CERN

CERN IT-Storage Operations:

EOS/CERNBox and future strategy

Luca Mascetti
CERN - IT Storage

luca.mascetti@cern.ch

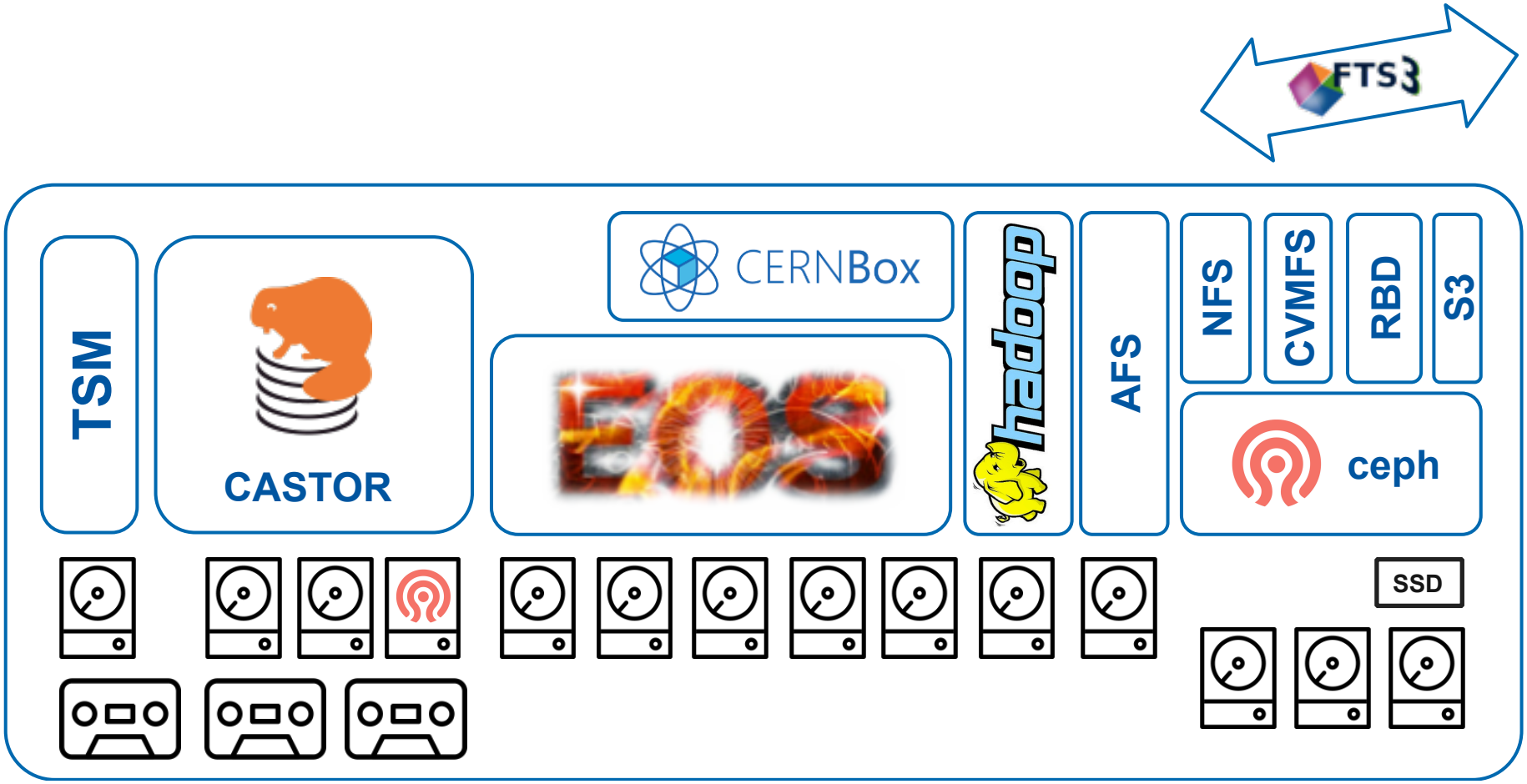
CERN IT-Storage

CERN IT-Storage Overview

Our goal

provide to experiments and users storage solution and tools for data management and data analysis, and operate them…

Storage Services



Vision for CERN Storage

Build a flexible and uniform infrastructure

Maintain and enhance (grid) data management tools

Vision for CERN Storage

Unlimited storage for LHC experiments at exabyte scale

EOS Disk pools (on-demand reliability, on-demand performance)

currently 165 PB deployed

CERN Tape Archive (high reliability, low cost, data preservation)

currently 190 PB stored

A generic home directory service

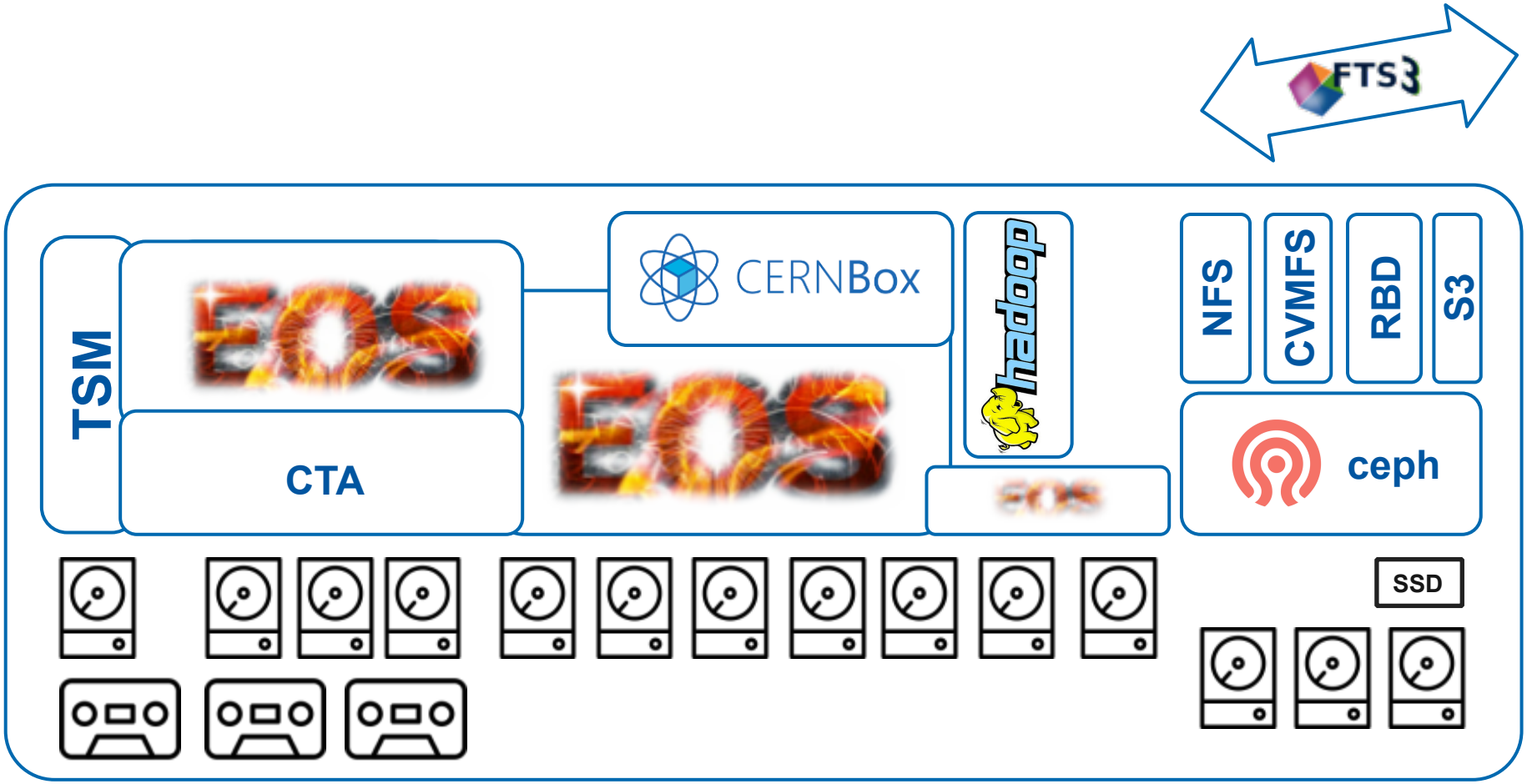
To cover all requirements of personal storage

Shared among all clients and services

Fuse mounts, NFS and SMB exports

Desktop Sync and Web/HTTP/DAV Access

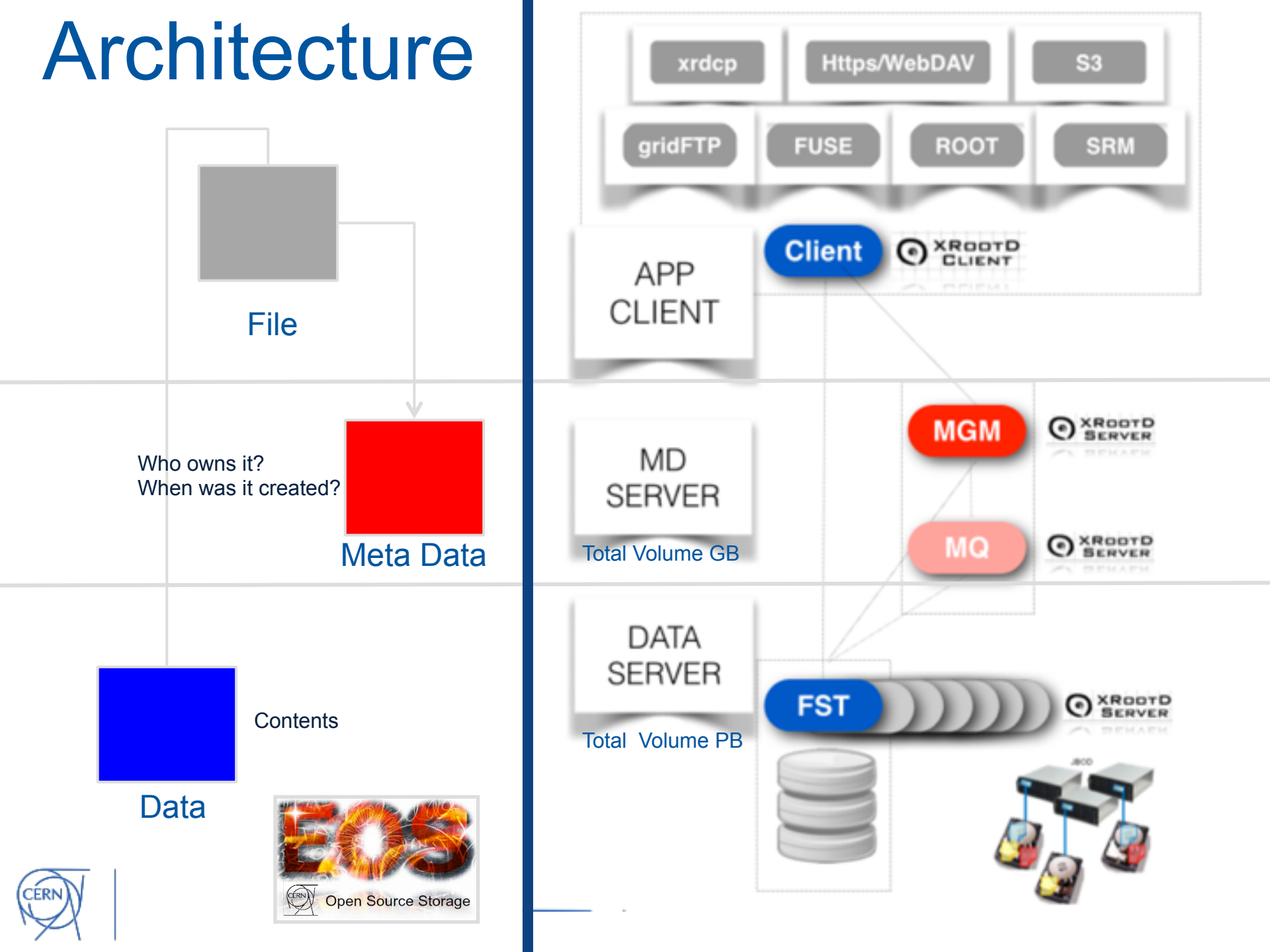
Storage Services Evolution



EOS

- Project started in 2010
- Licence free
- Simple and scalable solution, easy to operate
- Cheap, uses Network RAID (RAIN)
  - based on jbods, no raids
  - replication or erasure encoding
- In-memory namespace (low latency)
- Secure access (krb5, gsi)
- Quotas (user/group)
- Tuneable QoS
- Dev&Ops in CERN/IT-ST

Architecture



EOS Releases

named after gemstones… we have plenty of them ;)

Beryl Aquamarine

V 0.3.X

Citrine

V 4.X

XRootD V3 Server

IPV4

namespace in-memory data on attached disks

XRootD V4 Server

IPV6

plugins for meta data & data persistency

How is it used? CERN's mainstream usecase

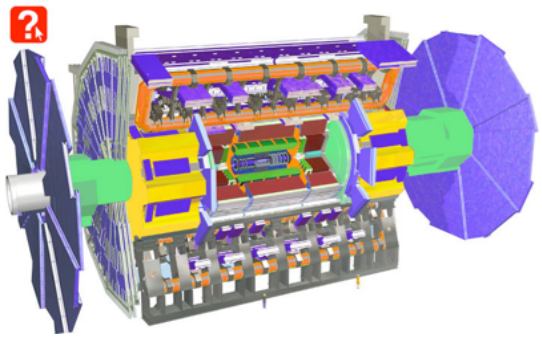
tape archive CASTOR

CERN Advanced STORage manager

local batch cluster

O(10^5) cores

LHC Detector



O(GB/s)

5-10 GB/s

EOS

Open Source Storage

peak 100 GB/s

5-10 GB/s

openstack™

Data Export to Worldwide Computing Grid

EOS Service @ CERN

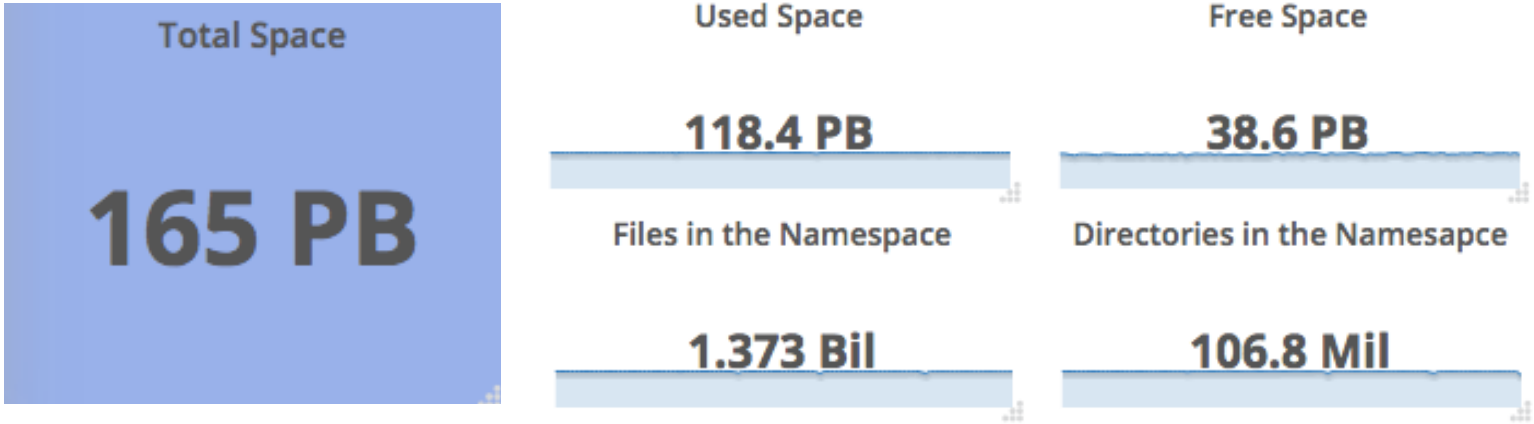
One instance per LHC Experiment

Public

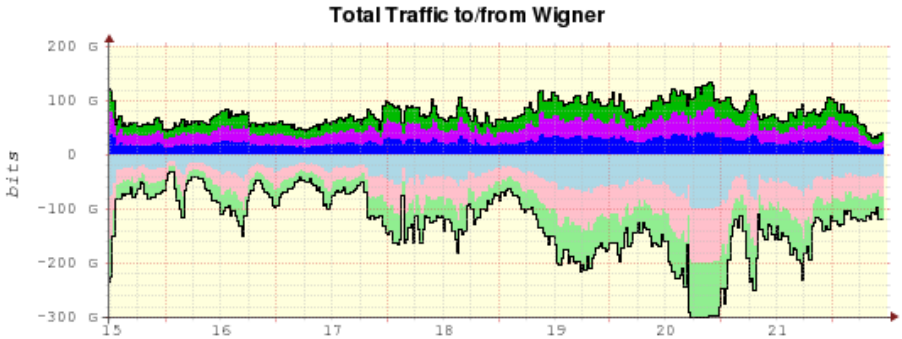
AMS
COMPASS
NA61/NA62
Others

CERNBox

EOS Service @ CERN



- Disk Storage for all LHC and physics data
  - and for CERNBox
- Deployment across two computer centres
  - CERN and Wigner RCP
  - Third link recently added



