



EOS Open Storage

the CERN storage ecosystem for
scientific data repositories



Dr. **Andreas-Joachim Peters**
for the EOS project
CERN IT-ST

International Symposium on Grids and Clouds 2018
in conjunction with
Frontiers in Computational Drug Discovery

16-23 March 2018
Academia Sinica, Taipei, Taiwan





Overview

- Introduction
- EOS at CERN and elsewhere
- Tapes, Clouds & Lakes
- Scientific Service Bundle
- EOS as a filesystem
- Vision, Summary & Outlook

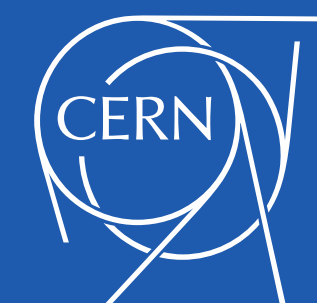




Everything about EOS



<http://eos.cern.ch>



Disclaimer: this presentation skips many interesting aspects of the core development work and focus on few specific aspects.



Introduction

What is EOS?

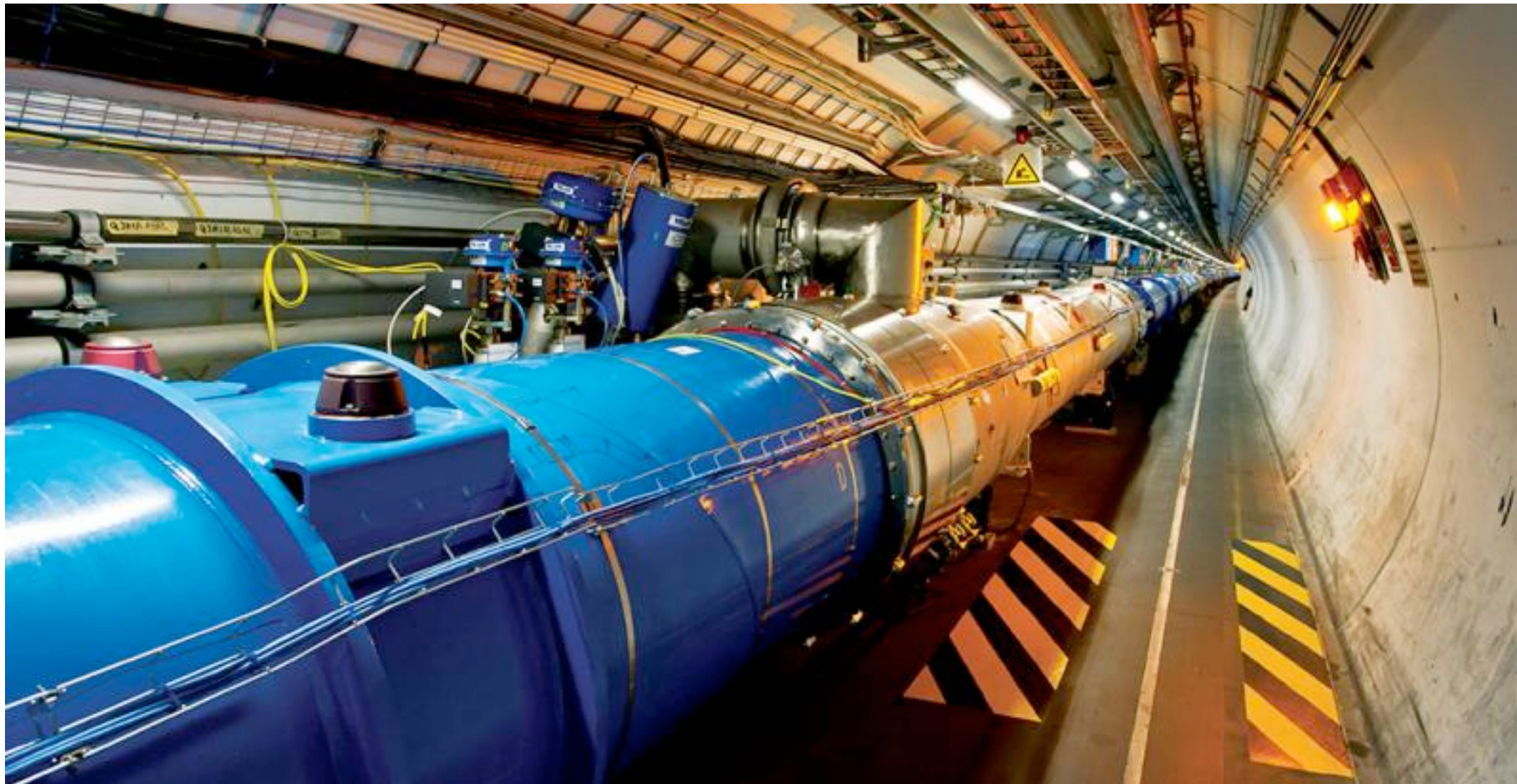
EOS is a storage software solution for

- central data recording
- user analysis
- data processing



Introduction

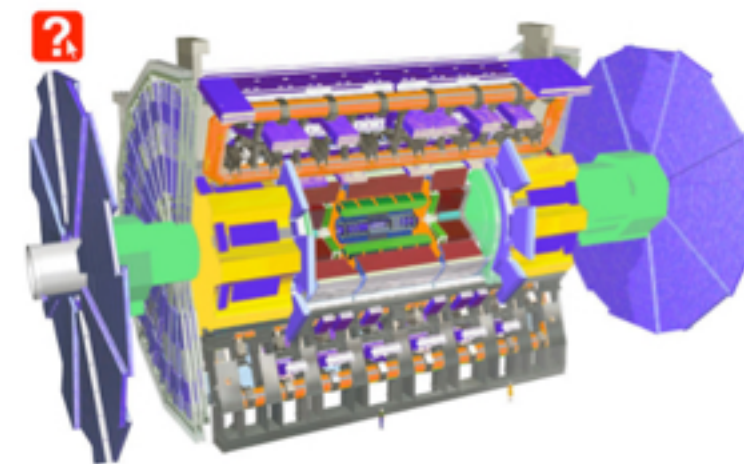
EOS and the Large Hadron Collider LHC



CERN mainstream use case

tape archive **CASTOR**
CERN Advanced STORage manager

LHC Detector



O(GB/s)



5-10 GB/s

5-10 GB/s



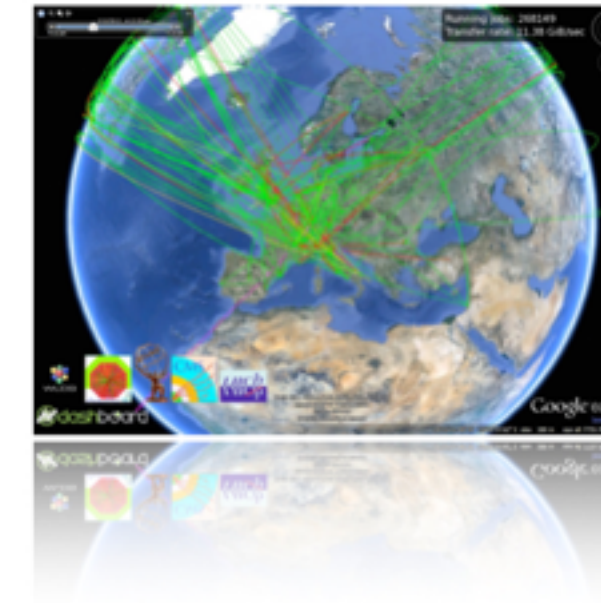
peak 100 GB/s

local batch cluster
O(10⁵ cores)



