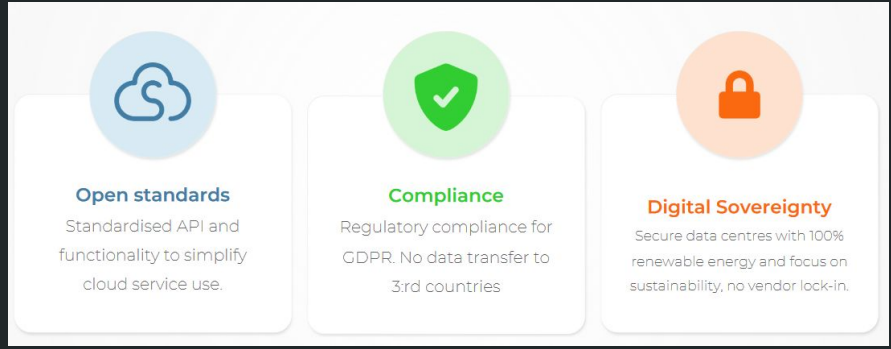
Narges Zarrabi - SURF

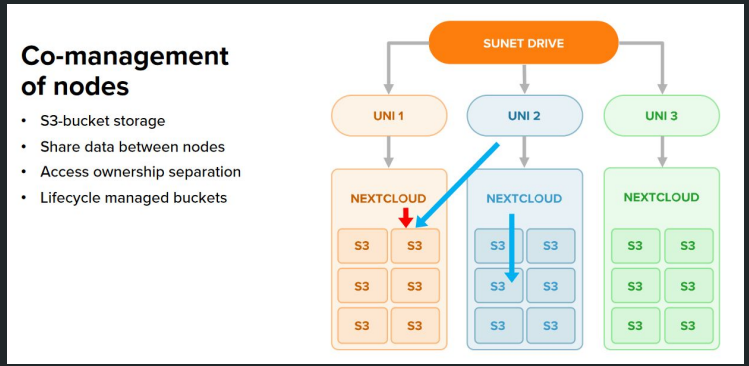


	Feb 2021	Feb 2022	Feb 2023
Total number of instances	22	31	35
Total users in all instances	4,500	10,880	16,000+
Total project folders	750	1750	2600+
Storage in TB	100	195	300+

Sunet Drive - An Academic EFSS Packaged for EOSC

Richard Freitag





FUTURE APPLICATION
ScienceMesh

Federated infrastructures and clouds

ScienceMesh: an interoperable federation of EFSS services

Pedro Ferreira - CERN

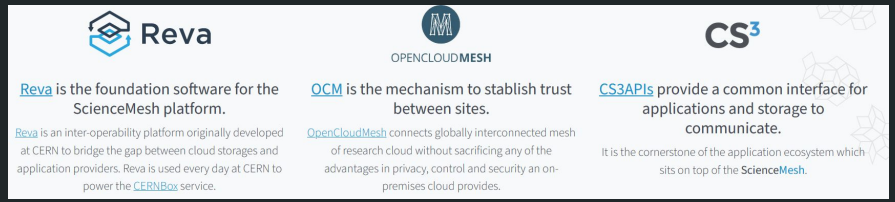
https://sciencemesh.io/

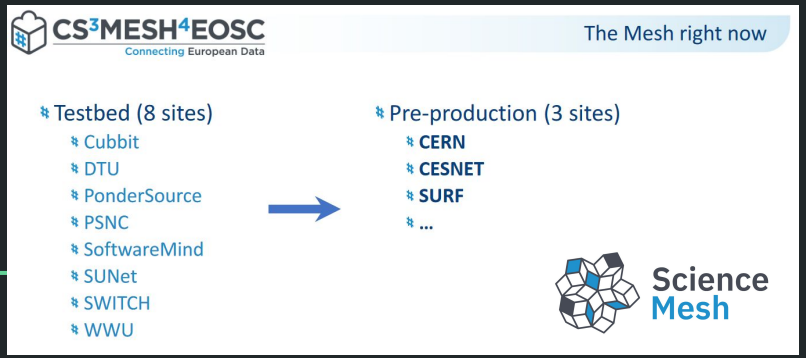
https://cs3mesh4eosc.eu/

https://wiki.geant.org/display/OCM/Open+Cloud+Mesh

An interoperable research platform for seamless sharing and collaboration on data across different EFSS systems.