EOS Service @ CERN

Total Space

165 PB

Used Space

118.4 PB

Free Space

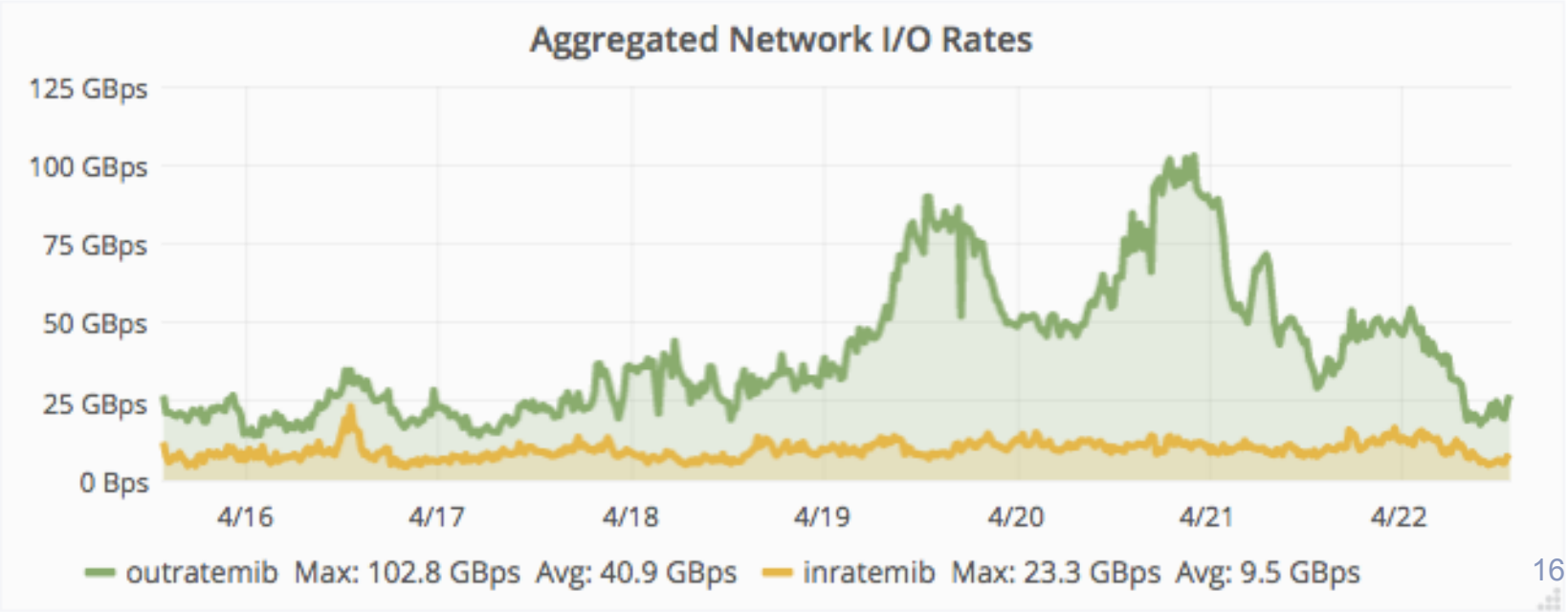
38.6 PB

Files in the Namespace

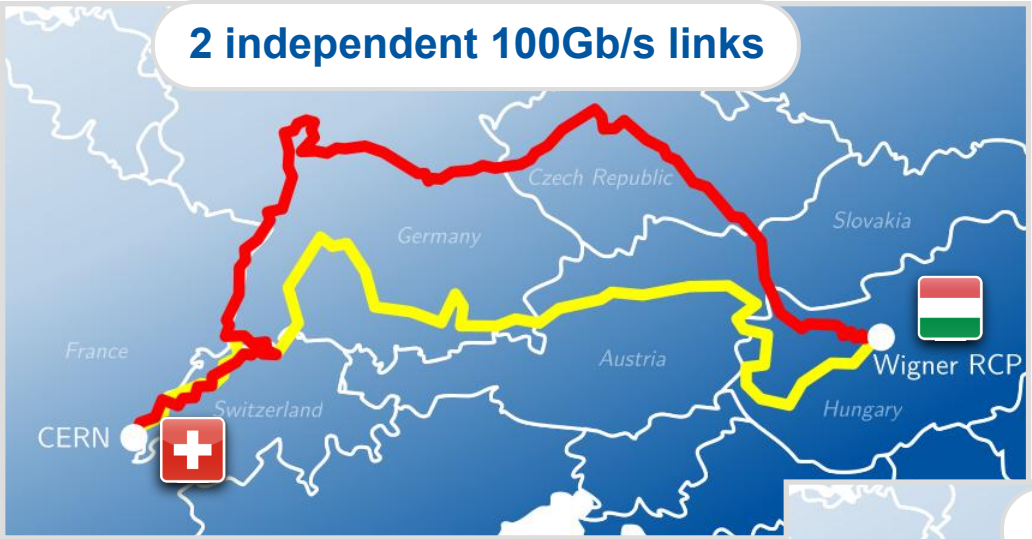
1.373 Bil

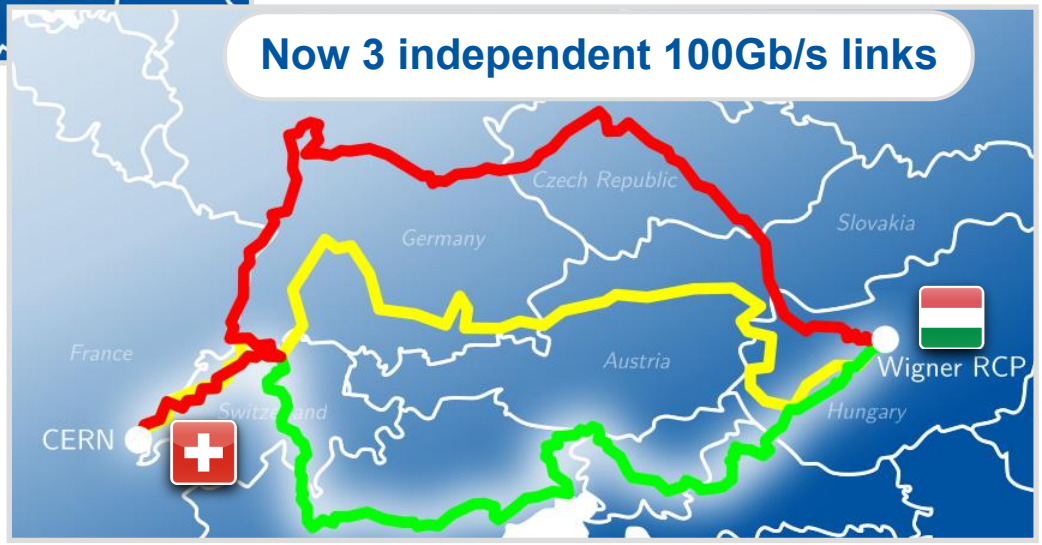
Directories in the Namesapce

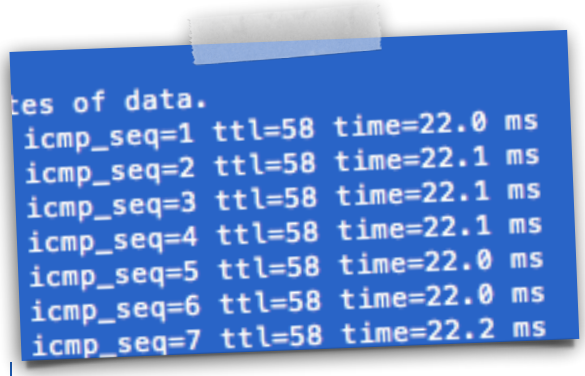
106.8 Mil



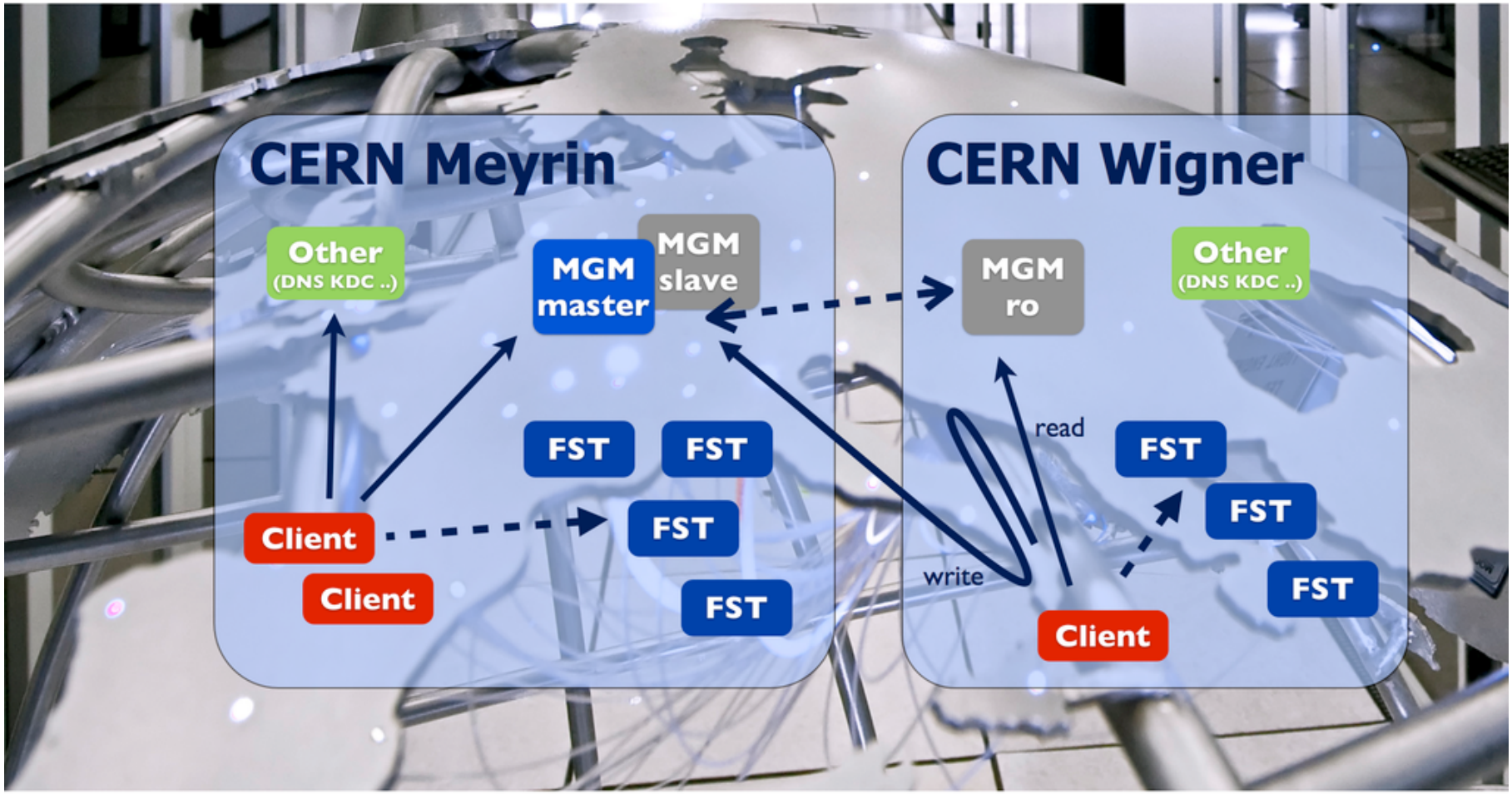
Wigner Computer Centre



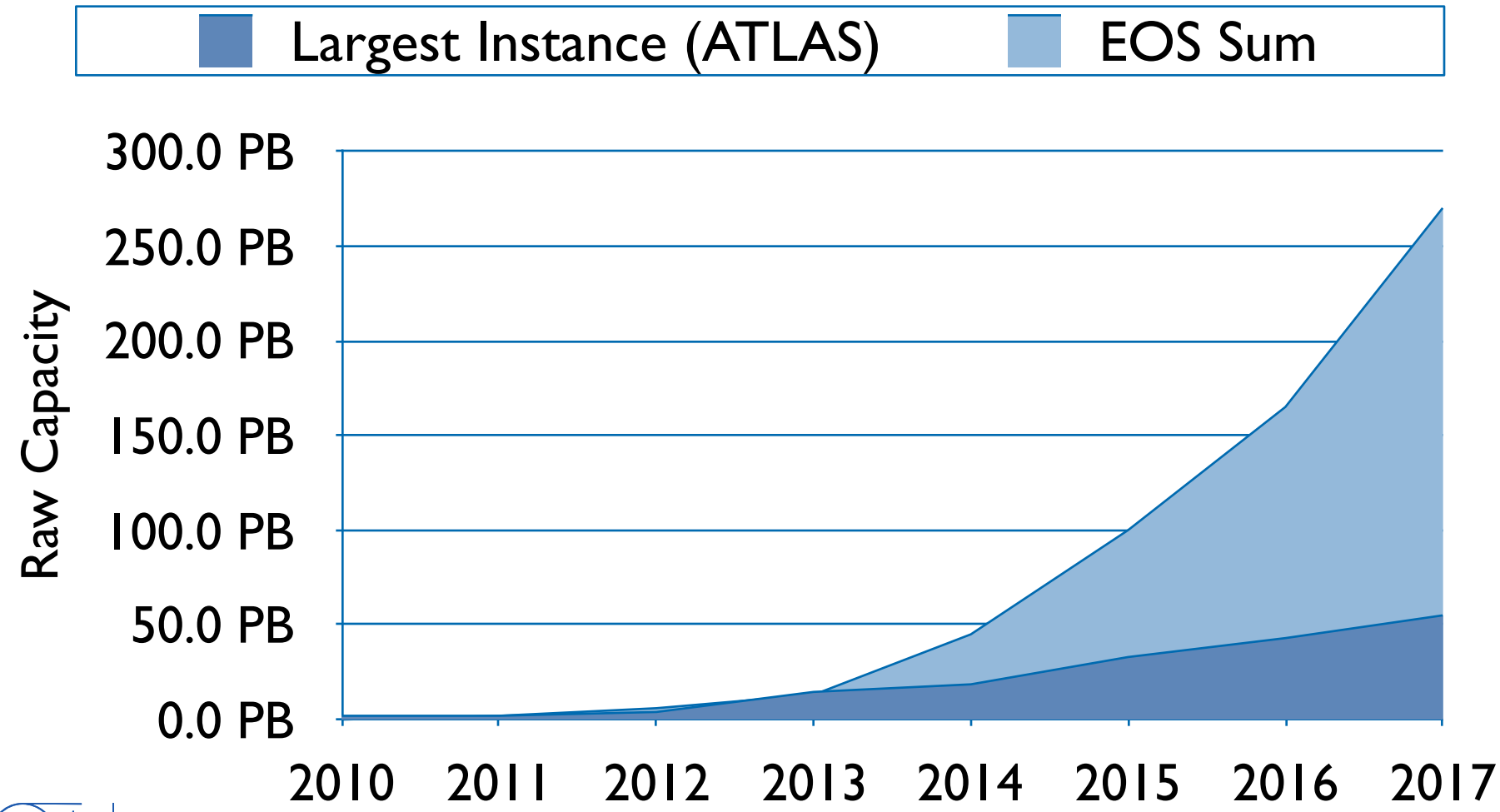




EOS 2015-2017 Deployment

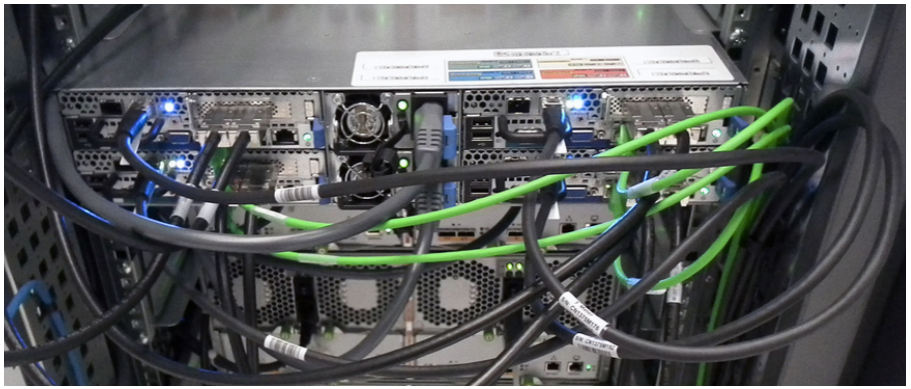


EOS Storage Deployment



Hardware Evolution

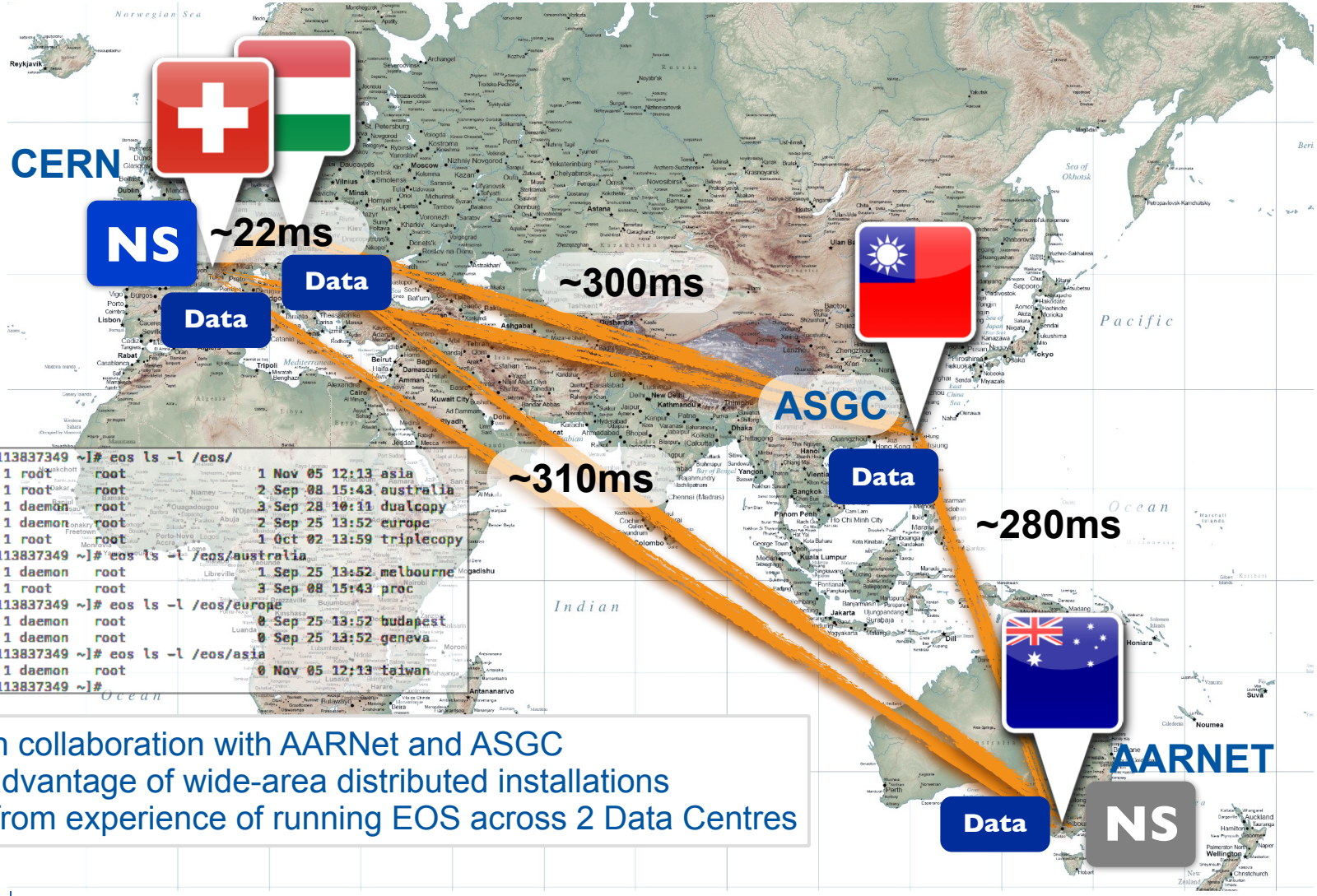
- Uniform storage for all our main services
- Profiting from economy of scale
  - minimise price per GB
- System Unit:
  - 8 physical cores (16 virtual) 64/128GB of RAM
  - 2x disk-tray of 24x 6TB HDDs (~**290 TB** brick)



- Prepared to receive additional 100PB on EOS (this week…)
- Planned test for the next generation
  - 4 trays per system unit + cascading (~600 TB brick)
  - 8 trays per system unit + cascading (~1.2 PB brick)

Global EOS

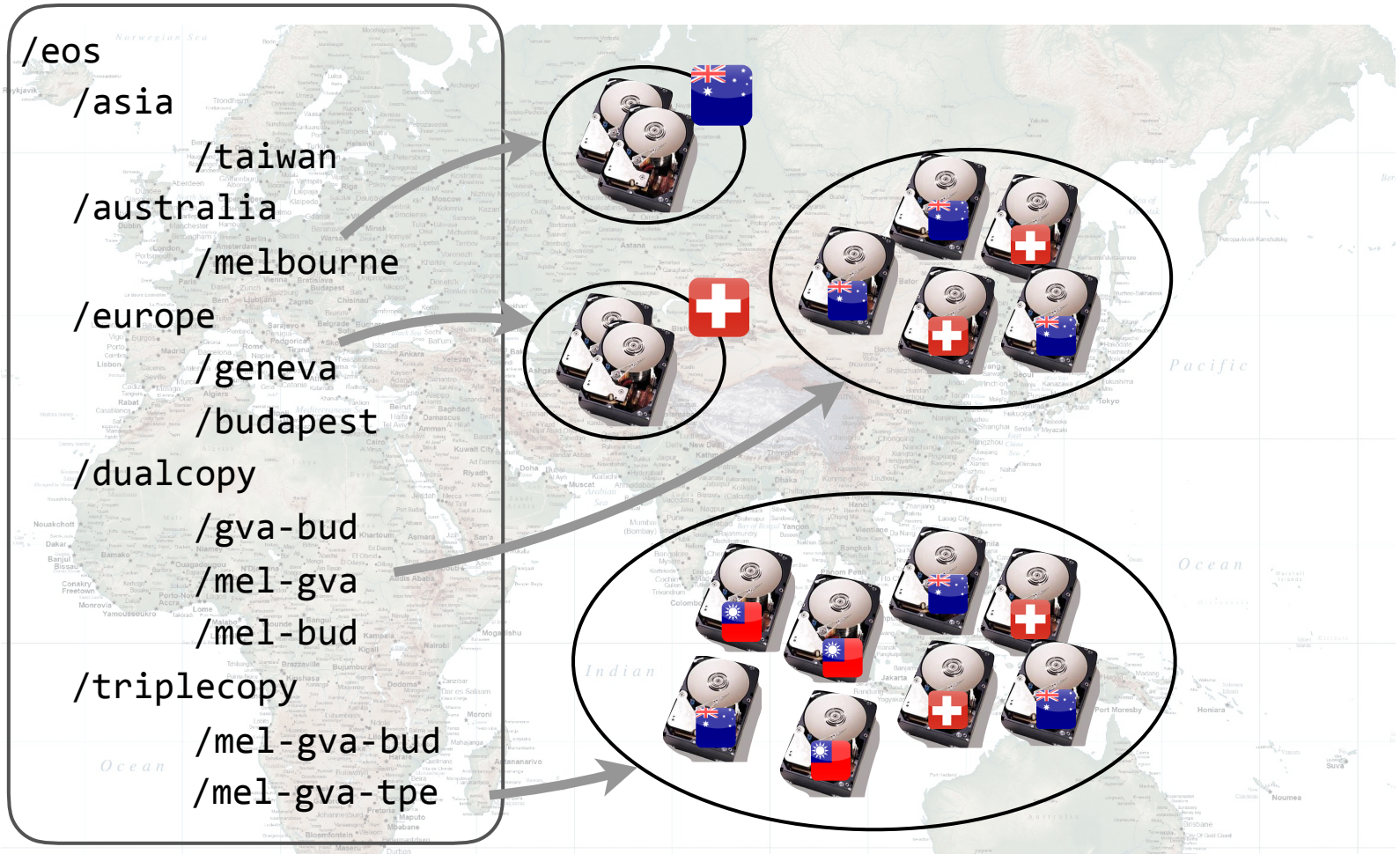
Global EOS: World-Wide Deployment




```
[root@p05151113837349 ~]# eos ls -l /eos/