openstack™

Data Export to Worldwide Computing Grid

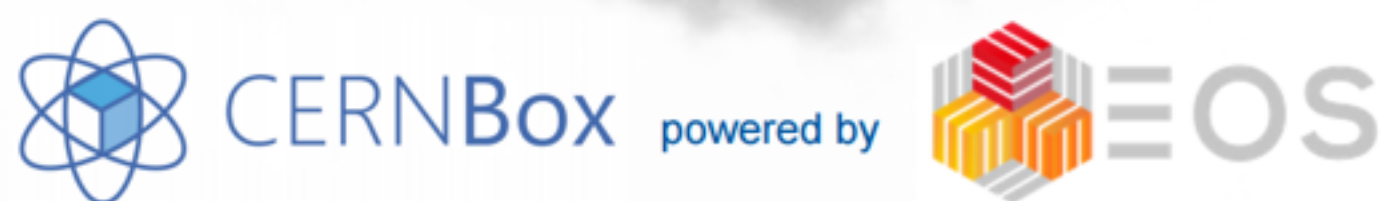
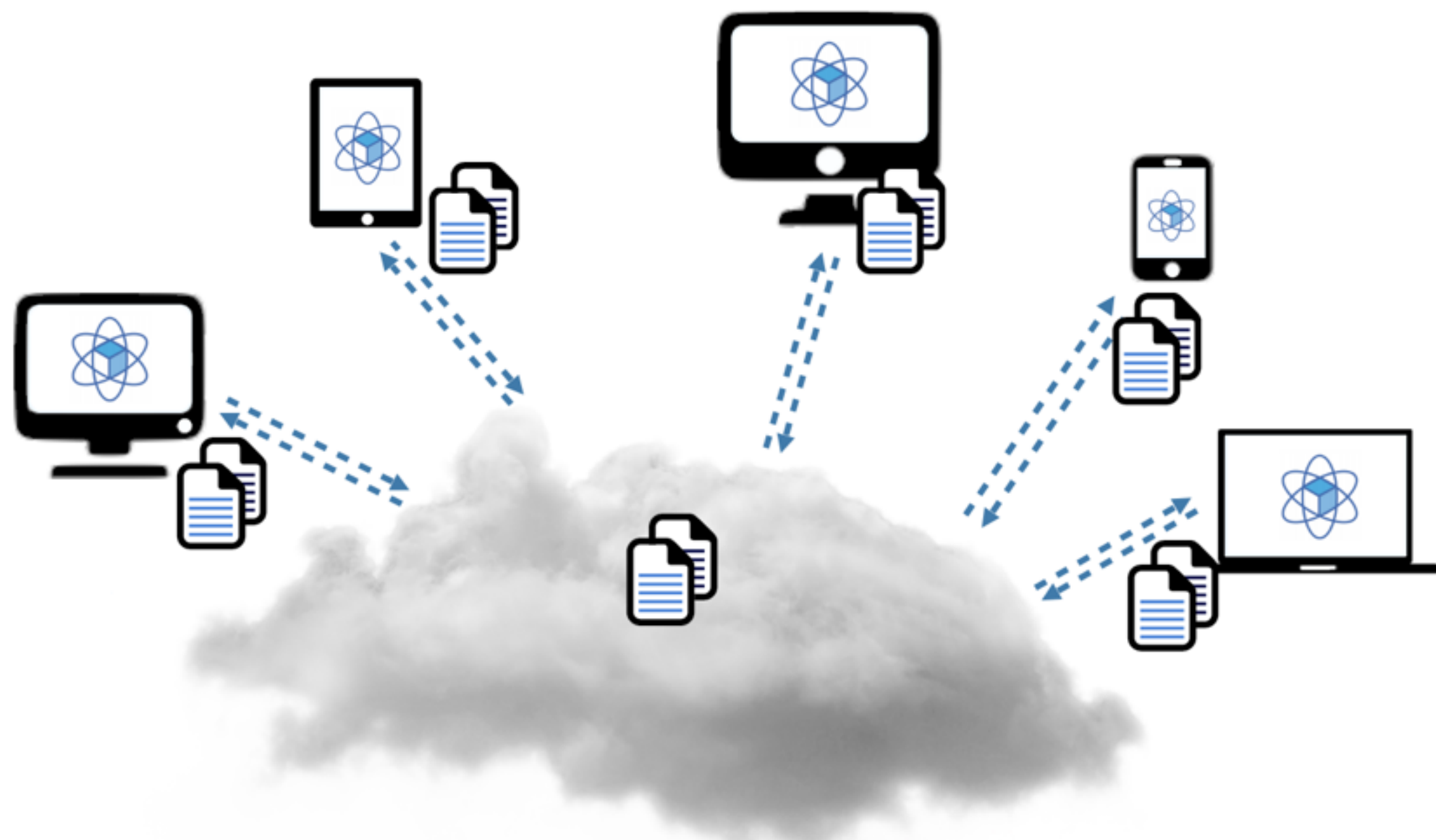




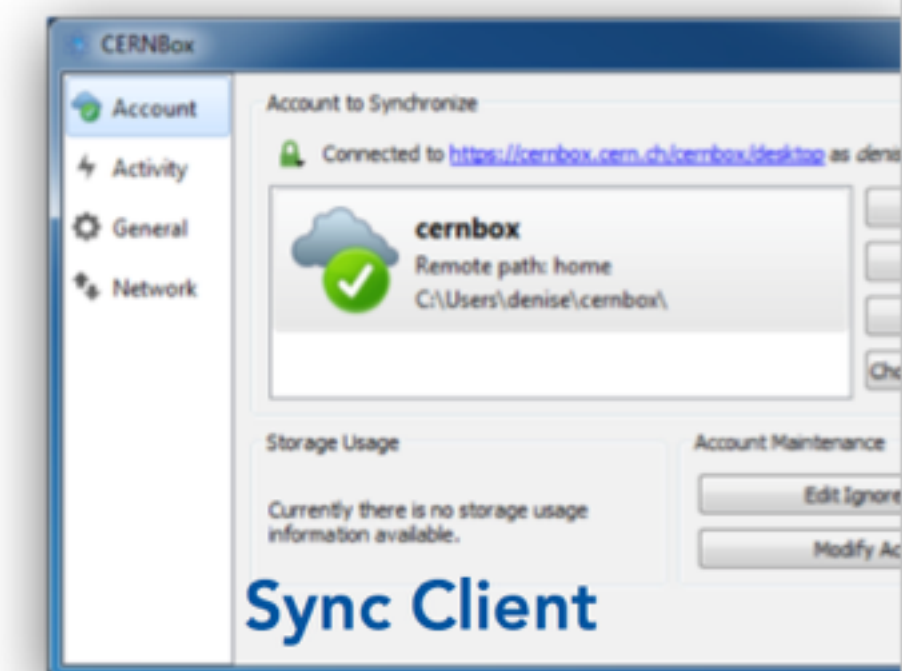
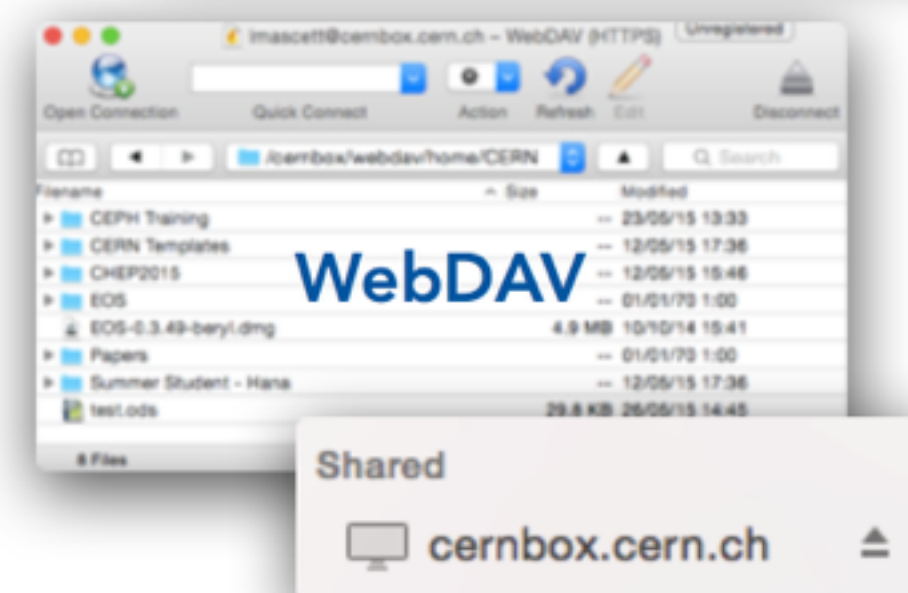
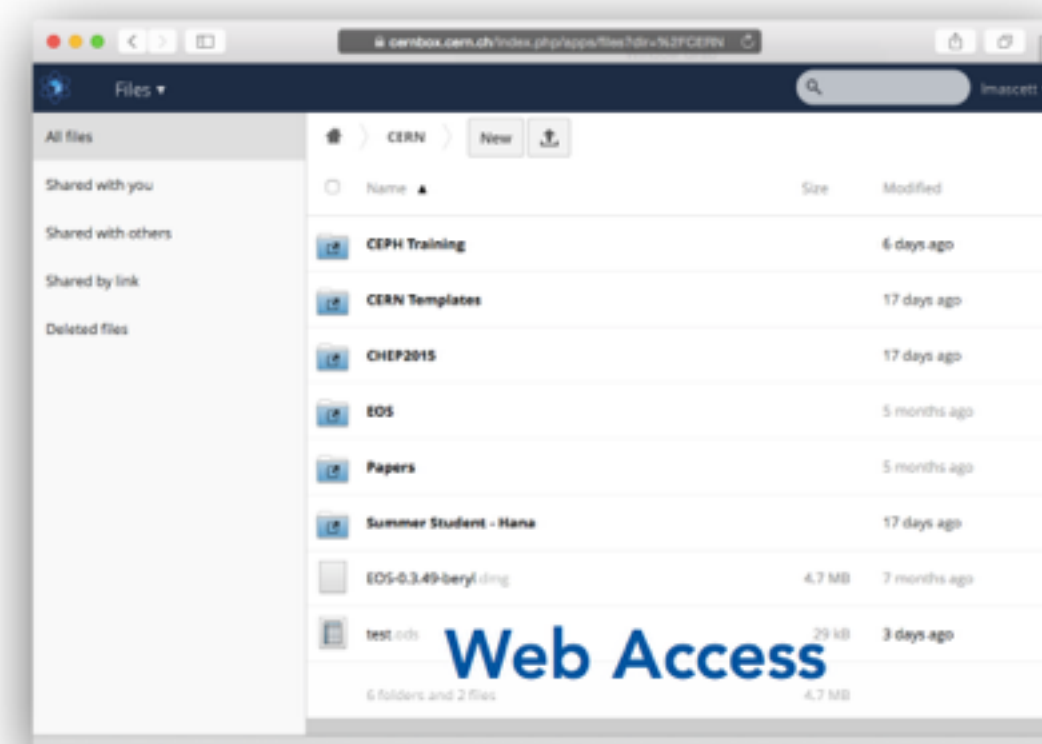
Introduction