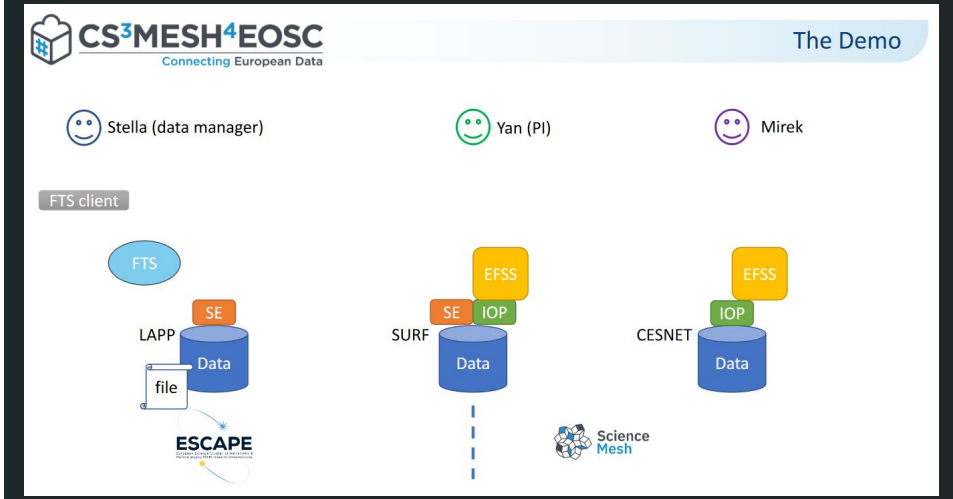
- Born out of 3.5-year EU Project
- Federated research space for Europe
- Decentralized Mesh of EFSS nodes
- OpenCloudMesh + discovery mechanism
- Data transfers: Rclone for point-to-point, FTS-Rucio between VOs