drwxrwsr-+   1 root     root          1 Nov 05 12:13 asia
drwxr-xr-x   1 root     root          2 Sep 08 15:43 australia
drwxrwsr-+   1 daemon   root          3 Sep 28 10:11 dualcopy
drwxrwsr-+   1 daemon   root          2 Sep 25 13:52 europe
drwxrwsr-+   1 root     root          1 Oct 02 13:59 triplecopy
[root@p05151113837349 ~]# eos ls -l /eos/australia
drwxr-xr-+   1 daemon   root          1 Sep 25 13:52 melbourne
drwxr-xr-x   1 root     root          3 Sep 08 15:43 proc
[root@p05151113837349 ~]# eos ls -l /eos/europe
drwxrwsr-+   1 daemon   root          0 Sep 25 13:52 budapest
drwxrwsr-+   1 daemon   root          0 Sep 25 13:52 geneva
[root@p05151113837349 ~]# eos ls -l /eos/asia
drwxrwsr-+   1 daemon   root          0 Nov 05 12:13 taiwan
[root@p05151113837349 ~]#
```

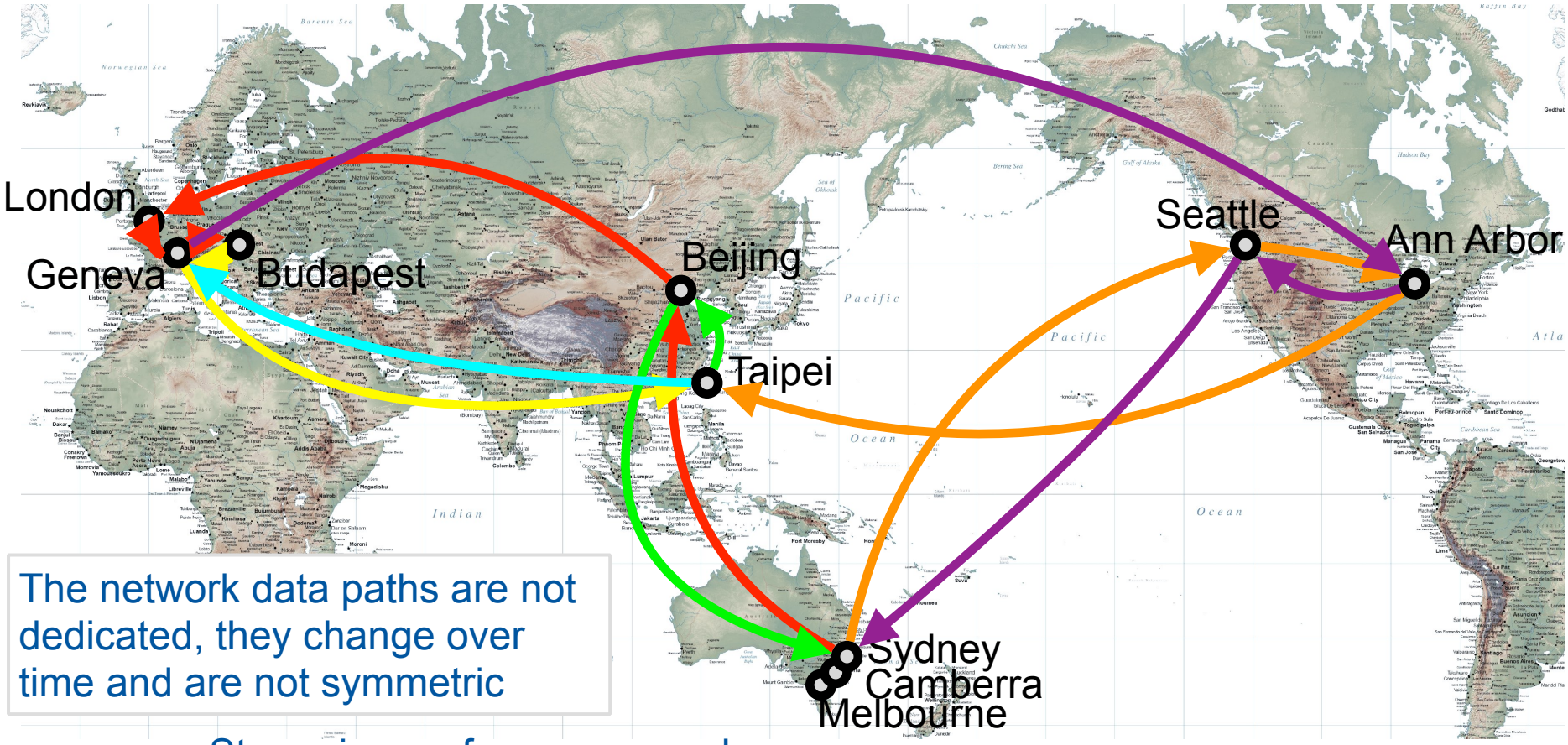
R&D in collaboration with AARNet and ASGC
Take advantage of wide-area distributed installations
Profit from experience of running EOS across 2 Data Centres

Data replication in storage pools



Storage pools were created with filesystems from all four sites. Files were replicated according to the different configured policy (e.g. 3 replicas: MEL-GVA-TPE).