EOS and CERNBox

Sync & Share Platform
with collaborative editing



Available Access Methods



EOS
Directly from the
storage backend
EOSUSER
(xroot, http, s3, ...)



Architecture

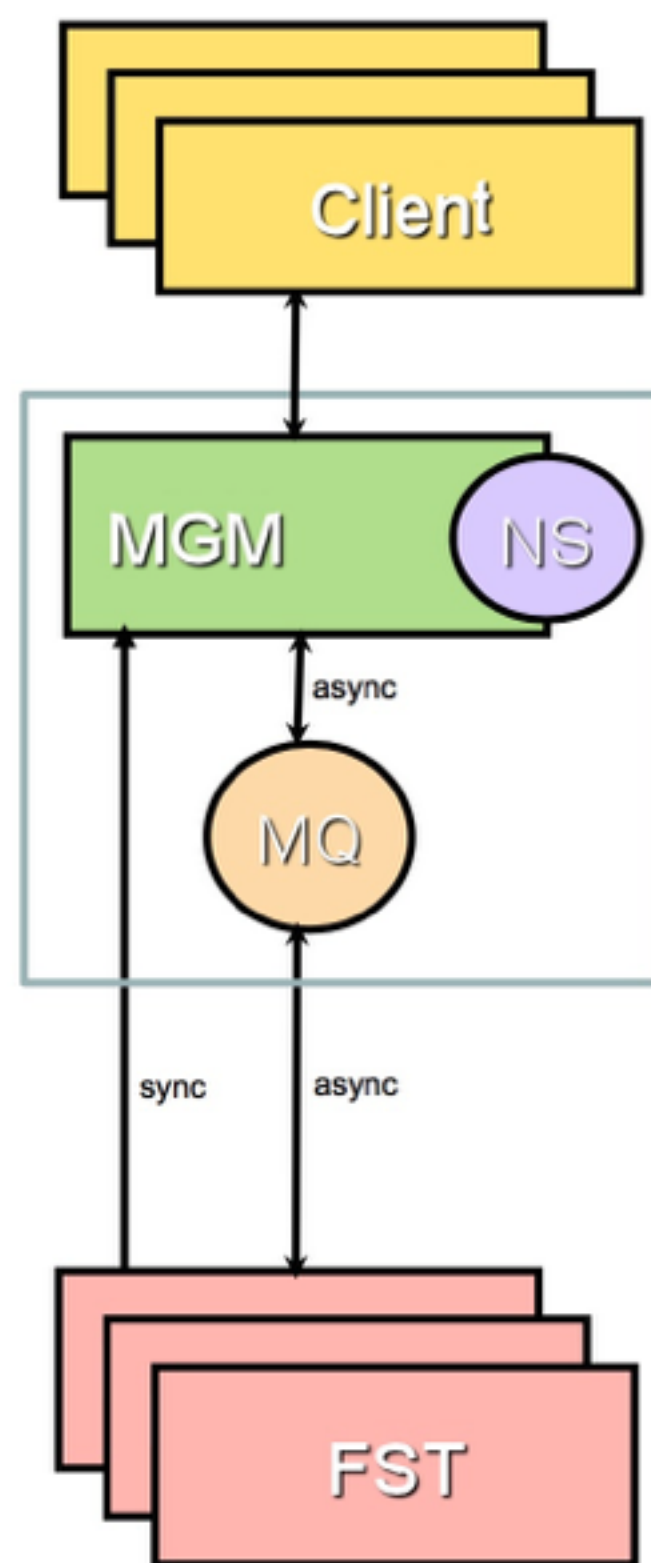


Storage Clients:
Browser, Applications, Mounts

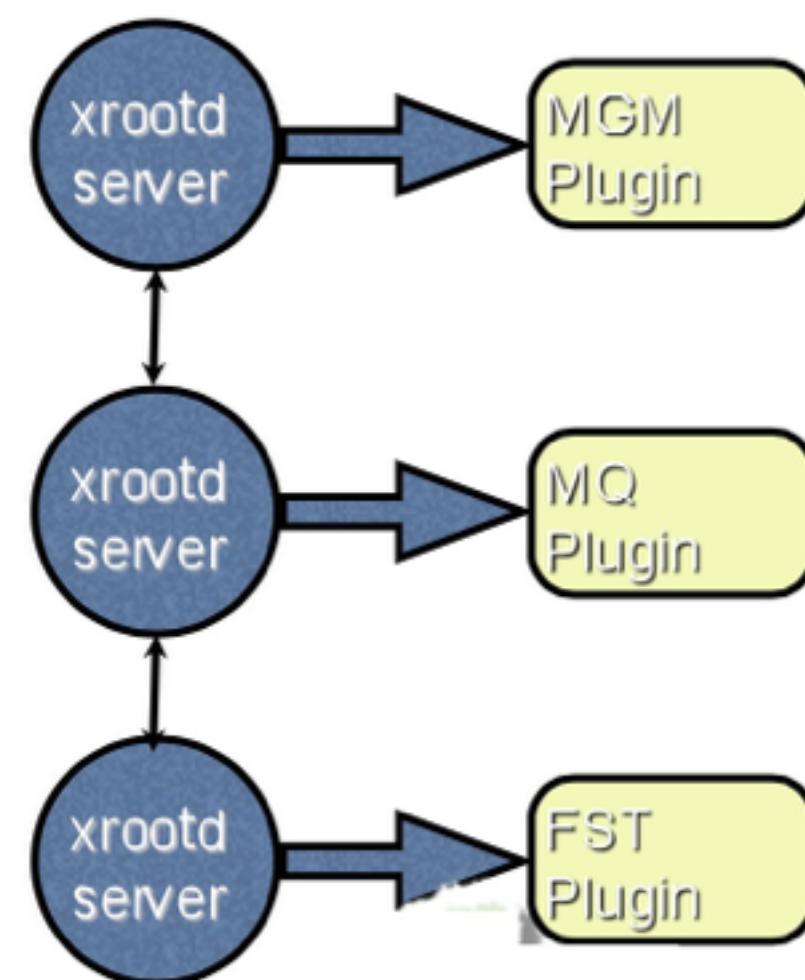
Meta Data Service / Namespace

Asynchronous Messaging Service

Data Storage



Implemented as plugins in **xrootd**



EOS is implemented in C++ using the XRootD framework

XRootD provides a client/server protocol which is tailored for data access

- third party transfer
- WAN latency compensation using vectored read requests
- pluggable authentication framework
- ...