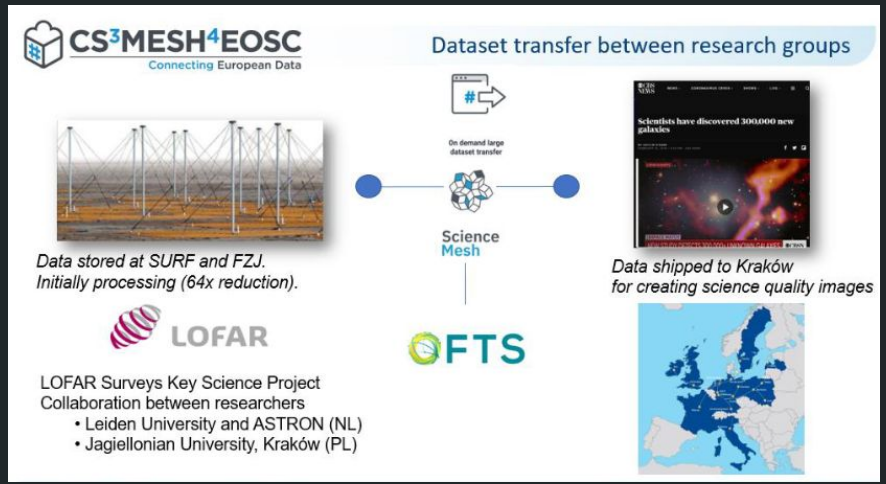




Data Transfers: Connecting Science Mesh and ESCAPE Data Lakes

Ron Trompert - SURF



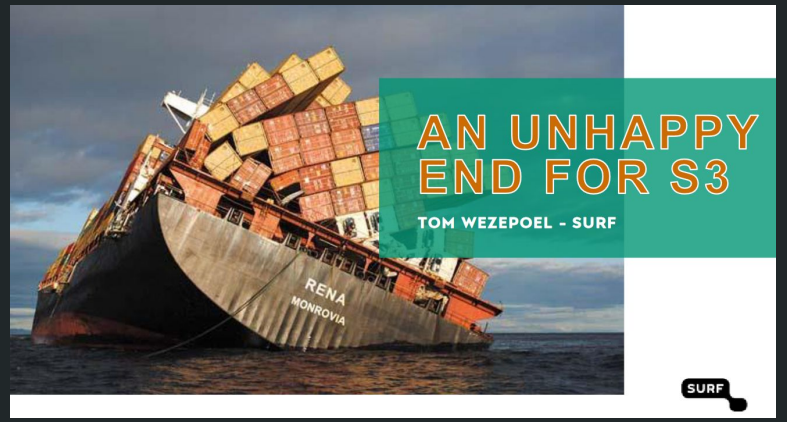


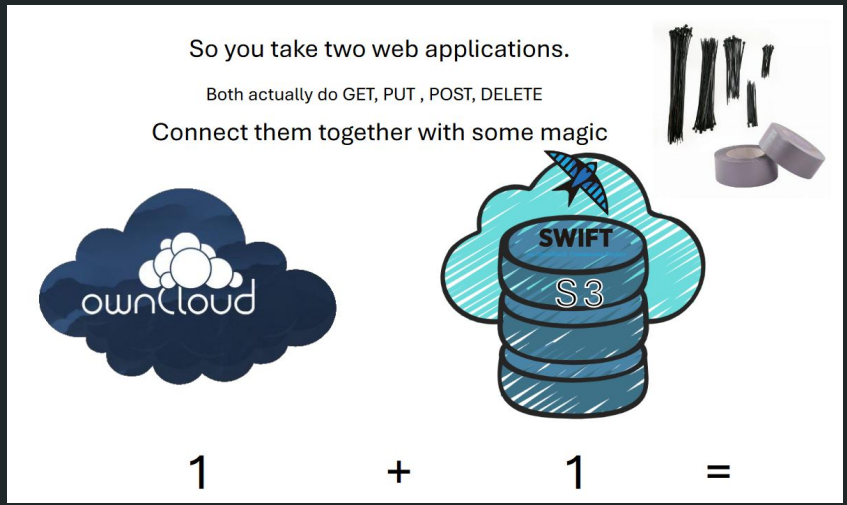
Scalable storage backends

An unhappy end for S3

Tom Wezepoel - SURF

Research Drive, the Dutch Sync & Share service based on ownCloud, uses OpenStack Swift S3 as its storage backend. The integration of S3 within the software is not that good, so migrate back to a posix compliant file system, namely CephFS (almost 2 PB of data)





Comparison between CephFS, CFFS (Comtrade FastFS), HDFS (Apache Hadoop), GPFS (IBM Spectrum Scale), Lustre

Gregor Molan - Comtrade

CDFS, appliance of CERN EOS by Comtrade (they provide hw and installation), uses RAIN instead of RAID.

... not very meaningful comparison... network fs all tested on one node...

Advantages of **RAIN**

Advantages

- Scalability
- Reliability
- Cost (JBOD without RAID controller)
- Geotag policies are applied during file placement to improve data loss prevention and IO performance.

Drawbacks

- All communication is done via the network
- Increased is IO and computational effort for non-sequential writes and server draining

The best

- Small files
  1. EOS on Linux
  1. Ceph on Linux
  1. GPFS on Linux
- Medium files
  1. EOS on Linux
  2. GPFS on Linux
- Large files
  1. EOS on Linux
  2. GPFS on Linux

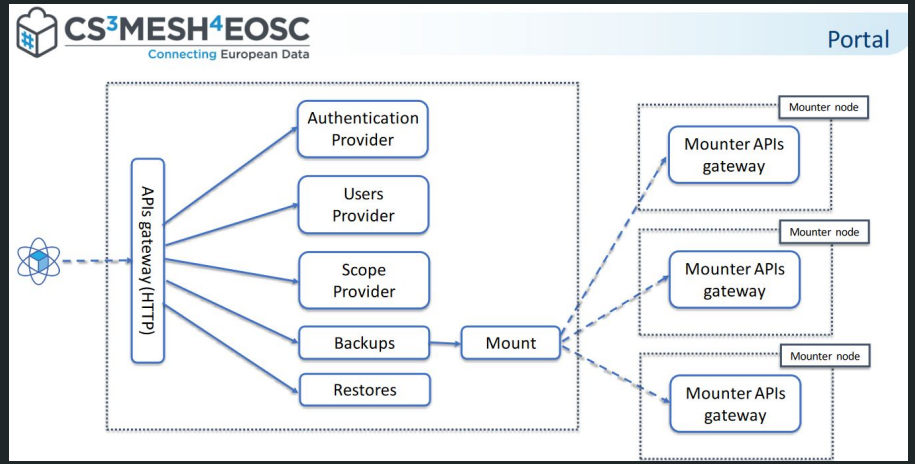
Not the best

- All file sizes
  - Hadoop on Win
  - Hadoop on Linux
  - Samba

C(ERN) BACK(UP): consolidated multi-petabyte backup solution for heterogenous storage and filesystems