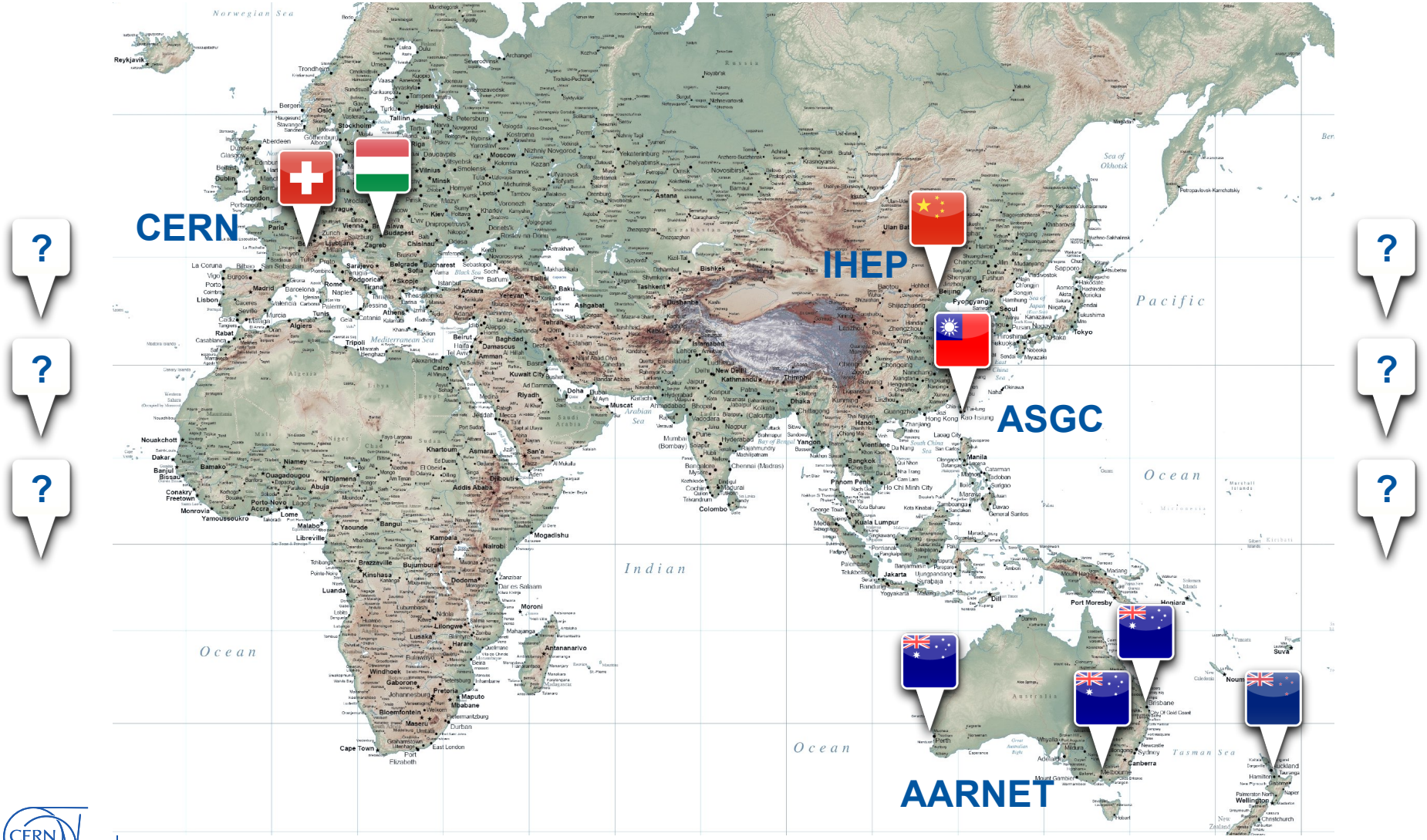
Network Topology



The network data paths are not dedicated, they change over time and are not symmetric

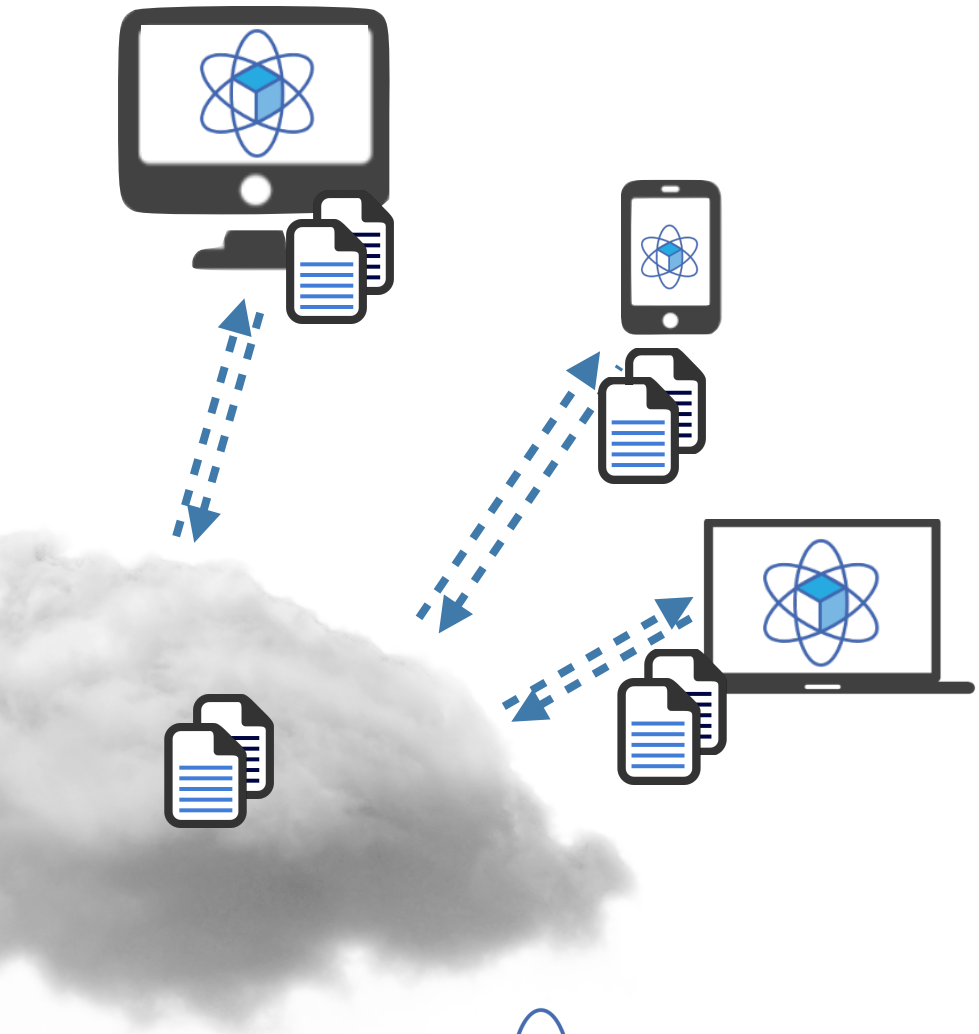
- Streaming performance good
  - possible problems in case of packet drops (tcp window)
  - tcp settings could be optimised
- Latency in read hidden by the read-only NS
- Latency in write to contact the read-write NS

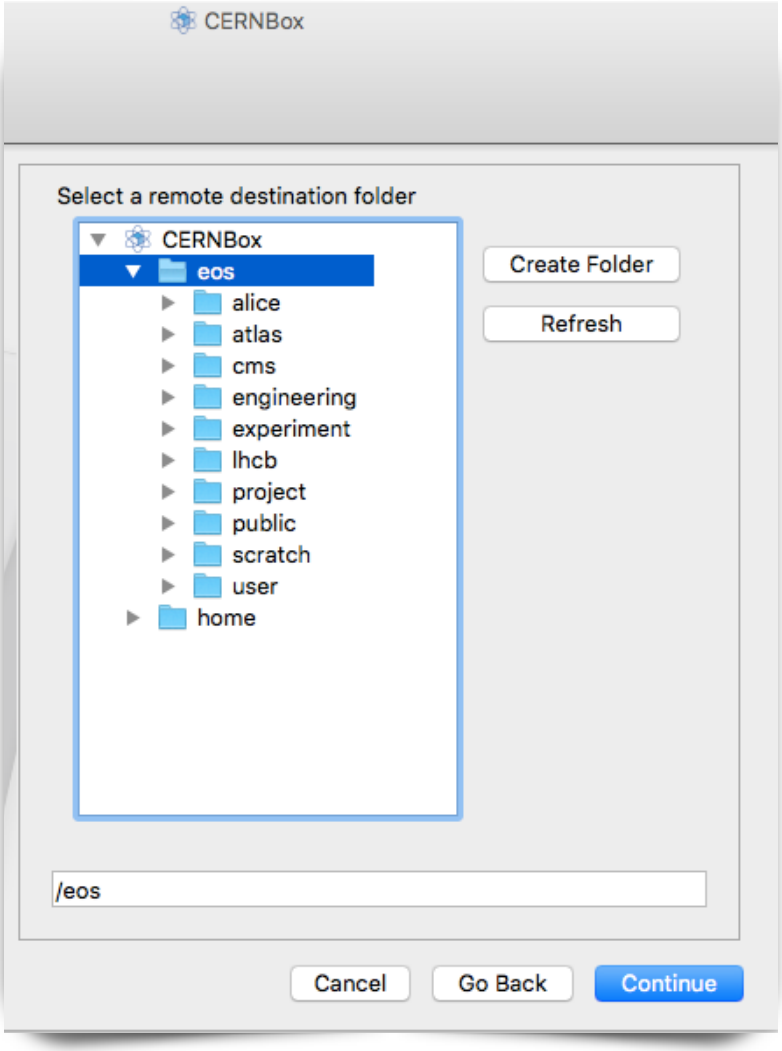
R&D - EOS World Wide Deployment - Future