Architecture Transition 2017/18

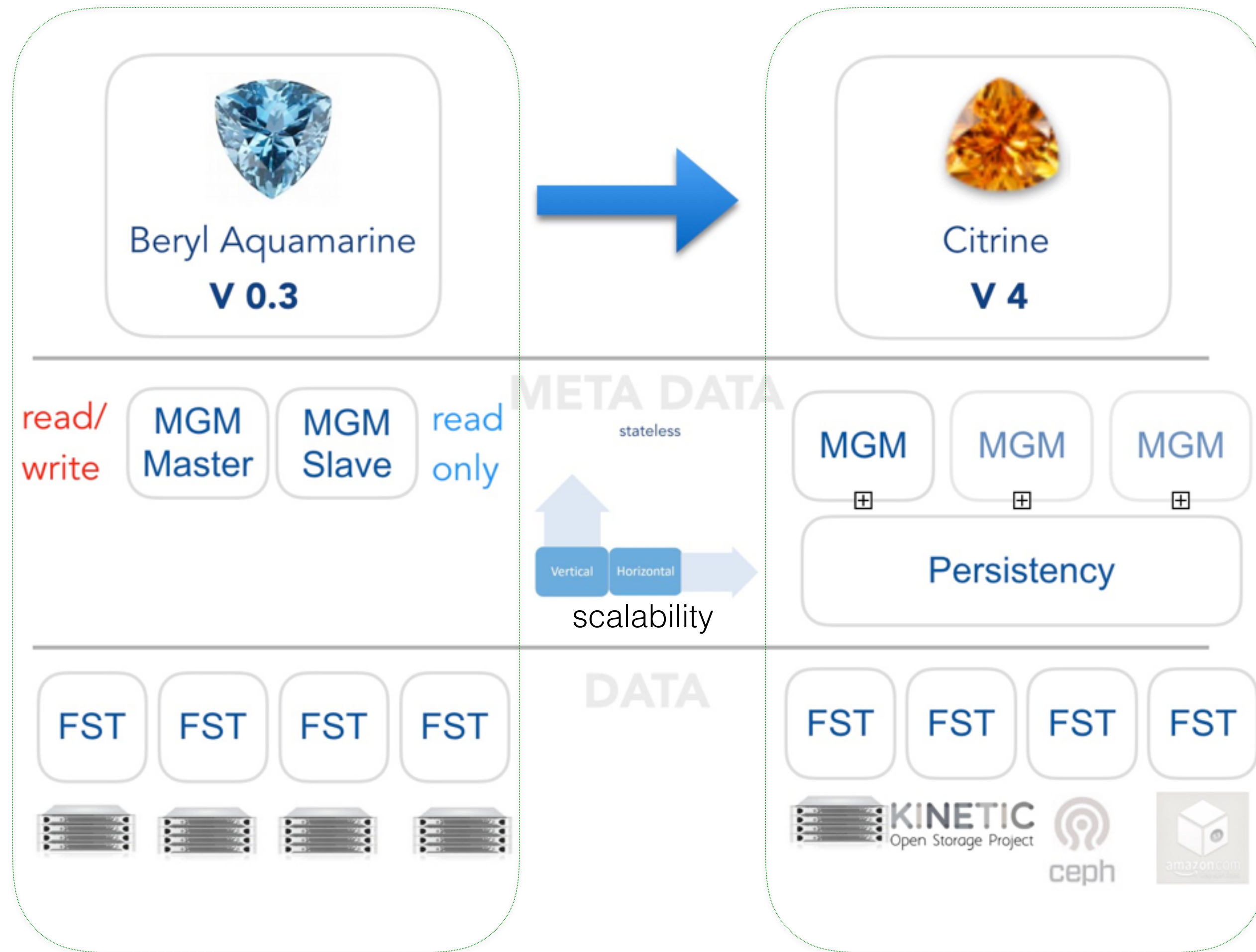
EOS releases are named after **gemstones**

AQUAMARINE Version

- production <= 2017
- in-memory namespace

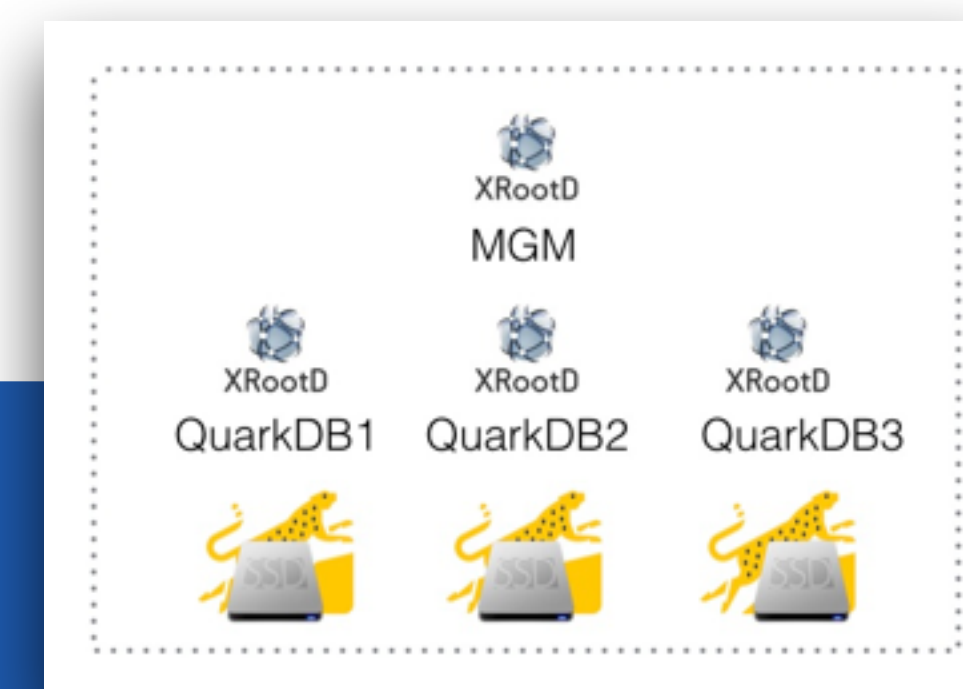
CITRINE Version

- production >=2017
- in-memory & scale-out KV persistency



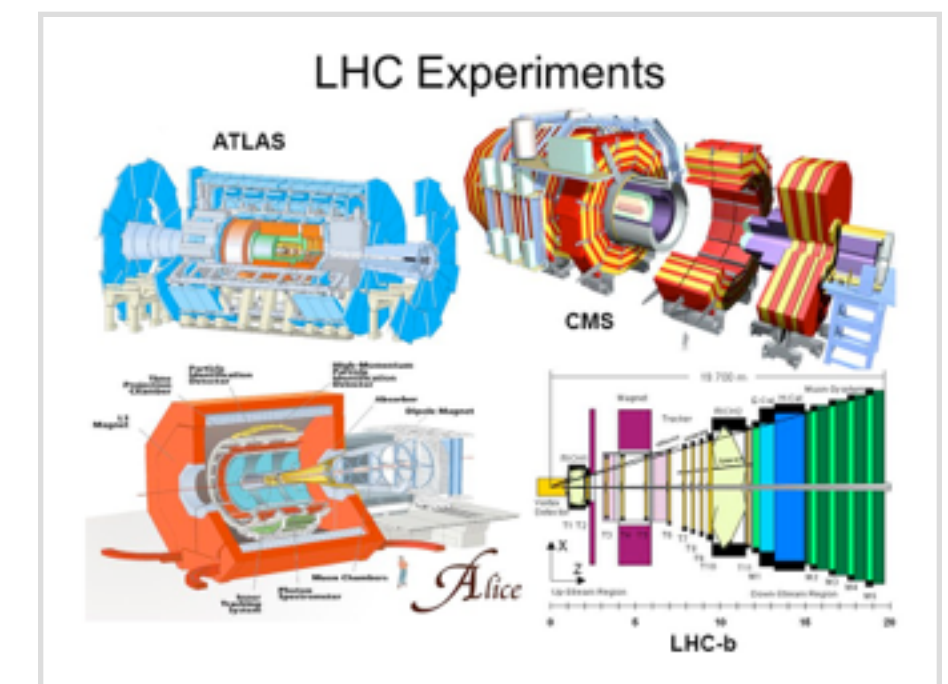
- during 2017 CERN services exceeded design limits - lower service availability