Gianmaria Del Monte - CERN

For CERNBox, 7 PB data, 5.5 PB backup



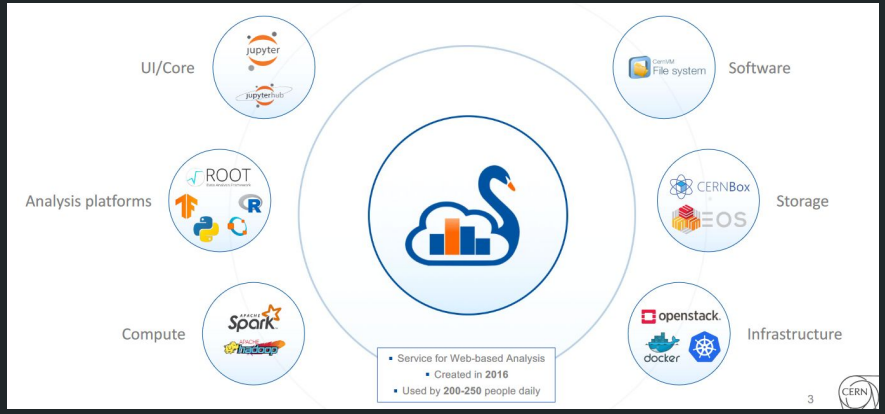
- Backup orchestration tool developed @CERN
- daily snapshots for all CERNBox users and projects and some services cephfs based (GitLab)
- Based on *restic*
- Expose to CERNBox users their daily snapshots, self-service restore

Collaborative data science and visualization

Evolving SWAN towards an Analysis Facility system

Diogo Castro - CERN

SWAN, CERN's Service for Web based ANalysis, is evolving from a plain notebook-based service into a fully fledged Analysis Facility, a single entry point to the multiple and heterogeneous storage, software and computational resources provided to CERN's research community.



AF @ CERN:

- A platform for single node and distributed analysis
- Connected to the SPARK clusters
- Allowing interactive use of GPUs
- Batch submission to HT Condor (also interactive with DASK)
- Integration ongoing with HPC service, CephFS, Kubeflow and ML

Running Interactive analysis on INFN Cloud Platform

Diego Ciangottini - INFN

Then we needed a shareable and scalable storage...

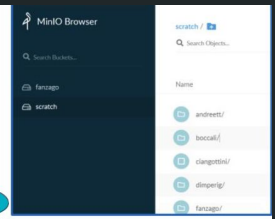
INFN Cloud provides object storage based on Openstack Swift, as centrally managed service.

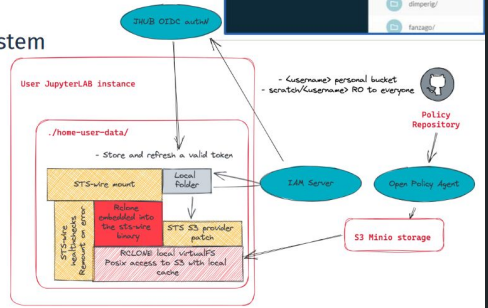
- The object storage is replicated in the two sites of backbone, Bari and CNAF
  - It guarantees the redundancy of data
- INFN Cloud is using the Minio-gateway software on Swift
  - Indigo-IAM OIDC authentication has been integrated and authorization policies are currently managed via OpenPolicyAgent
  - Ceph migration is under evaluation
    - Difficult to find good alternatives for Minio's WebUI tough
- Service URL https://minio.cloud.infn.it

Store JupyterLab persistent data

ll the services running notebooks
nount user areas store on the object
torage backend

- E.g. they are visible in the file system as posix directories.
- STS-wire: rclone mount+oidc shim library as been developed for the purpose





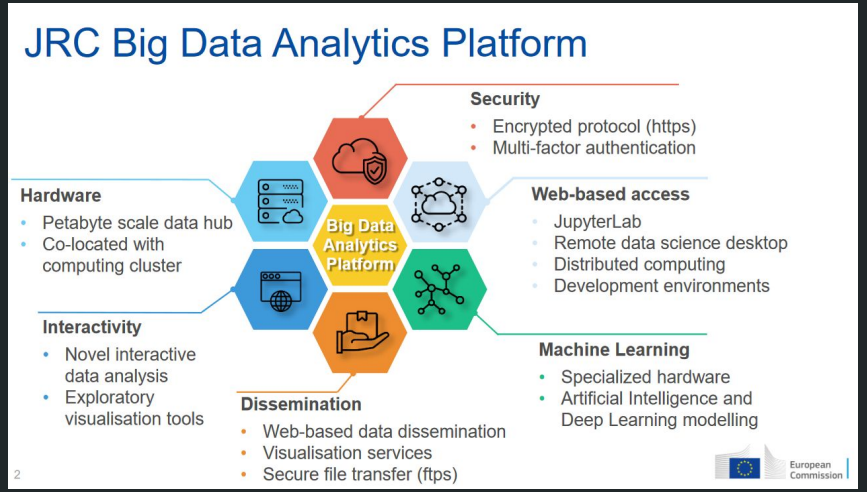
VOIS library: Pushing data science dashboards to the limits

Davide De Marchi - EC JRC

voilà

vois library

CS3 MESH4 EOSC
Connecting European Data



Voilà: a Jupyter notebook extension to automatically create standalone applications and dashboards
VOIS library: VOIlà Simplication library, partially developed in the context of CS3MESH4EOSC

Many more

https://indico.cern.ch/event/1210538