CERNBox

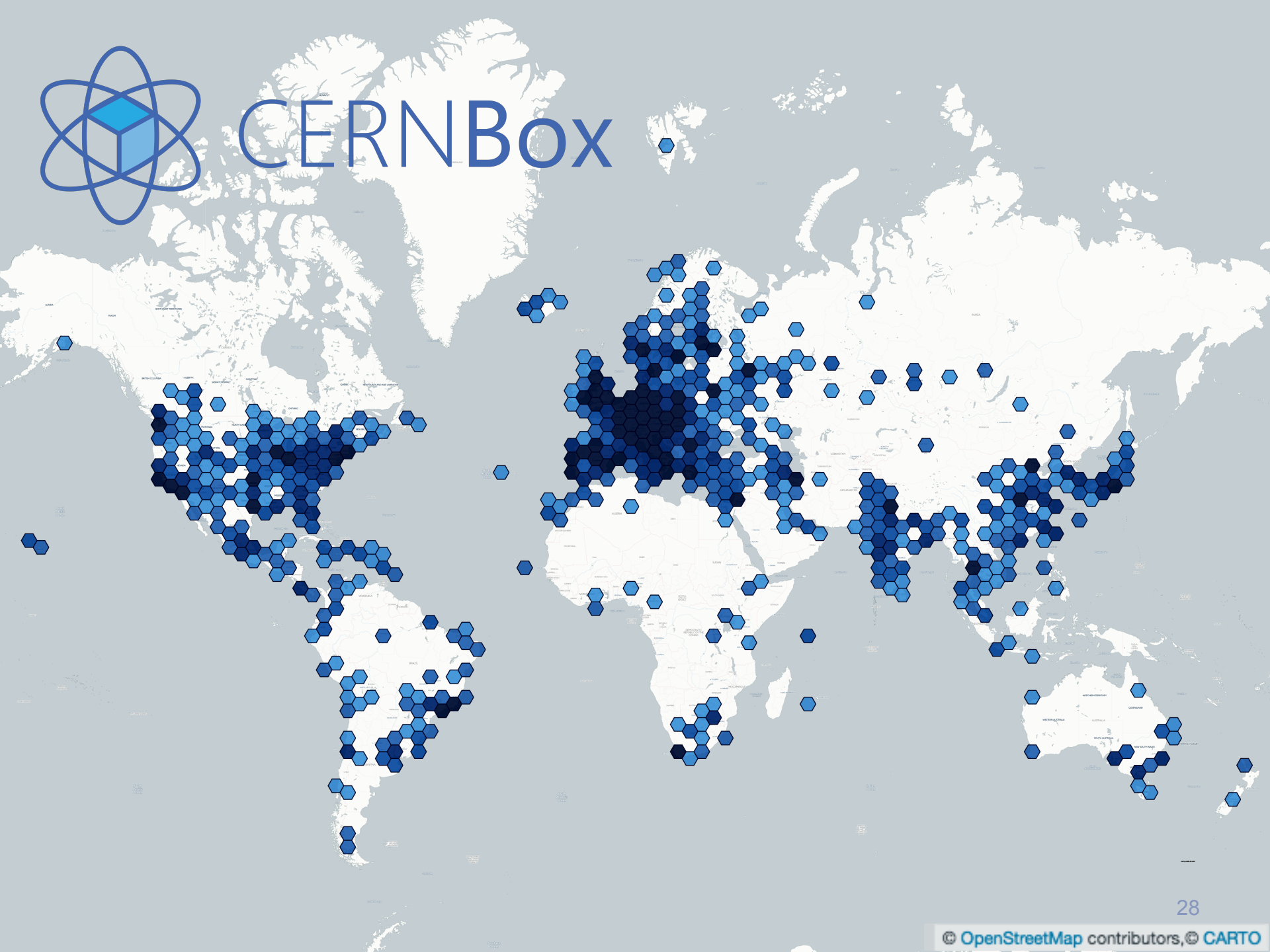
Bring data closer to our users: CERNBox

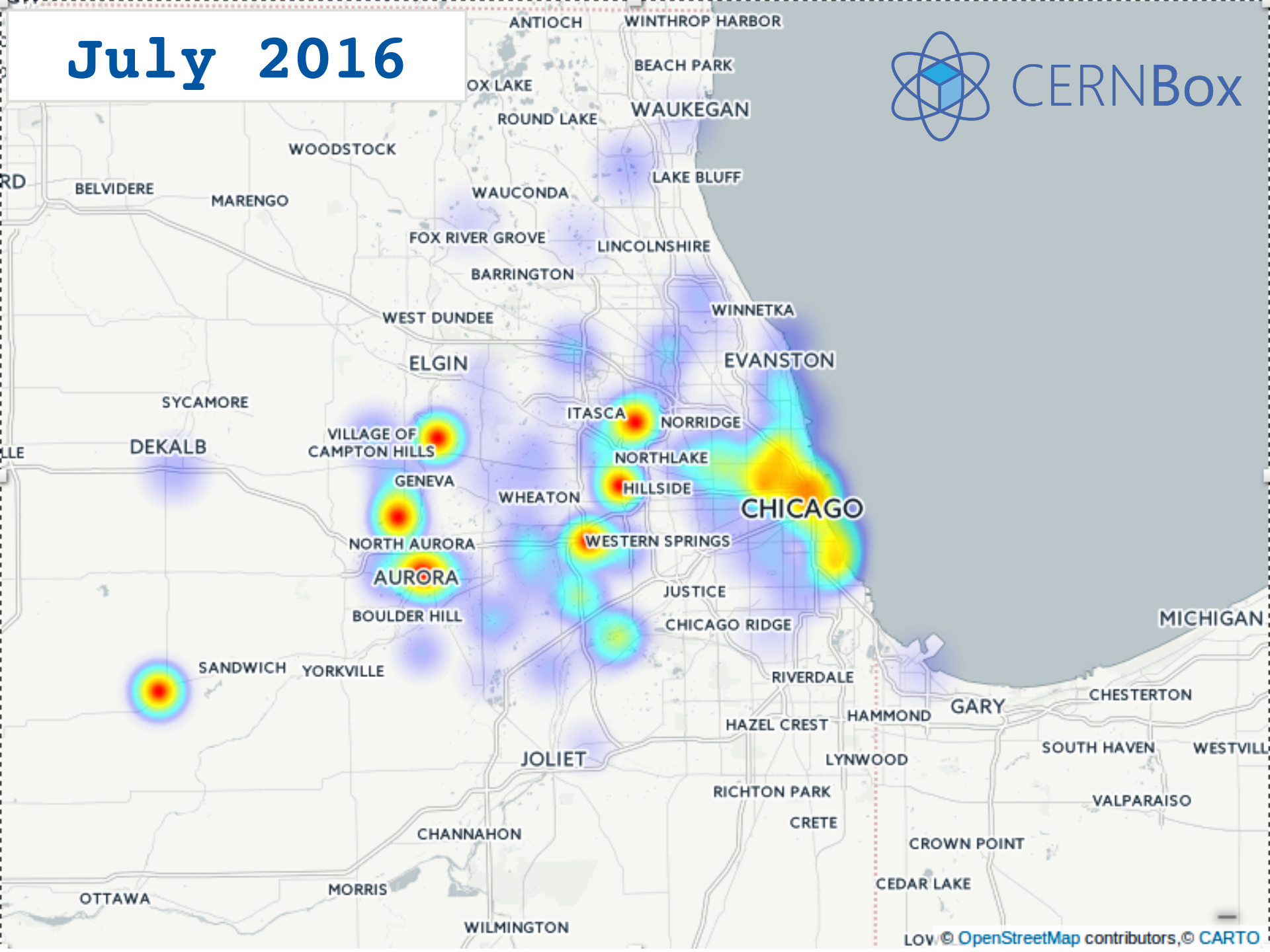




CERNBox powered by EOS

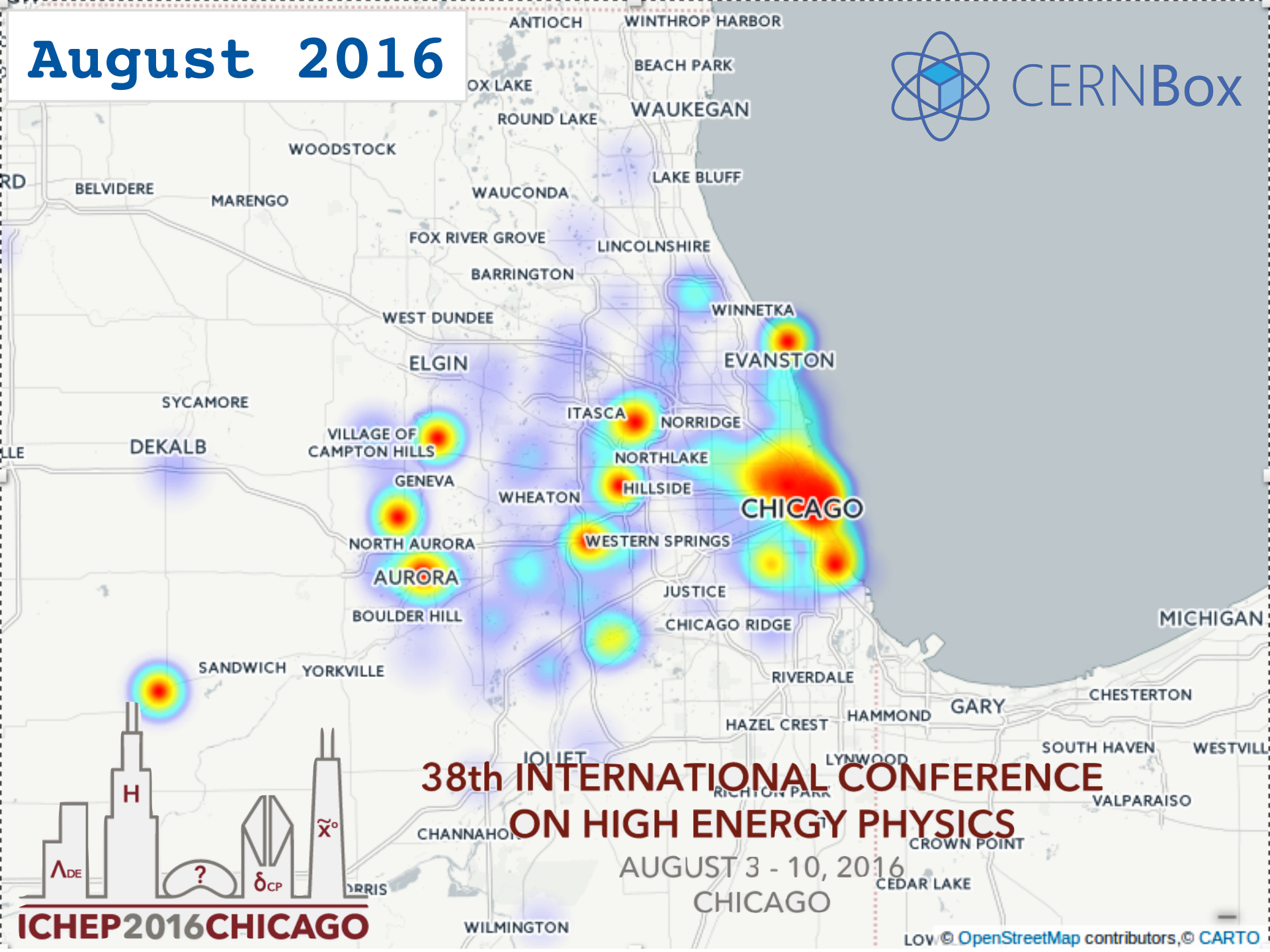
CERNBox




July 2016
CERNBox
ANTIOCH
WINTHROP HARBOR
BEACH PARK
OX LAKE
WAUKEGAN
ROUND LAKE
WOODSTOCK
LAKE BLUFF
BELVIDERE
MARENGO
WAUCONDA
FOX RIVER GROVE
LINCOLNSHIRE
BARRINGTON
WINNETKA
WEST DUNDEE
ELGIN
EVANSTON
SYCAMORE
ITASCA
NORRIDGE
VILLAGE OF CAMPTON HILLS
DEKALB
NORTHLAKE
GENEVA
HILLSIDE
WHEATON
CHICAGO
NORTH AURORA
WESTERN SPRINGS
AURORA
JUSTICE
BOULDER HILL
CHICAGO RIDGE
MICHIGAN
SANDWICH
YORKVILLE
RIVERDALE
CHESTERTON
GARY
HAMMOND
HAZEL CREST
SOUTH HAVEN
WESTVILL
JOLIET
LYNWOOD
RICHTON PARK
VALPARAISO
CRETE
CHANNAHON
CROWN POINT
CEDAR LAKE
OTTAWA
MORRIS
WILMINGTON
© OpenStreetMap contributors,© CARTO

August 2016

CERNBox



38th INTERNATIONAL CONFERENCE ON HIGH ENERGY PHYSICS

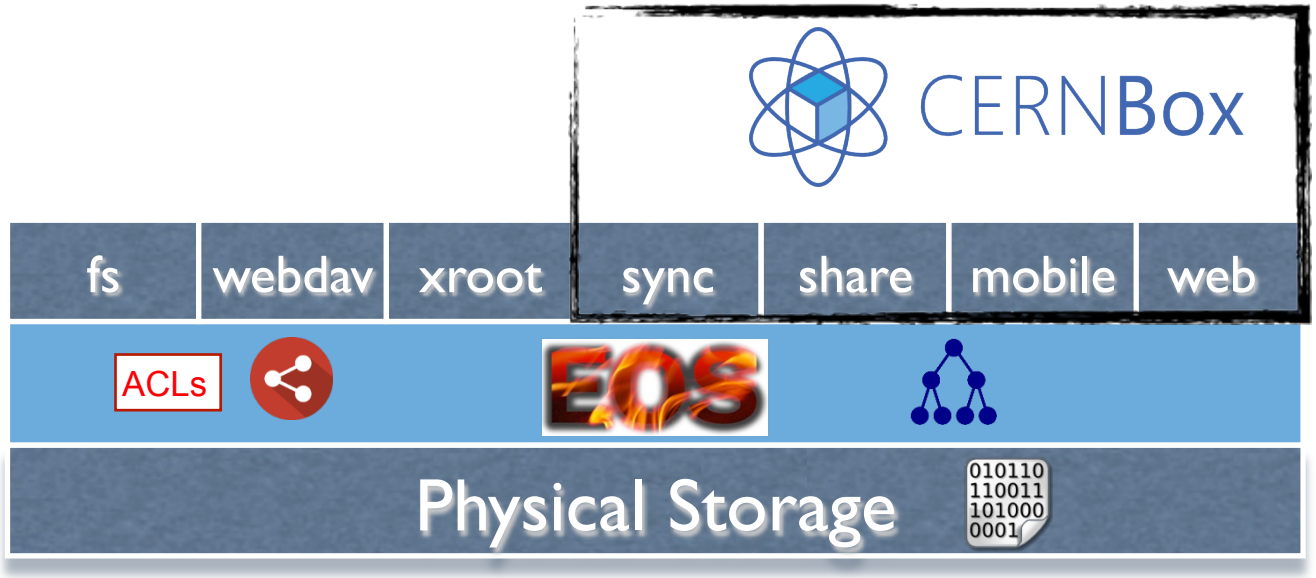
AUGUST 3 - 10, 2016
CHICAGO

ICHEP2016CHICAGO

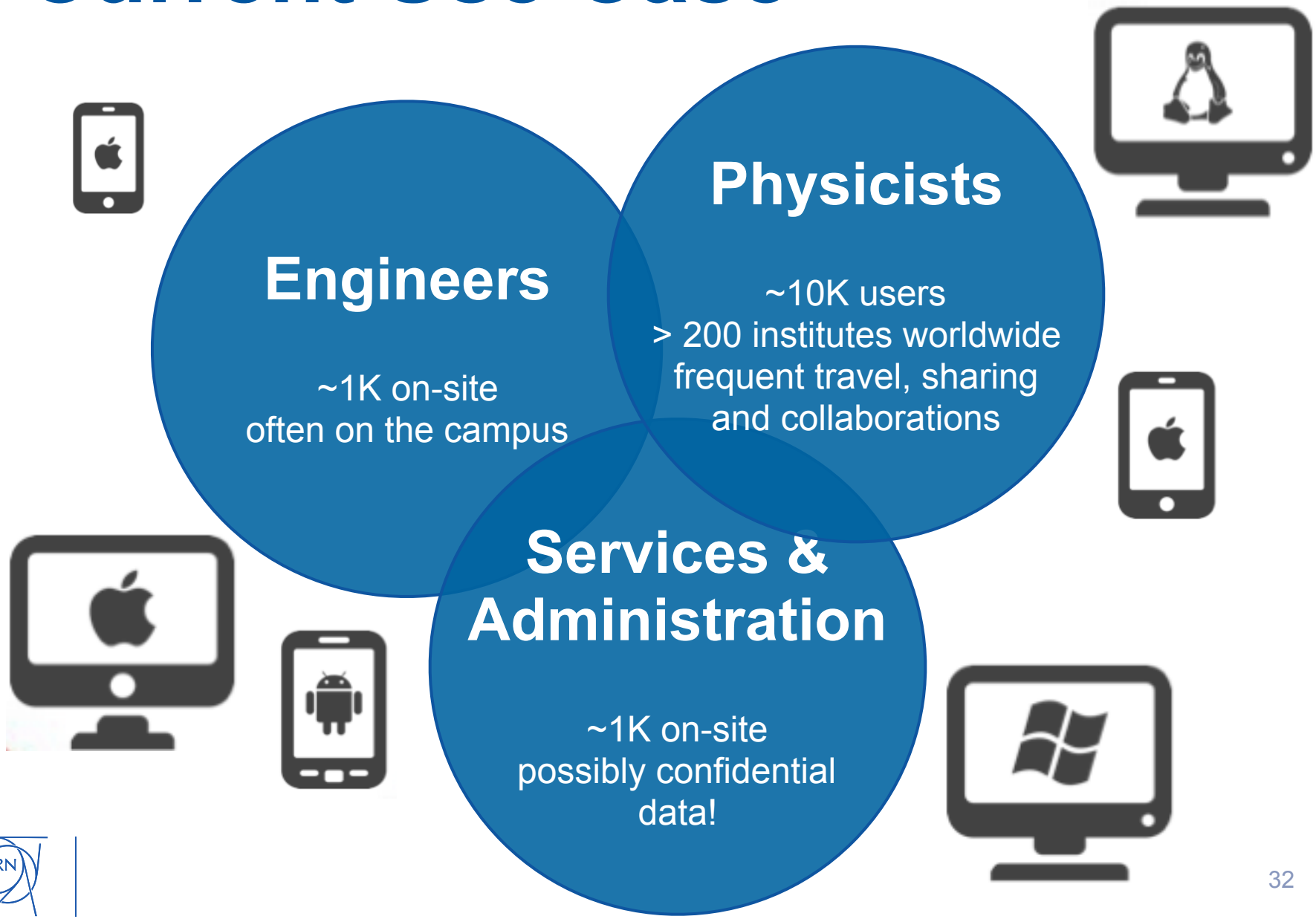
CERNBox

CERNBox provides a cloud synchronisation service

- Available for all CERN users (1TB/user)
- Synchronise files (data at CERN) and offline data access
- Easy way to share with other users
- All major platforms supported
- Based on **ownCloud integrated with EOS**

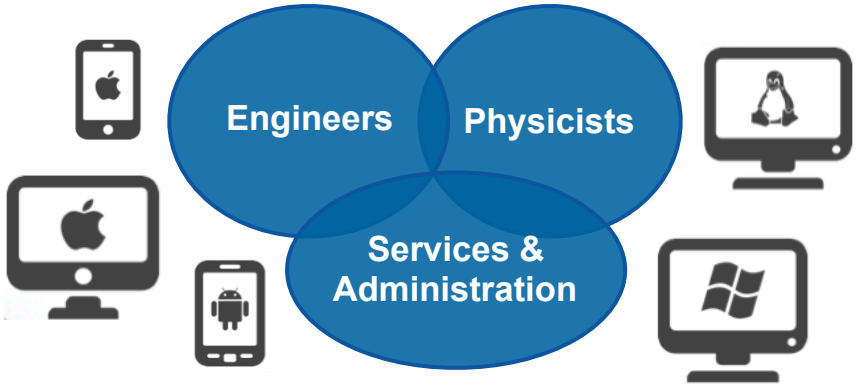


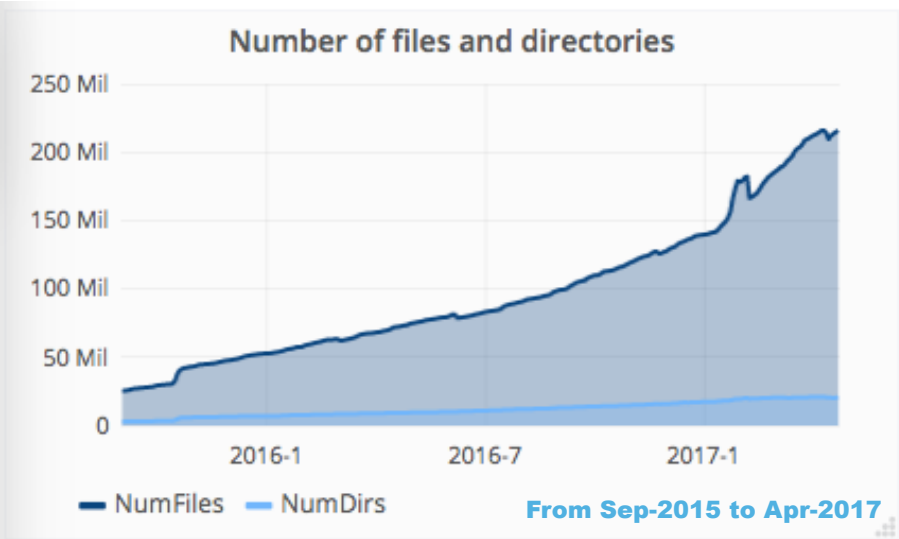
Current Use Case

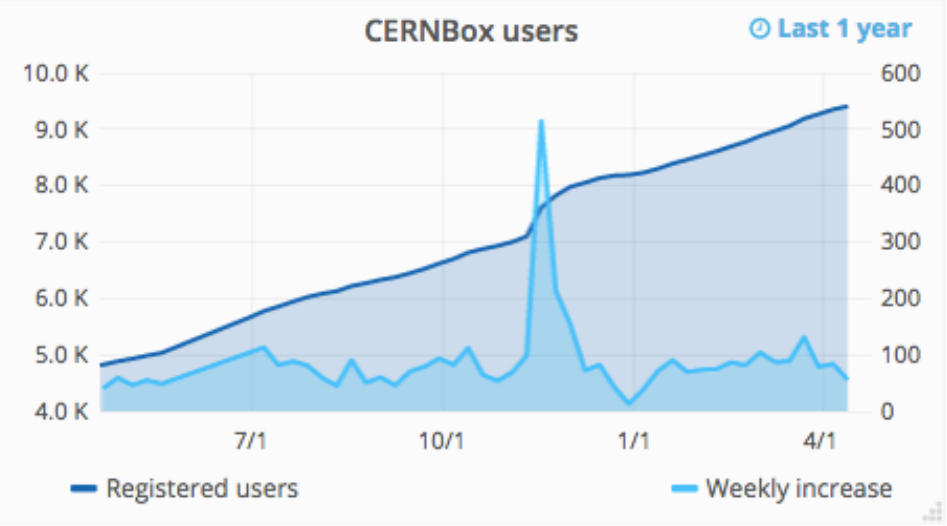


CERNBox Service Numbers

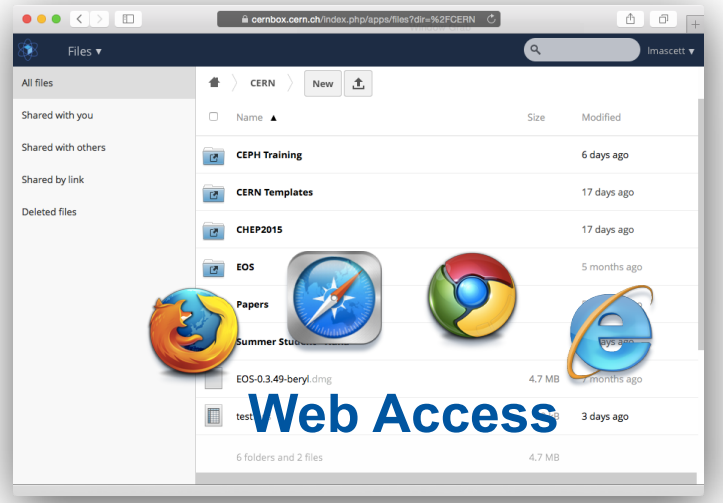
	Jan 2016	Jan 2017	May 2017
Users	4074	8411	9741
# files	55 Mill	176 Mil	233 Mil
# dirs	7.2 Mil	19 Mil	21 Mil
Used Raw	208 TB	806 TB	1110 TB
Deployed Raw	1.3 PB	3.2 PB	3.3 PB