- leading effort to commission new architecture in 2018 with namespace cache in-memory & KV store persistency in **QuarkDB**





EOS at CERN



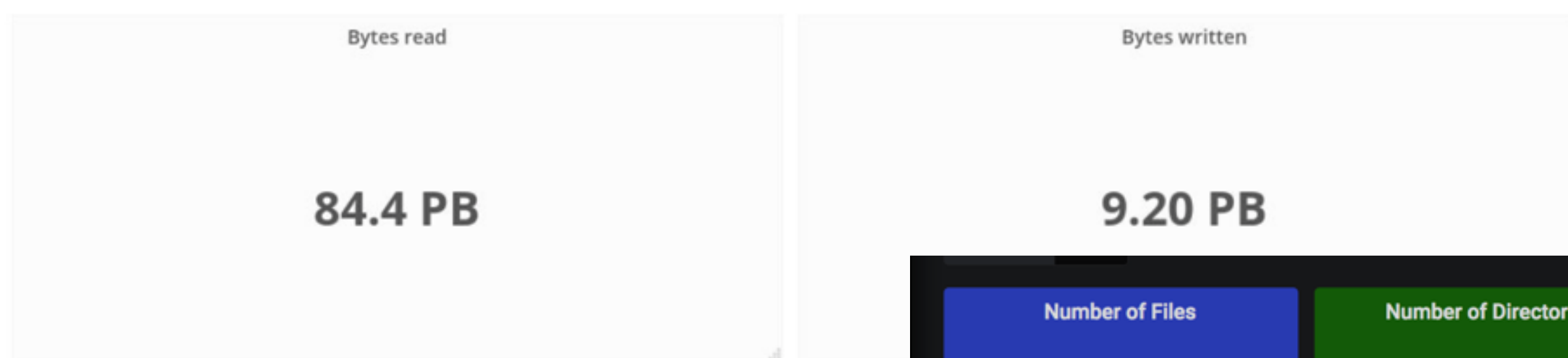
	2017	2018
Nodes	~1200	~1400
Disks	~40000	~50000
Raw capacity	~150PB	~250PB



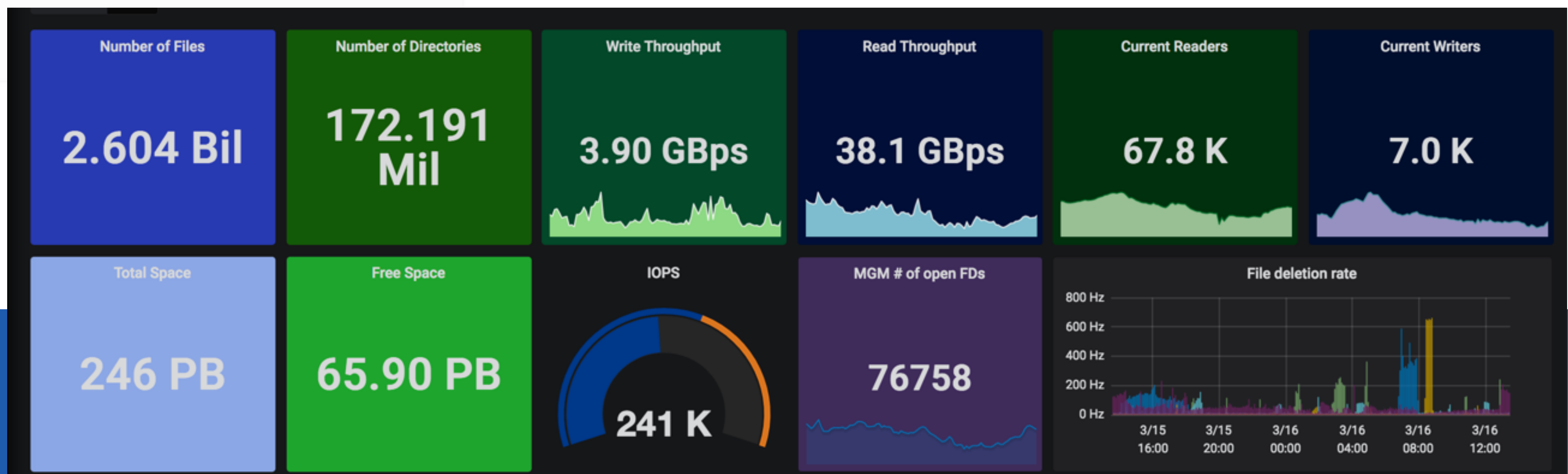
15 EOS instances

- 4 LHC
- 2 CERNBox (new home)
- EOSMEDIA (Foto, Video)
- EOSPUBLIC (non-LHC Experiments)
- EOSBACKUP (backup for CERNBox)
- 6 for various test infrastructures