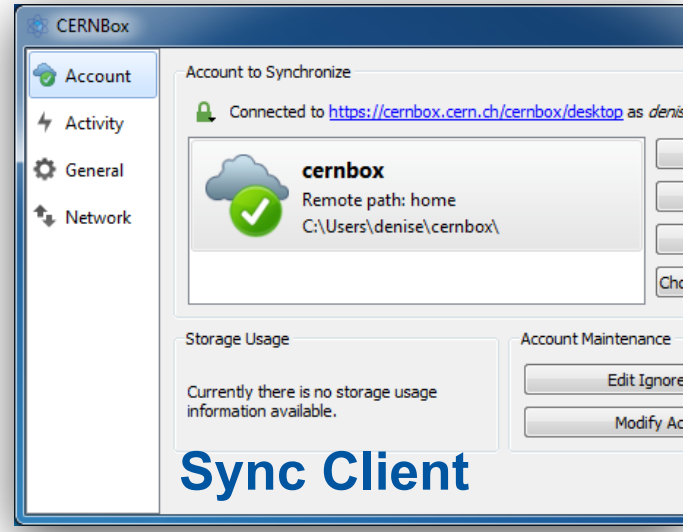


Available Access Methods

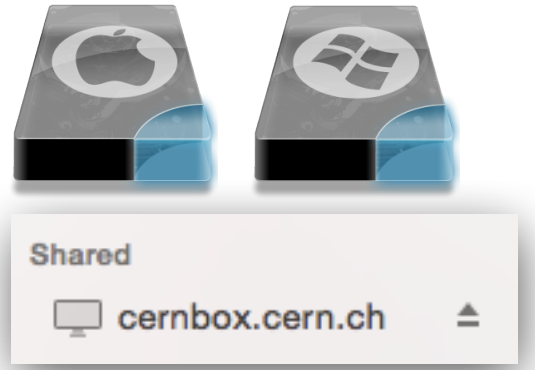


Web Access

Mobile App



Sync Client



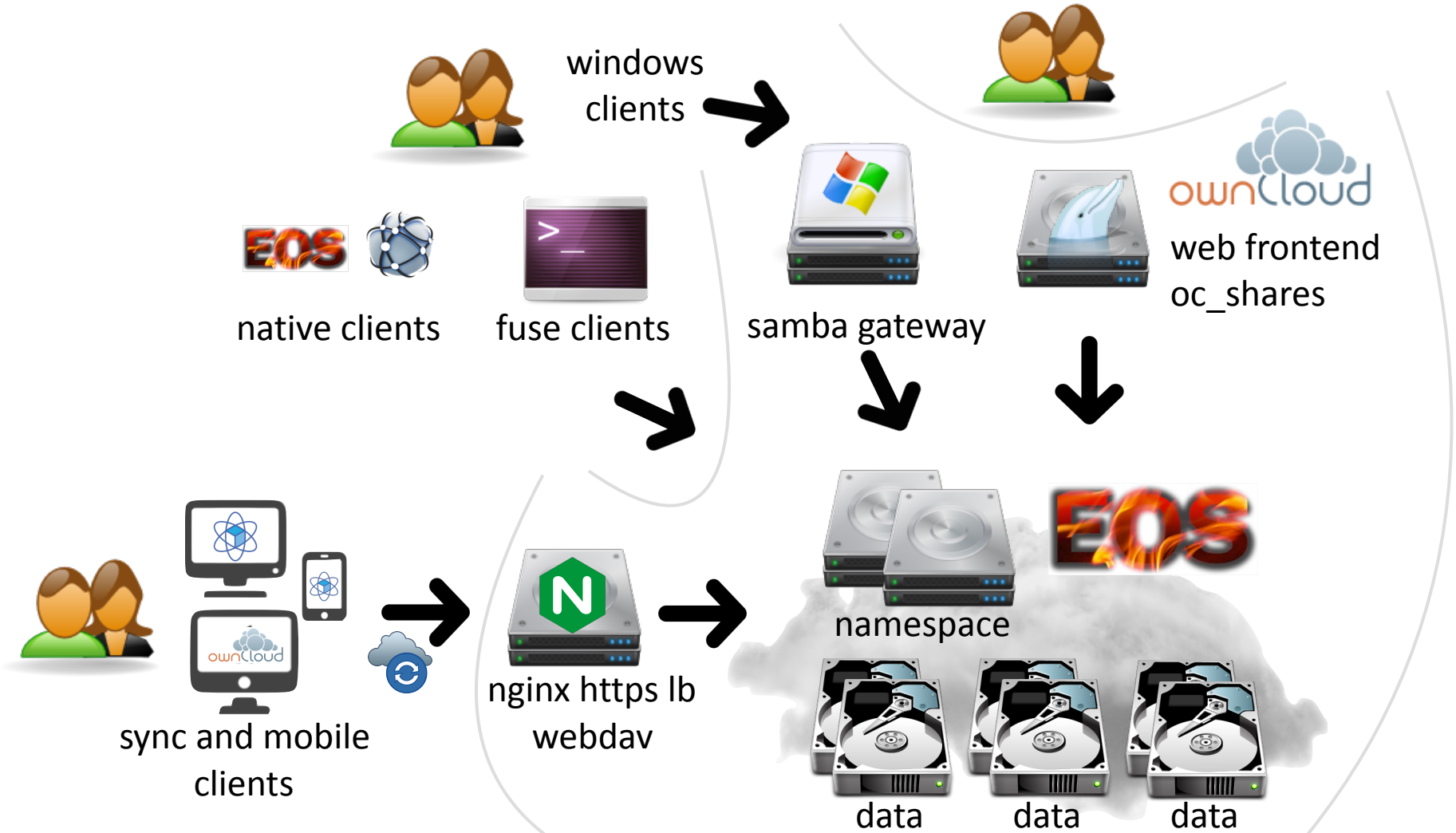
WebDAV

SAMBA

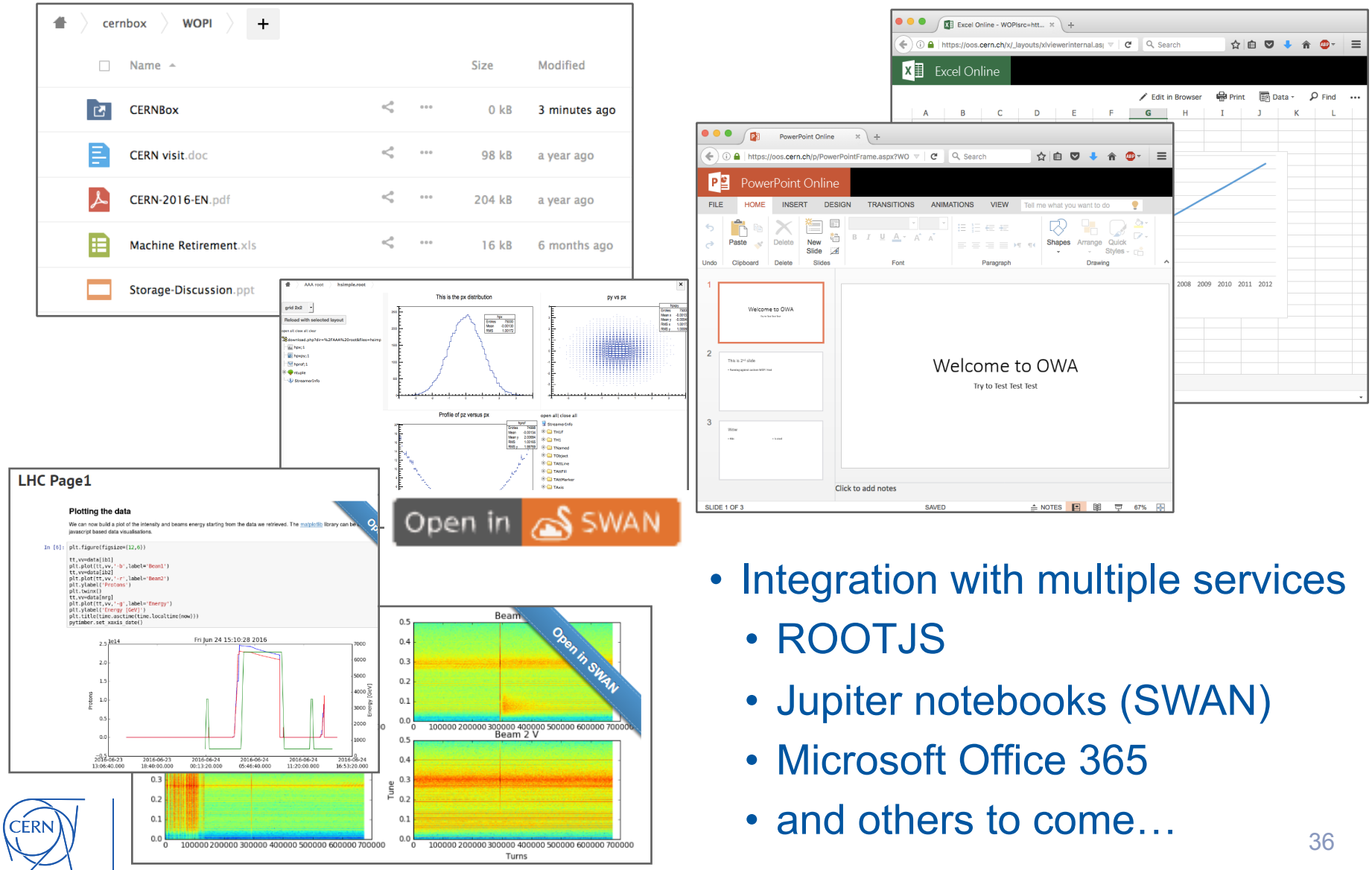
FUSE

Directly from the storage backend (xroot, http, s3, …)

CERNBox Architecture

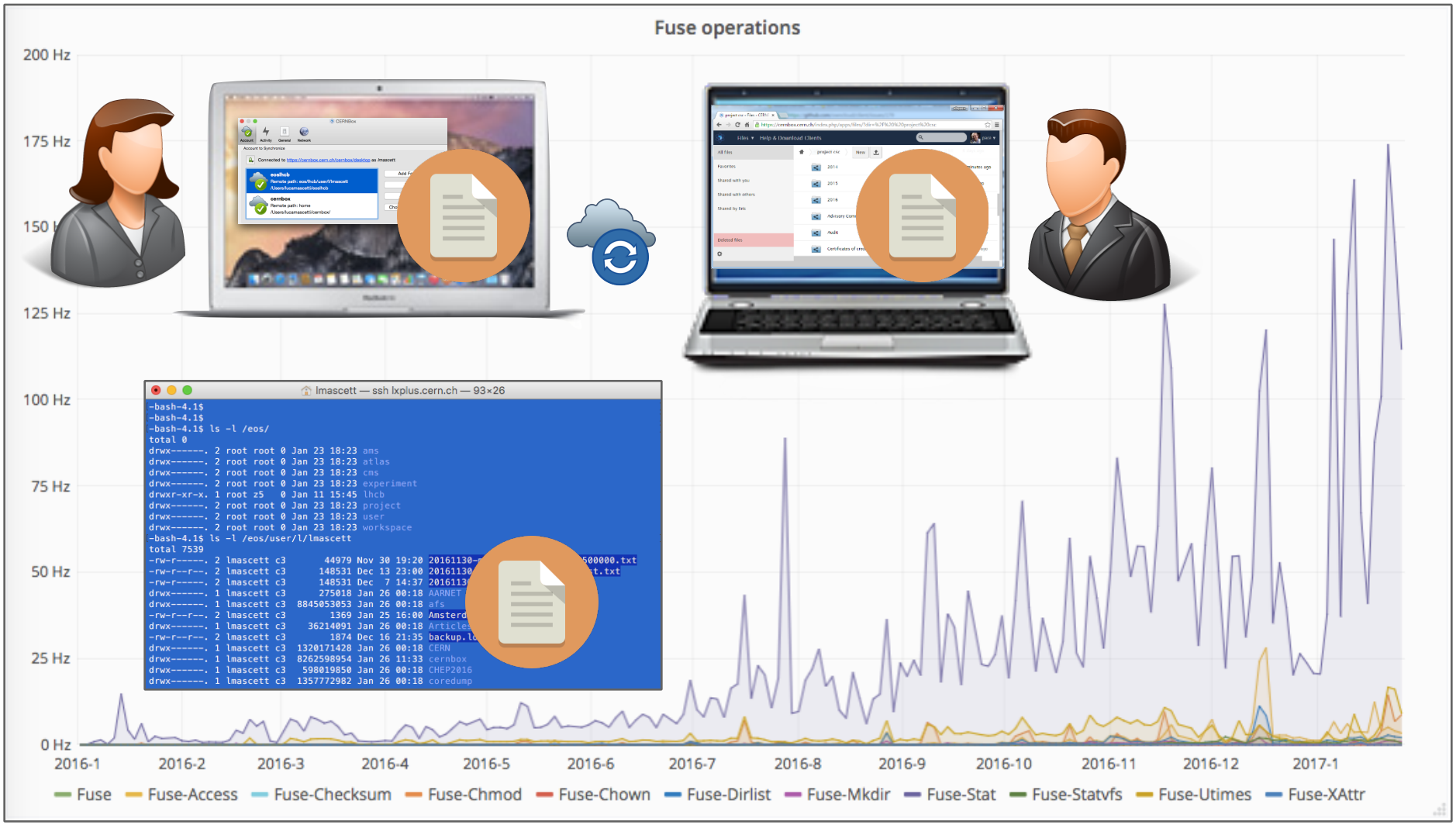


CERNBox as main entry-point



- Integration with multiple services
  - ROOTJS
  - Jupiter notebooks (SWAN)
  - Microsoft Office 365
  - and others to come…

Filesystem Access (FUSE)



Python Notebook Integration (SWAN)

Files | Help & Download Clients | Luca Mascetti (lmascett)

All files | SWAN > swanExamples > notebooks

Favorites
Shared with you
Shared with others
Shared by link
All projects
Your projects
Eos Browser

Name	Size	Modified
.~FastjetInteractive.ipynb	53 kB	8 months ago
CMSDimuon_py.ipynb	341 kB	9 months ago
CppFromPython_pycpp.ipynb	7 kB	9 months ago
FastjetInteractive.ipynb	53 kB	8 months ago
Geometry_Visualisation_cpp.ipynb	598 kB	8 months ago
pandas.ipynb	94 kB	9 months ago
Simple_ROOTbook_cpp.ipynb	165 kB	8 months ago
	325 kB	9 months ago
	169 kB	3 months ago
	165 kB	9 months ago
	4 kB	9 months ago
	1.9 MB	

https://swan001.cern.ch/user/lmascett/notebooks/SWAN/swanExamples/n

CMSDimuon_py (autosaved) | Control Panel | Logout

File Edit View Insert Cell Kernel Help | Python 2

Markdown | CellToolbar

Dimuon spectrum

This ROOTbook produces a plot of the dimuon spectrum starting from a subset of the CMS collision events of Run2010B.

Dataset Reference:
McCauley, T. (2014). Dimuon event information derived from the Run2010B public Mu dataset. CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.CB8H.MFFA.

In [1]:
```
import ROOT
```

```
Welcome to JupyROOT 6.07/06
```

A little extra: JavaScript visualisation. This command will become a magic very soon.