Over 30 days



GRAFANA Dashboard 3/2018

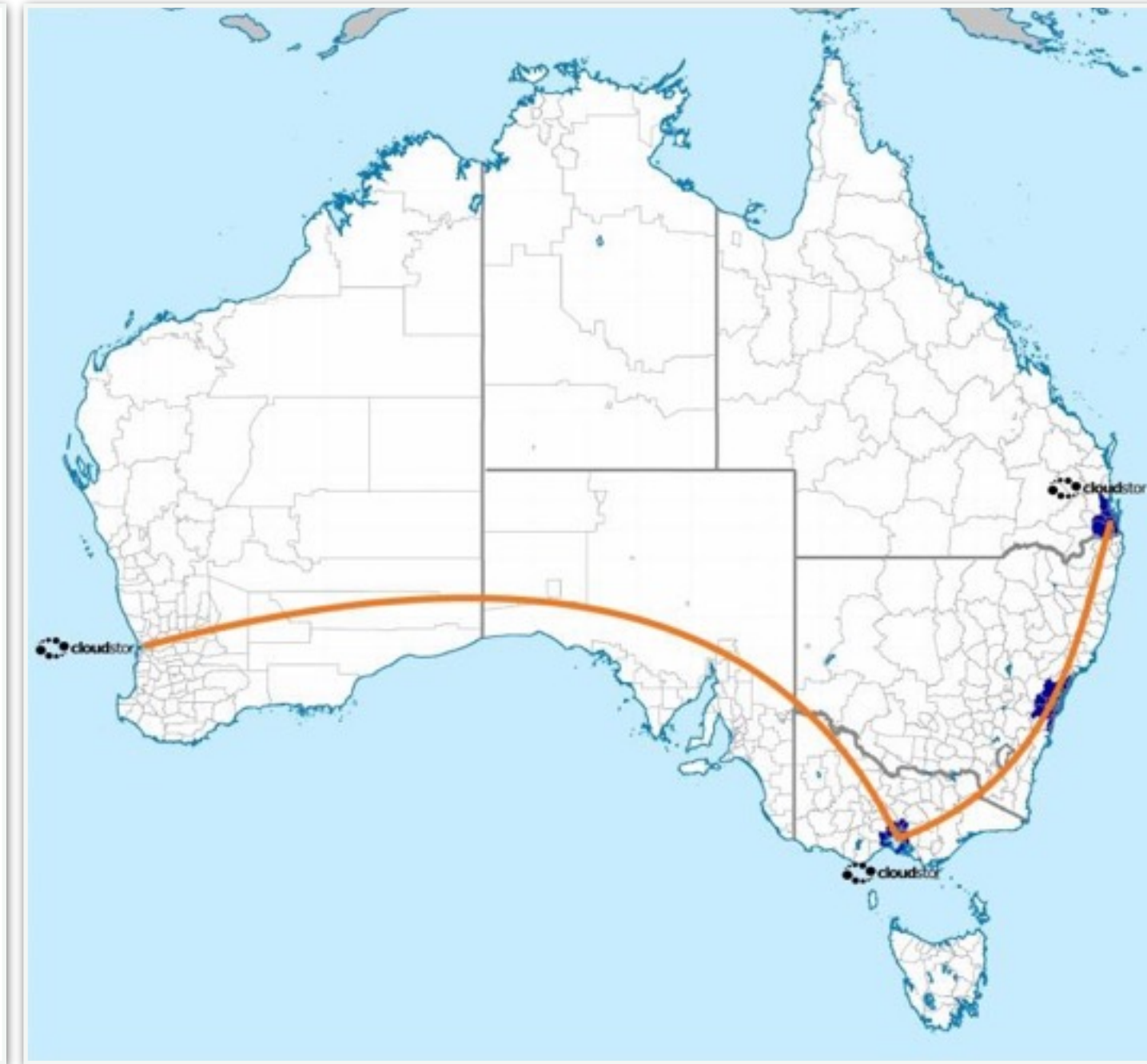
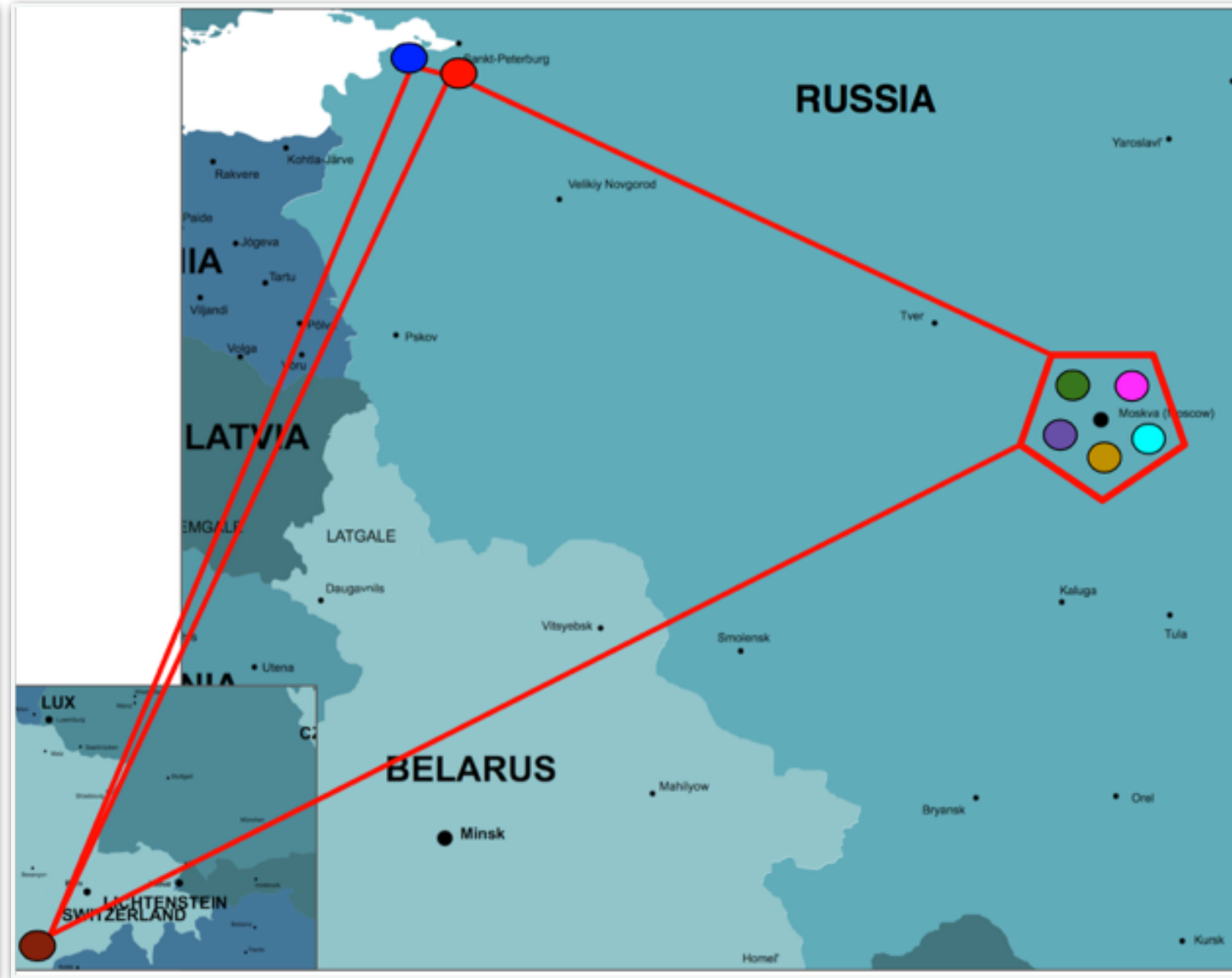
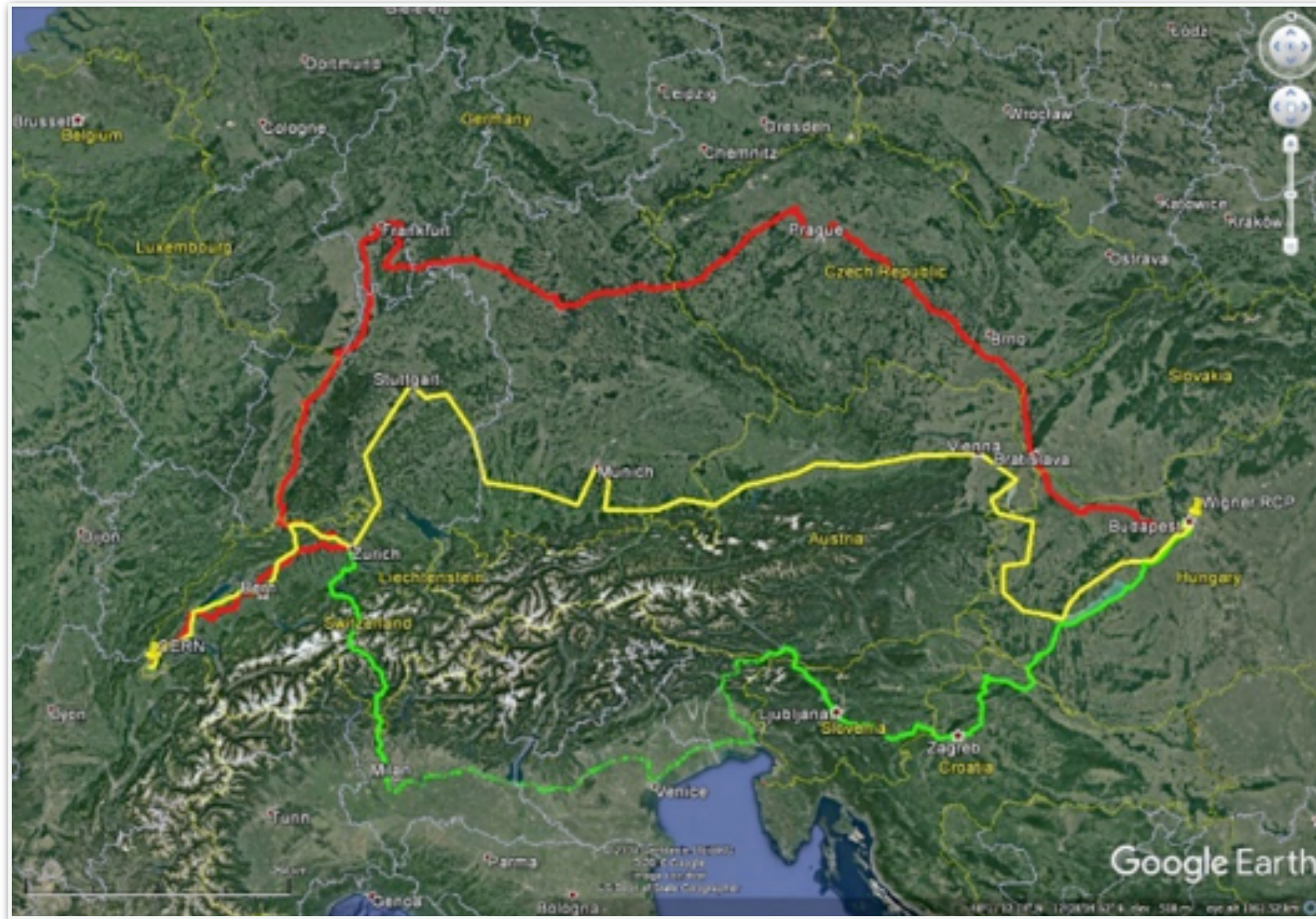




Distributed EOS

Latency 22 ms

60 ms



EOS@CERN
CERN & Wigner Data Center
3 x 100Gb

Russian Federation
Prototype

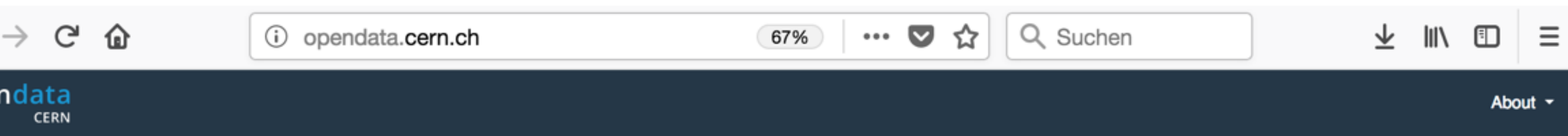
AARNet
CloudStor





EOS for OpenData

Invenio Digital Library Framework



Explore more than **1 petabyte** of open data from particle physics!