In [2]:
```
%jsroot on
```

Convert to ROOT format and analyse

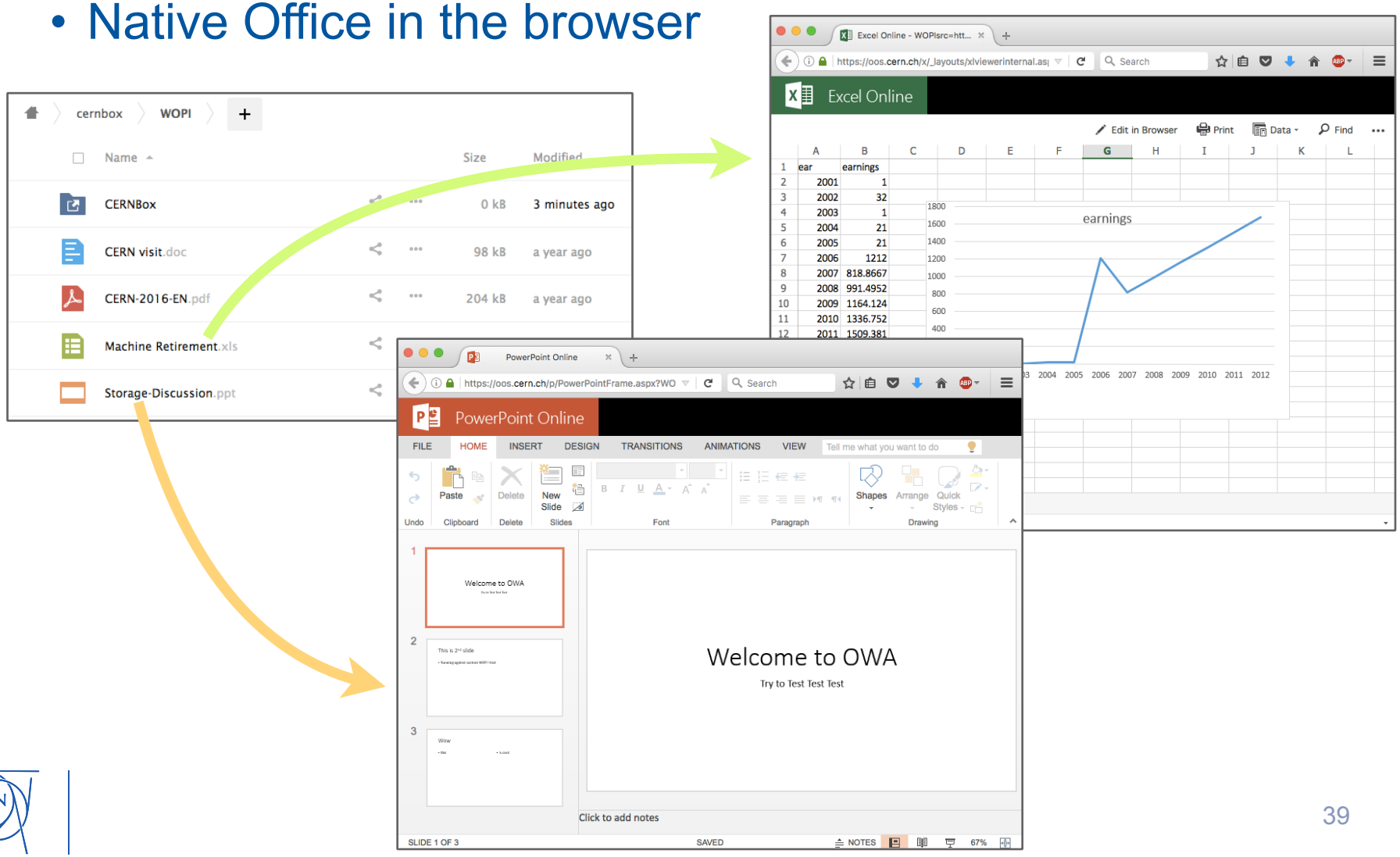
First of all we convert the csv file into ROOT format, i.e. filling up a TTree data structure. But first of all we uncompress it if it's not.

In [3]:
```
inputFileName = '../data/MuRun2010B.csv'
import os
if not os.path.exists(inputFileName):
    import gzip
    import shutil
    with gzip.open(inputFileName+'.gz', 'rb') as f_in, open(inputFileName, 'wb') as f_out:
        shutil.copyfileobj(f_in, f_out)
```

Open in SWAN

Office Online Integration (WOPI)

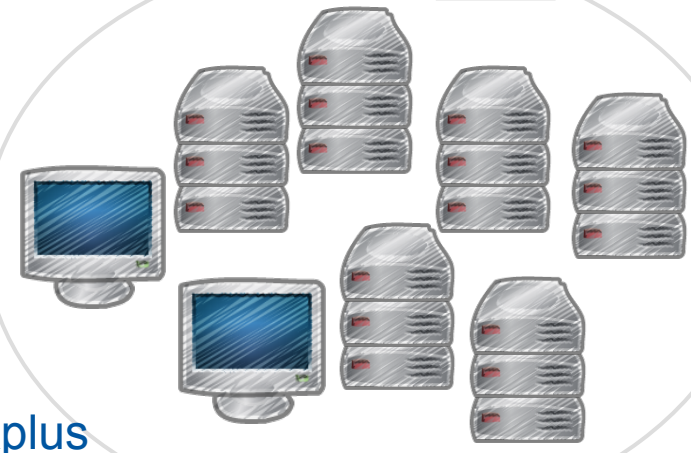
- Web Application Open Platform Interface (WOPI) protocol
  - Native Office in the browser



Generic HOME Directory Service

$HOME Current Status

AFS



lxplus
lxbatch
linux managed

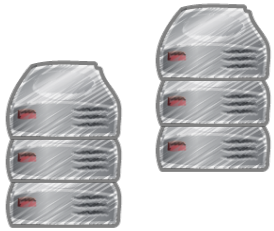
DFS

Terminal Servers
Windows managed

localdisk

majority of users' devices

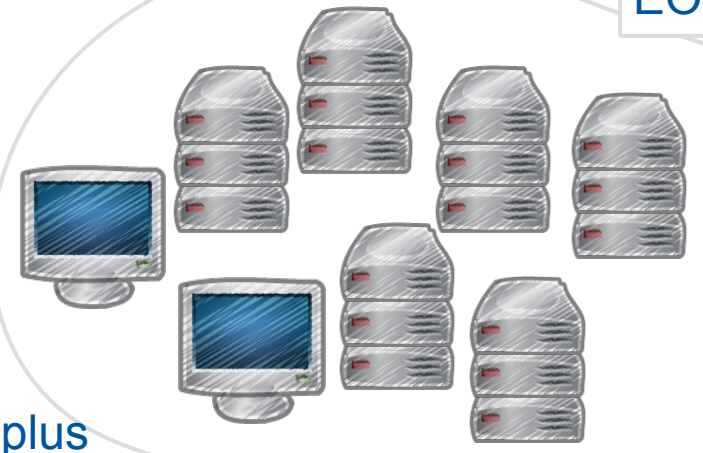
NFS



small private/experiment clusters

$HOME Future Vision

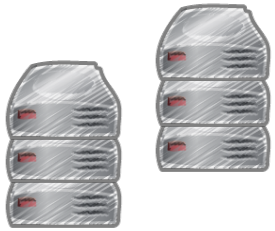
EOS\CERNBox



lxplus
lxbatch
linux managed

Terminal Servers
Windows managed

majority of users' devices



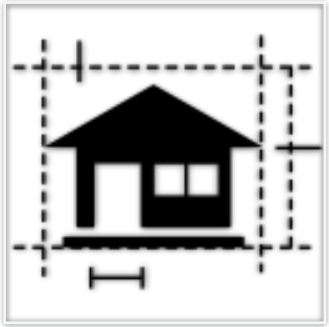
small private/experiment clusters

Trigger

Current AFS Status

- Slow decline of AFS
- Concern for the long-term viability of OpenAFS
  - Future highly visible issues (IPv6, DES)
  - Increasing issues for client support
    - Data corruption or low performance on latest Linux versions
- 2015 IT Dept. level decision phaseout AFS
  - EOS FUSE as replacement strategy

CERN $HOME



```
[jdoe@lxplus066 ~]$
[jdoe@lxplus066 ~]$ pwd
/eos/user/j/jdoe
```

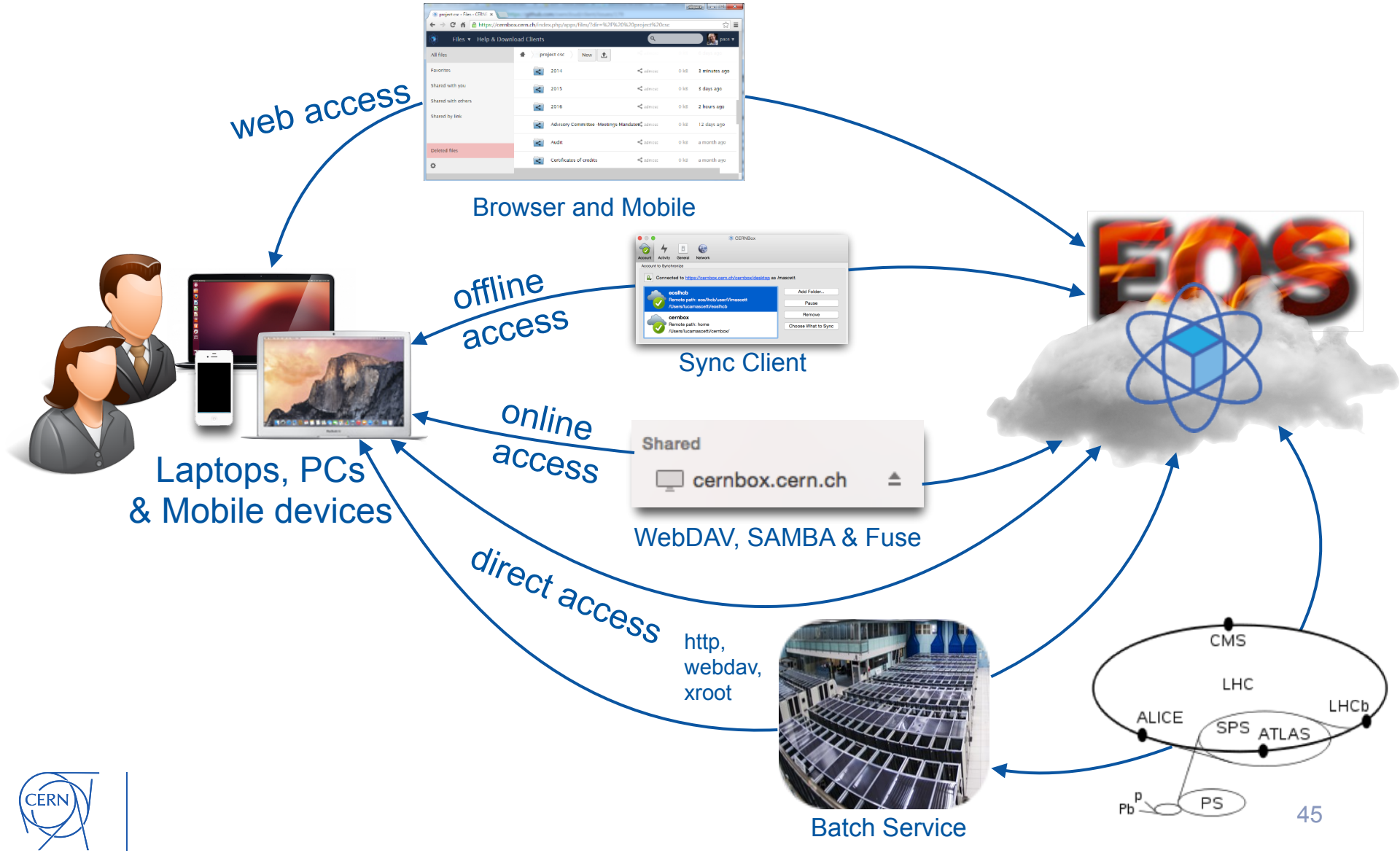
- Strategic choice to use EOS\CERNBox
  - flexible and powerful storage infrastructure
    - master the full storage development
  - supported by existing user base (> 9k users)

- Same HOME for multiple operating systems
  - enable synchronisation and disconnected access
  - improve user experience (no split-brain home)

- Serve with the same technology multiple community
  - Avoid to run separate/isolate storage cluster
  - better system interoperability and profit from future synergy (DFS)

$HOME => **`/eos/user/<u>/<username>/`**

CERN $HOME


web access
Browser and Mobile
offline access
Sync Client
online access
Shared
cernbox.cern.ch
WebDAV, SAMBA & Fuse
direct access
http, webdav, xroot
Batch Service
Laptops, PCs & Mobile devices
EOS
CMS
LHC
LHCb
ALICE
SPS
ATLAS
PS
Pb
p

Desktop/Laptop

Each directory is a Sync Folder for CERNBox

```
Desktop   -> /eos/user/j/jdoe/Desktop
Documents -> /eos/user/j/jdoe/Documents
Pictures  -> /eos/user/j/jdoe/Pictures
Music     -> /eos/user/j/jdoe/Music
Videos    -> /eos/user/j/jdoe/Videos
```

Clusters

lxplus, lxbatch and terminals

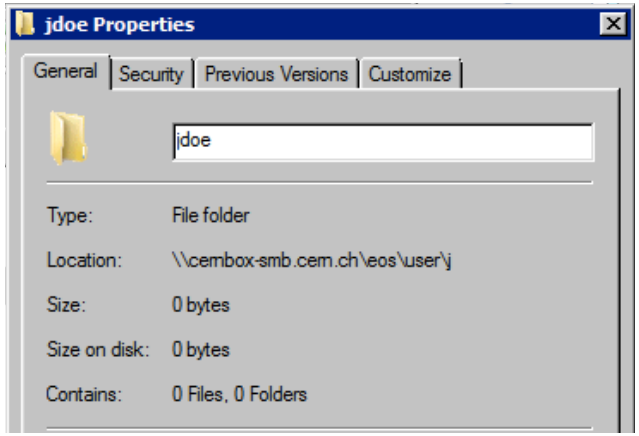
$HOME is set to /eos/user/j/jdoe/
prototype lxplus-eos.cern.ch available soon

Cluster Desktops

If needed cluster desktop could be configured with $HOME in /eos/user/j/jdoe/ like lxplus

Windows Terminal Server

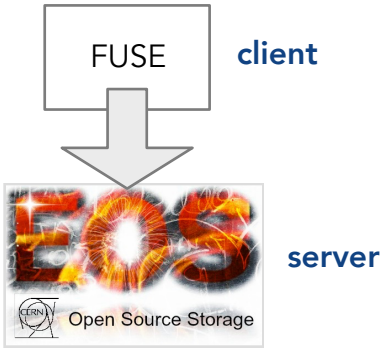
File access is via SAMBA



EOS FUSE Current Status

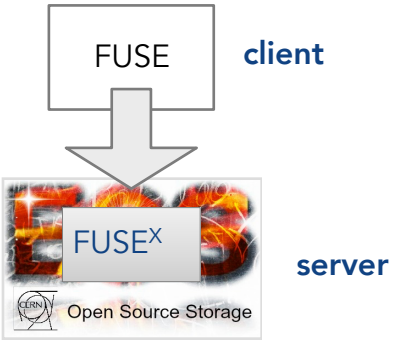
- /eos already mounted on lxplus and lxbatch
  - adding features + bugfix during the last months
- Most of the issues will be solved with the new eos fuse client
  - decision to start with a clean rewrite (EOSFUSEng)
- Behaviour will be similar to a local filesystem