Start typing...

search examples: collision datasets, keywords:education, energy



Dataset x

Filter by type

- Dataset 997
 - Collision 100
 - Derived 173
 - Simulated 723
- Documentation 56
 - About 8
 - Activities 19
 - Authors 3
 - Guide 16
 - Help 2
 - Policy 4
 - Report 1
- Environment 19
 - Condition 5
 - VM 11

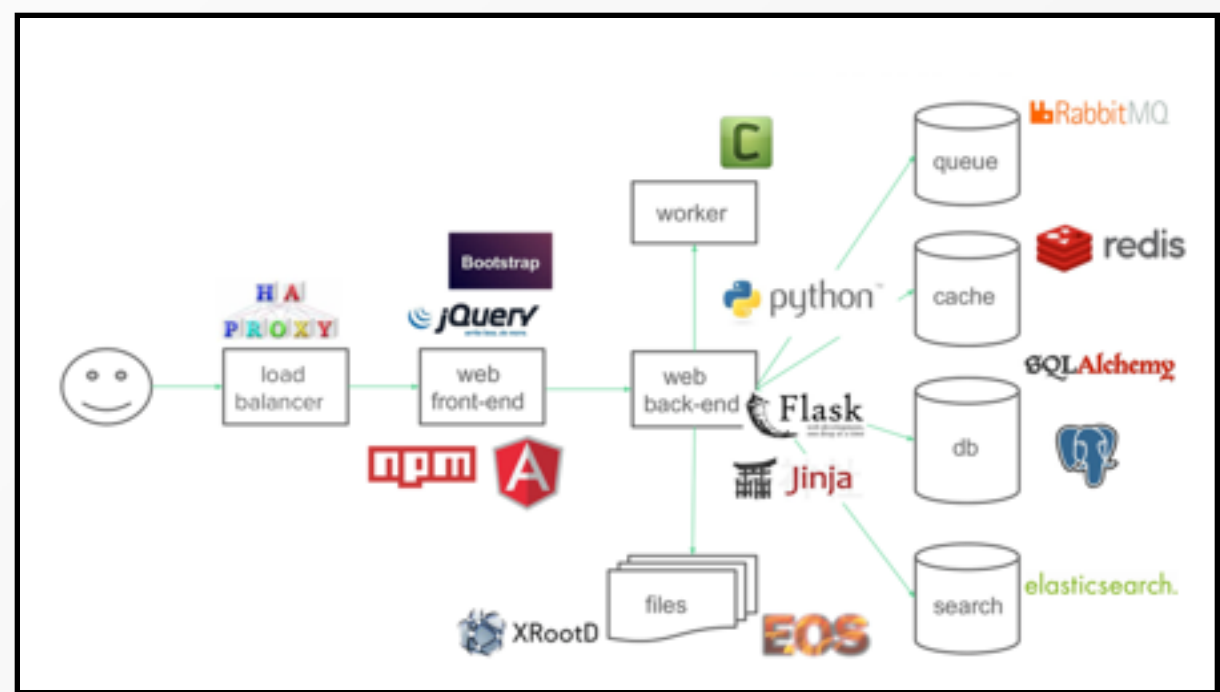
Sort by: Best match asc. Found 997 results.

/MinimumBias/Run2011A-v1/RAW
A sample from MinimumBias primary dataset in RAW format from P

Muons and electrons in PAT candidate format derived from 12Oct2013-v1/AOD primary dataset
Preprocessed data for the two-lepton/four-lepton analysis example

Muons and electrons in PAT candidate format derived from Apr21ReReco-v1/AOD primary dataset
Data preprocessed for the two-lepton/four-lepton analysis example

LHC10h_PbPb_ESD_139038
Pb-Pb ESD data sample at 3.5 TeV from RunH of 2010. Run period



opendata CERN

Higgs-to-four-lepton analysis example using 2011-2012 data

Cite as: Jomhari, Nur Zulaiha; Geiser, Achim; Bin Anuar, Afiq Aizzuddin; (2017). Higgs-to-four-lepton analysis example using 2011-2012 data. CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS_JKB8_RR42

Software Analysis CMS Accelerator CERN LHC

Description

This research level example is a strongly simplified reimplementation of parts of the original CMS Higgs to four lepton analysis published in Phys.Lett. B716 (2012) 30-61, arXiv:1207.7235.

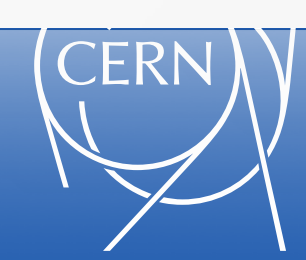
The published reference plot which is being approximated in this example is https://inspirehep.net/record/1124338/files/H4L_mass_3.png. Other Higgs final states (e.g. Higgs to two photons), which were also part of the same CMS paper and strongly contributed to the Higgs boson discovery, are not covered by this example.

The example consists of different levels of complexity. The highest level minimal understanding of the content of this paper and of the meaning educational exercises. The lower levels might also be interesting for ed with the linux operating system and the ROOT analysis tool.

Use with

The example uses legacy versions of the original CMS datasets in the A publication due to improved calibrations. It also uses legacy versions o but not identical to, the ones in the original publication. These legacy d in many later CMS publications.

/DoubleElectron/Run2011A-12Oct2013-v1/AOD
/DoubleMu/Run2011A-12Oct2013-v1/AOD



Get started





CERN Open Source for Open Data



Invenio

- <http://inveniosoftware.org>
- <http://github.com/inveniosoftware>
- @inveniosoftware



CERN Open Data

- <http://opendata.cern.ch>
- <http://github.com/cernopendata>



Tapes . . .





EOS + Tape = EOSCTA

CERN
TAPE
ARCHIVE

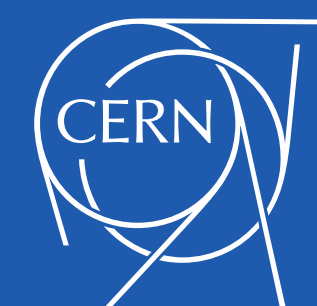
- in 2017 tape storage passed **200 PB** with CERN CASTOR storage system
- **CTA** modularises and splits tape functionality from the disk cache implementation and can be adapted to the disk technology
- Tape copies are treated in EOS as offline disk replicas
- **EOS** & **CTA** communicate via GOOGLE protocol buffer messages, which can be configured synchronous or asynchronous using the EOS workflow engine
- first production CTA code available in **2018** - continuous testing & improvements currently on the way



participating
in



Extreme (Data) Clouds



<http://www.extreme-datacloud.eu/>



EOS as a **datalake** technology

Bird, Campana, Espinal, Girona (CERN-IT/LCG)

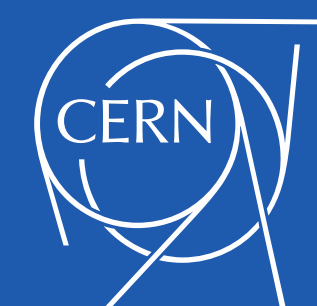


pleiadian-starseed.com

in



WLCG
Worldwide LHC Computing Grid



WLCG
Worldwide LHC Computing Grid

<http://wlcg.web.cern.ch/>



Datalakes

Datalakes: evolution of distributed storage

- Datalakes are an extension of storage consolidation where geographically distributed storage centers are operated and accessed as a single entity

Goals

- Optimise storage usage to lower the cost of stored data

technology requirements: geo-awareness, storage tiering and automated file workflows fostered by fa(s)t - **QOS**