V2 implementation



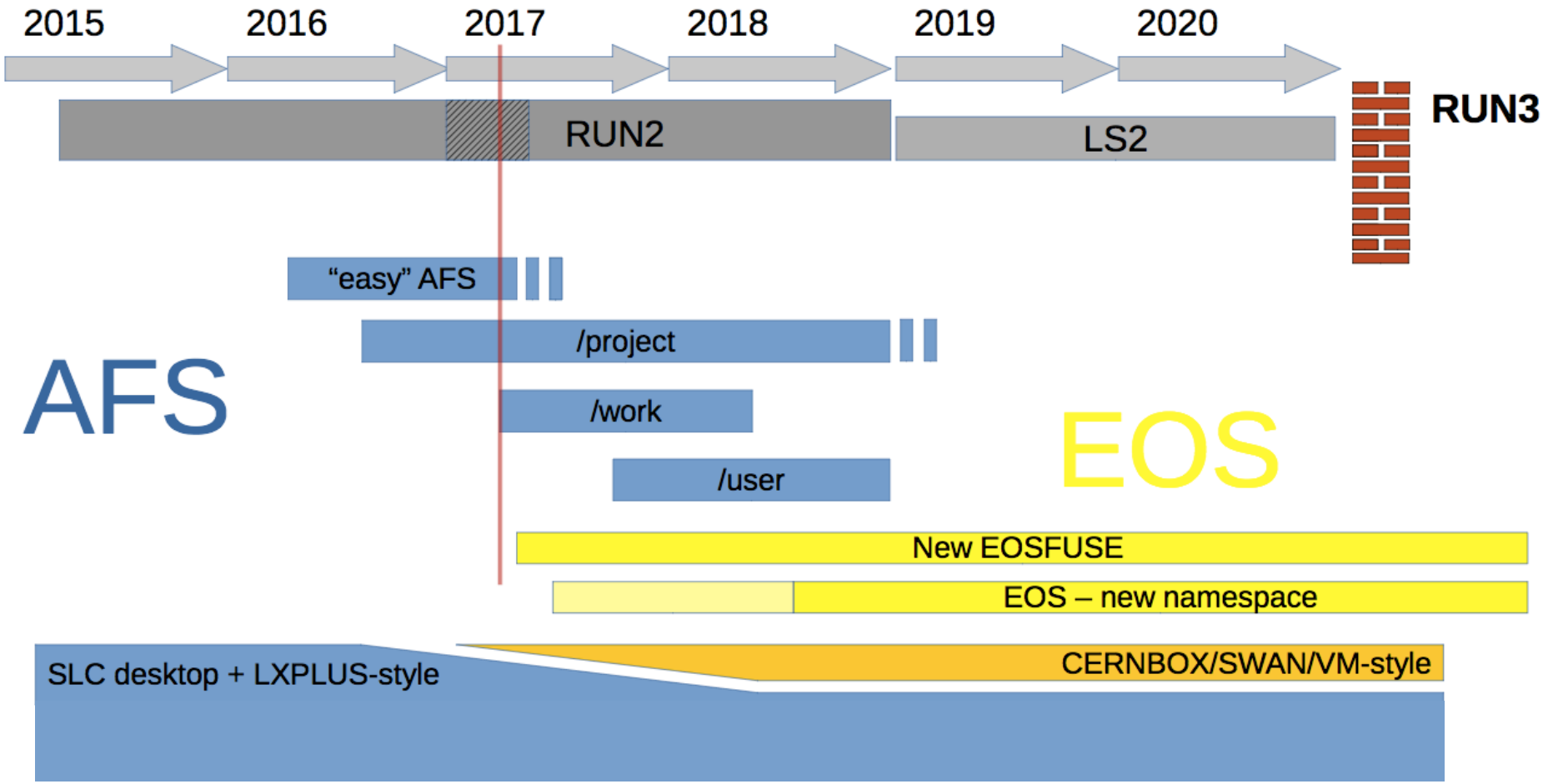
FUSE filesystem implemented as **pure client side** application without dedicated server side support.

V3 implementation



Dedicated server-side support providing a fully asynchronous server->client communication, leases, locks, file inlining, local meta-data and data caching

AFS Migration Plan

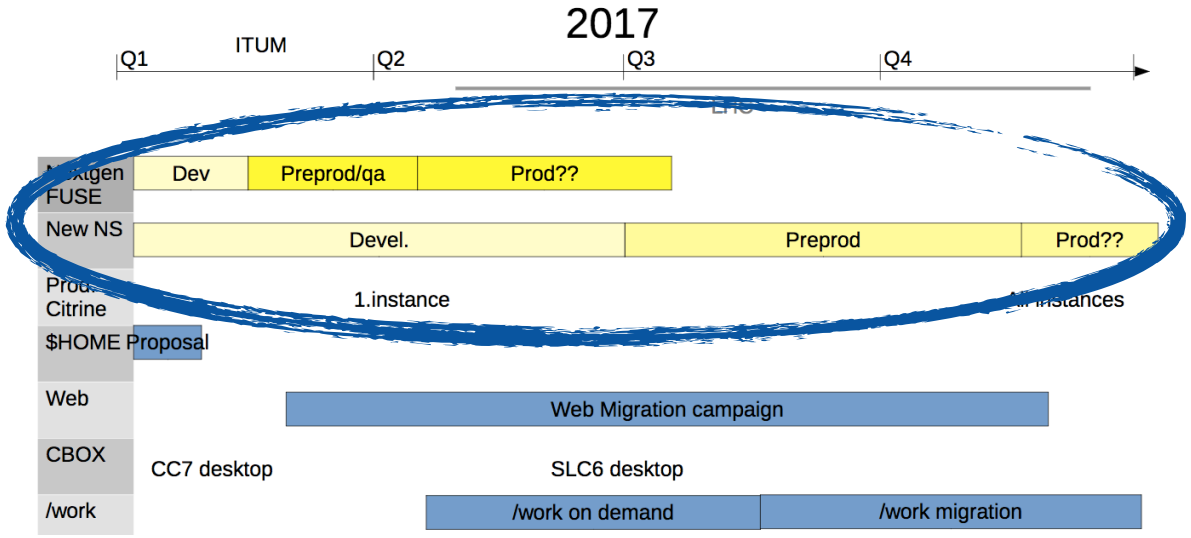


Major Milestones - Roadmap

- FUSE performance and localFS behaviour
  - development and deployment of **EOSFUSEng**
    - **lxplus-eos.cern.ch** testbed
- Namespace scalability (> 1 Billion Files)
  - Boost booting performance (x4 achieved)
  - New EOS server major upgrade

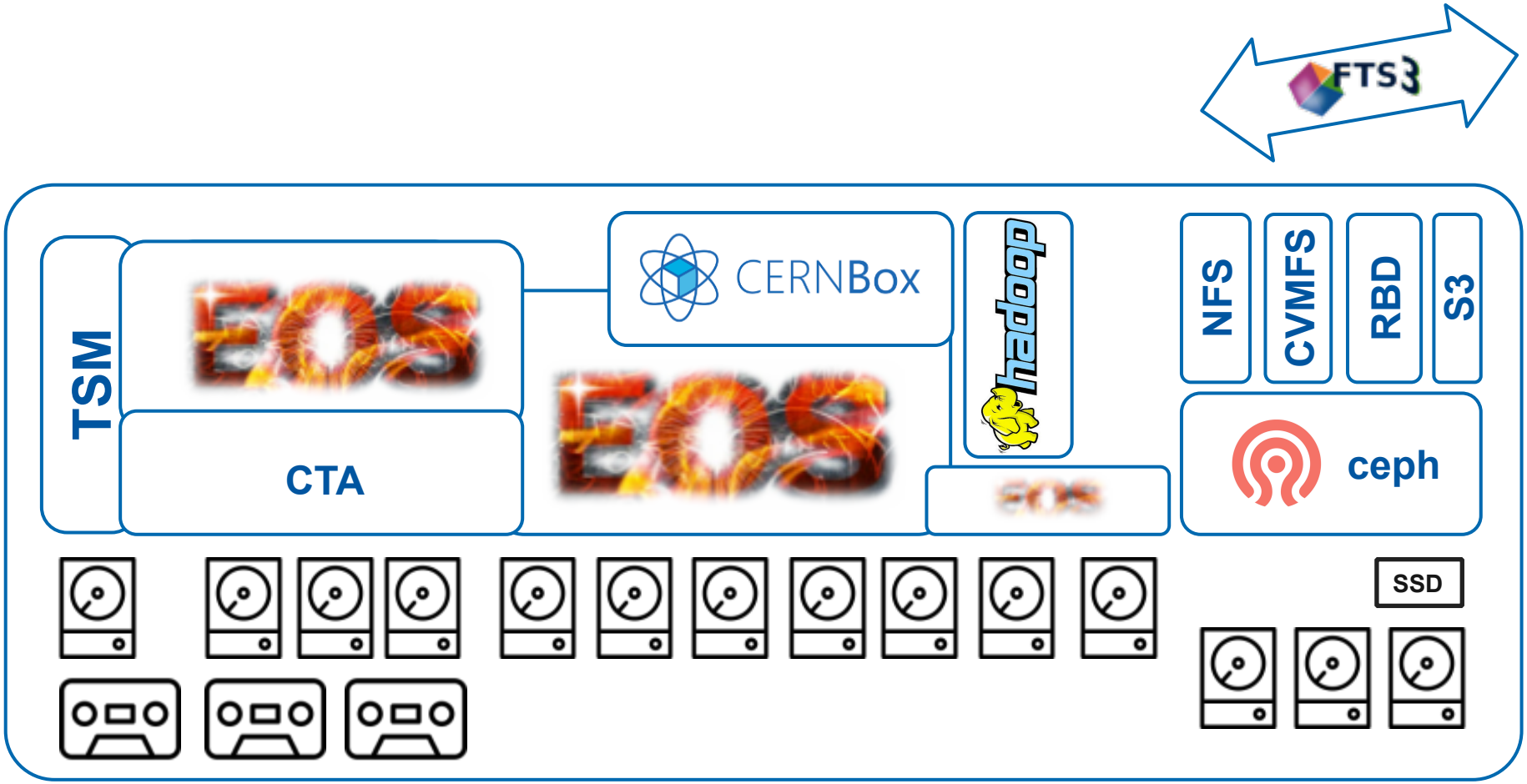
Deployed in few instances

Scheduled on May 15th for eoslhcb



Summary and Outlook

Storage Services Evolution



EOS Development Strategy

Evolve **flexible** storage infrastructure as ***scale-out solution*** towards exabyte scale

- one storage – many views on data – analysis/analytics service integration
- suitable for LAN & WAN deployments
- enable cost optimized deployments

Evolve **file-system** interface /eos

- multi-platform access
  - using standard protocol bridges WebDAV, CIFS, NFS, HTTP

Evolve **namespace scalability**

- scale-out implementation

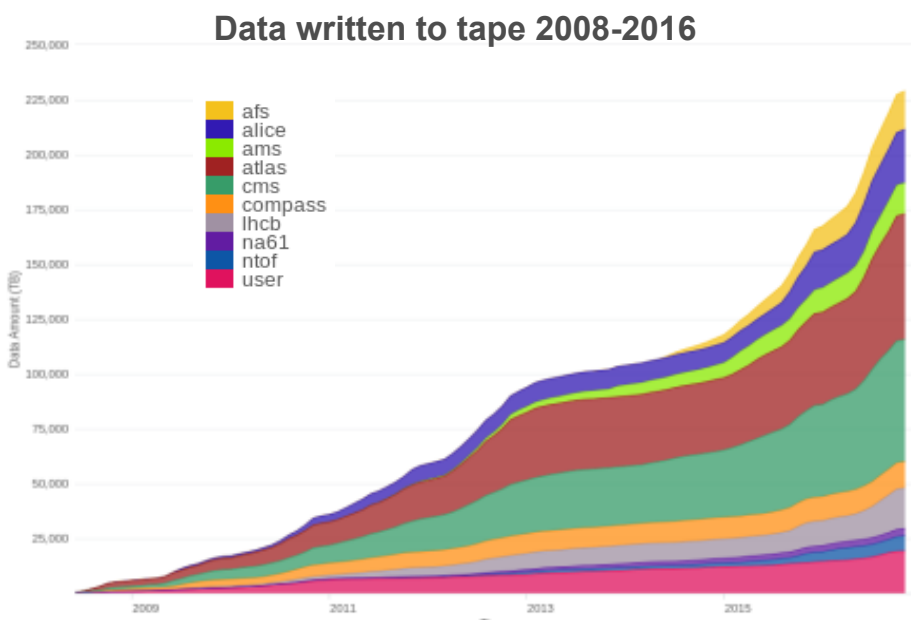
Evolve scale-out **storage back-end** support

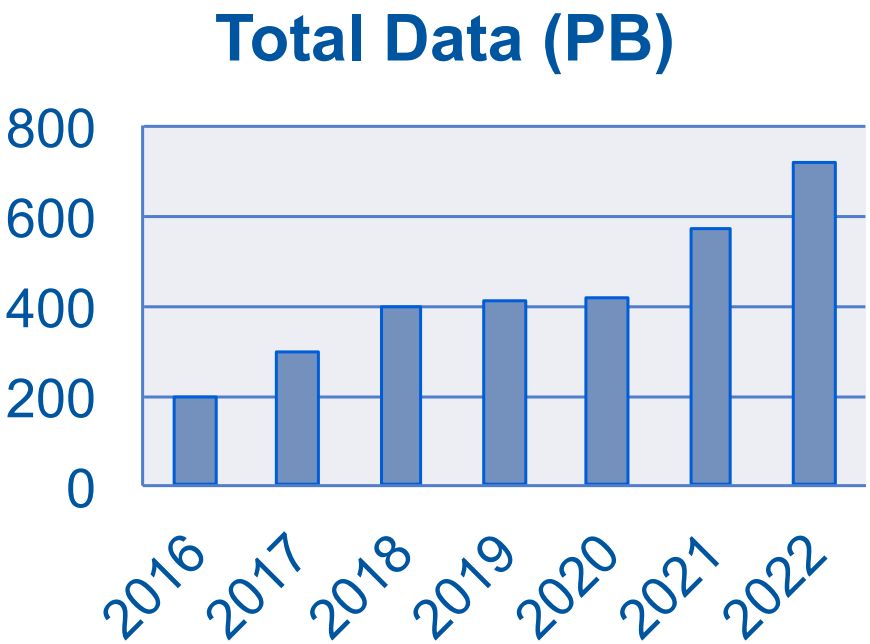
- object disks, object stores, public cloud, cold storage

Data Archiving evolution

Preserving the past … and anticipating the future

- 2017-2018: 100PB/year
- 2021++ : 150PB/year





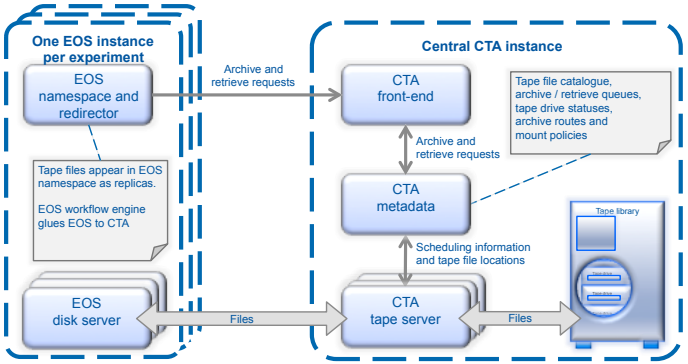
Data Archiving Evolution: CTA

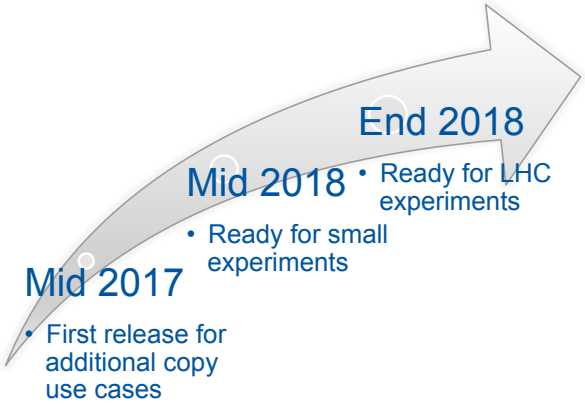
EOS + tape …

- EOS is the strategic storage platform
- Tape is the strategic long term archive medium

EOS + tape = ❤️

- Meet **CTA** : the **CERN Tape Archive**
- Tape backend for EOS
- Streamline data paths, software and infrastructure





EOS plus CTA is a "drop in" replacement for CASTOR

- **Same tape format as CASTOR – only need to migrate metadata**

Summary and Outlook

EOS provides a **very flexible** platform for large communities
- integrated in Tier-0 workflow by ATLAS & CMS
- more than 9k users storing data today via **CERNBox**

Demonstrated unprecedented scalability
- largest **low-cost** HEP storage installation site today

CERNBox as an extension of the Desktop
- Bring data closer to our users
- New ways to interact with the data
- Ongoing integration with multiple services

Strategic direction for CERN based disk storage
- for physics data (user/group/grid)
- as generic home directory via **EOS fuse** and **CERNBox**

- **CERN Tape Archive (CTA)**
  - tape backend for EOS
  - strategic long term archive medium

Thanks for the attention!

www.cern.ch

Questions?