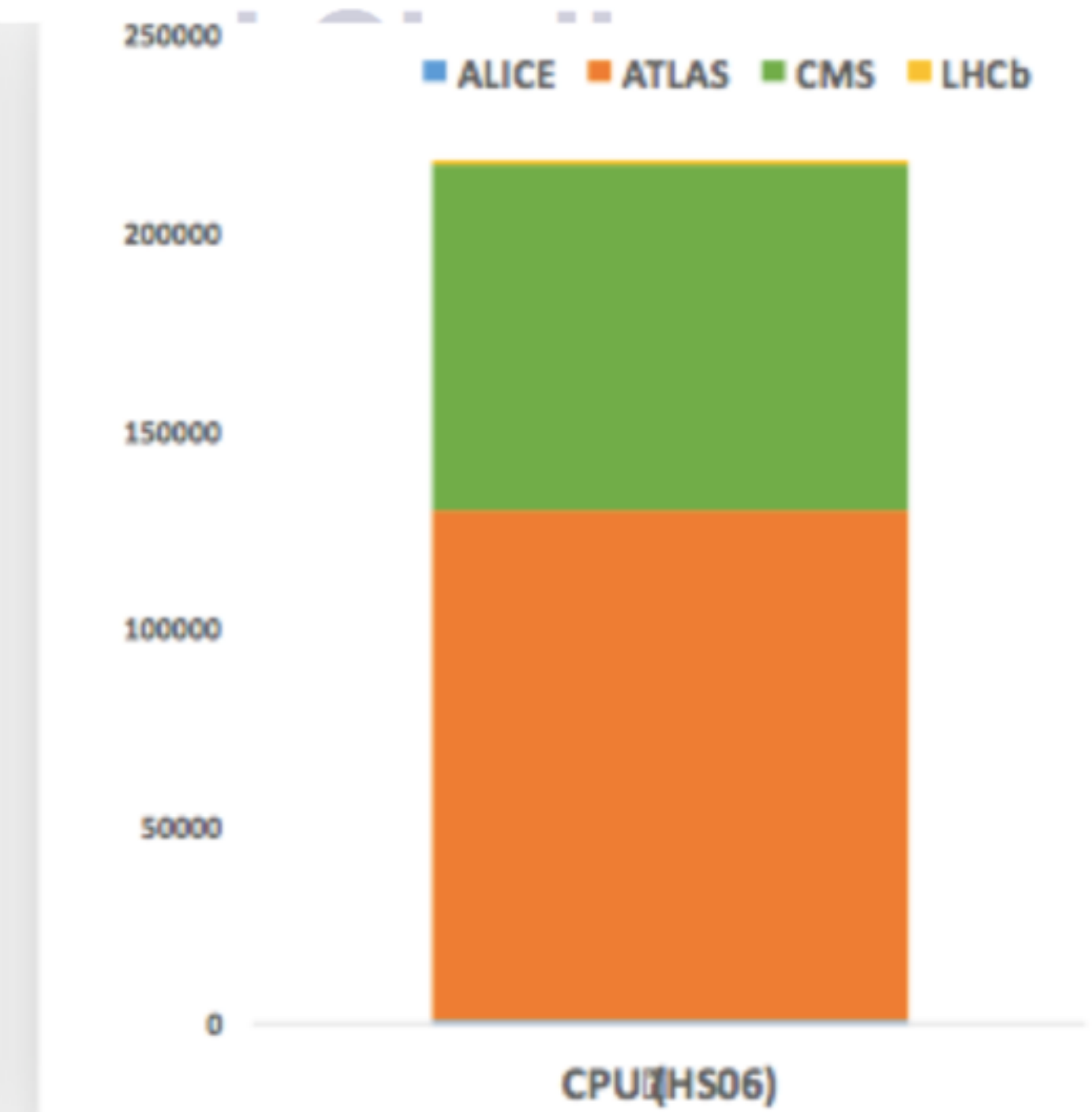


Datalakes

HL-LHC
deal with 12 more data

Datalakes motivation: High Luminosity LHC

Openlab Collaboration Board 2017



Data:

- Raw 2016: 50 PB → 2027: 600 PB
- Derived (1 copy): 2016: 80 PB → 2027: 900 PB

CPU:

- x60 from 2016



Techn



Enable Other Storage



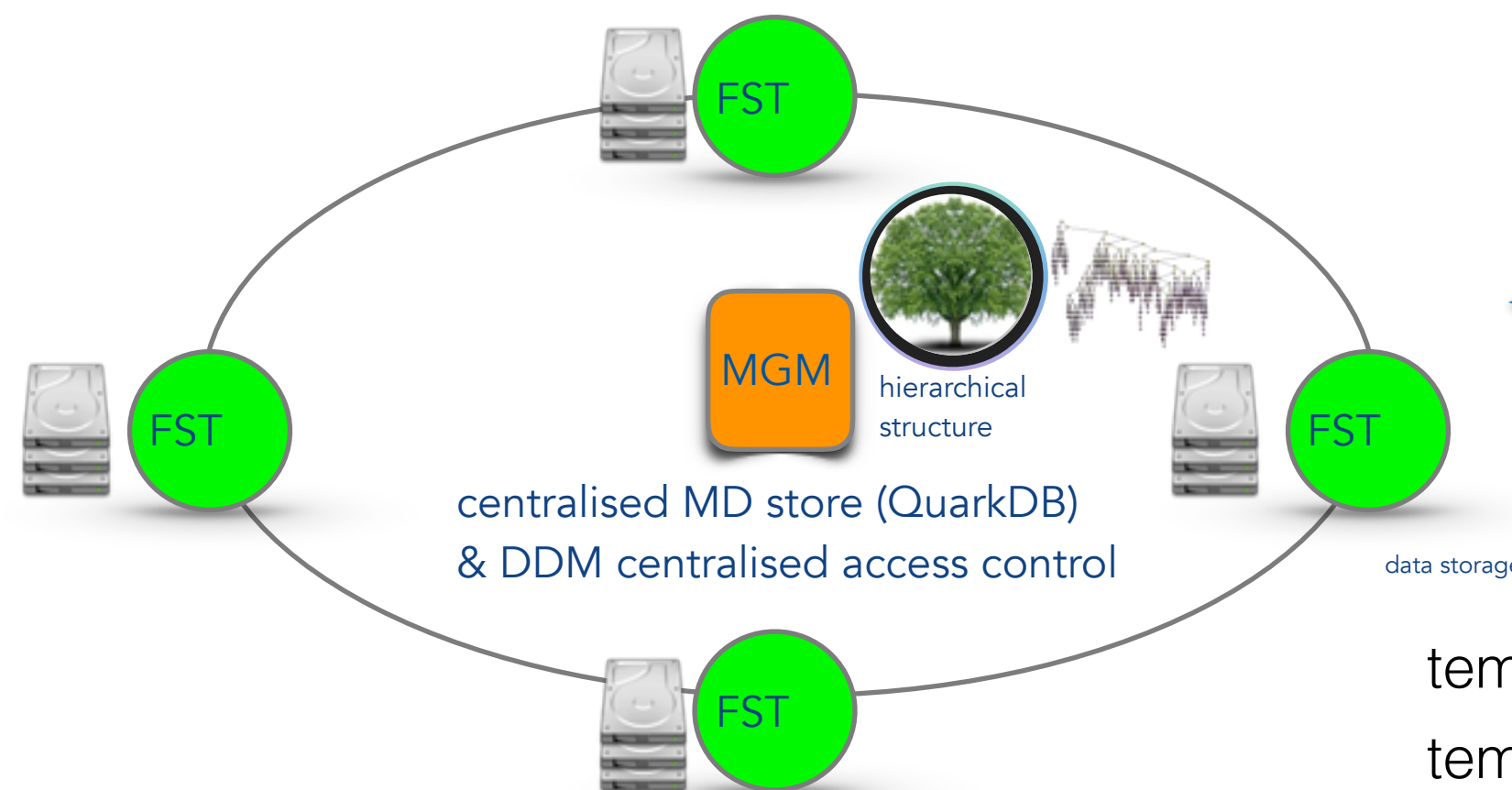
- Scope of EOS in XDC & WLCG Datalake project
 - enable storage **cache**s
 - enable **hybrid** storage
 - **distributed deployments** and storage **QOS** for cost savings
- What does this really mean?



Dynamic Caches

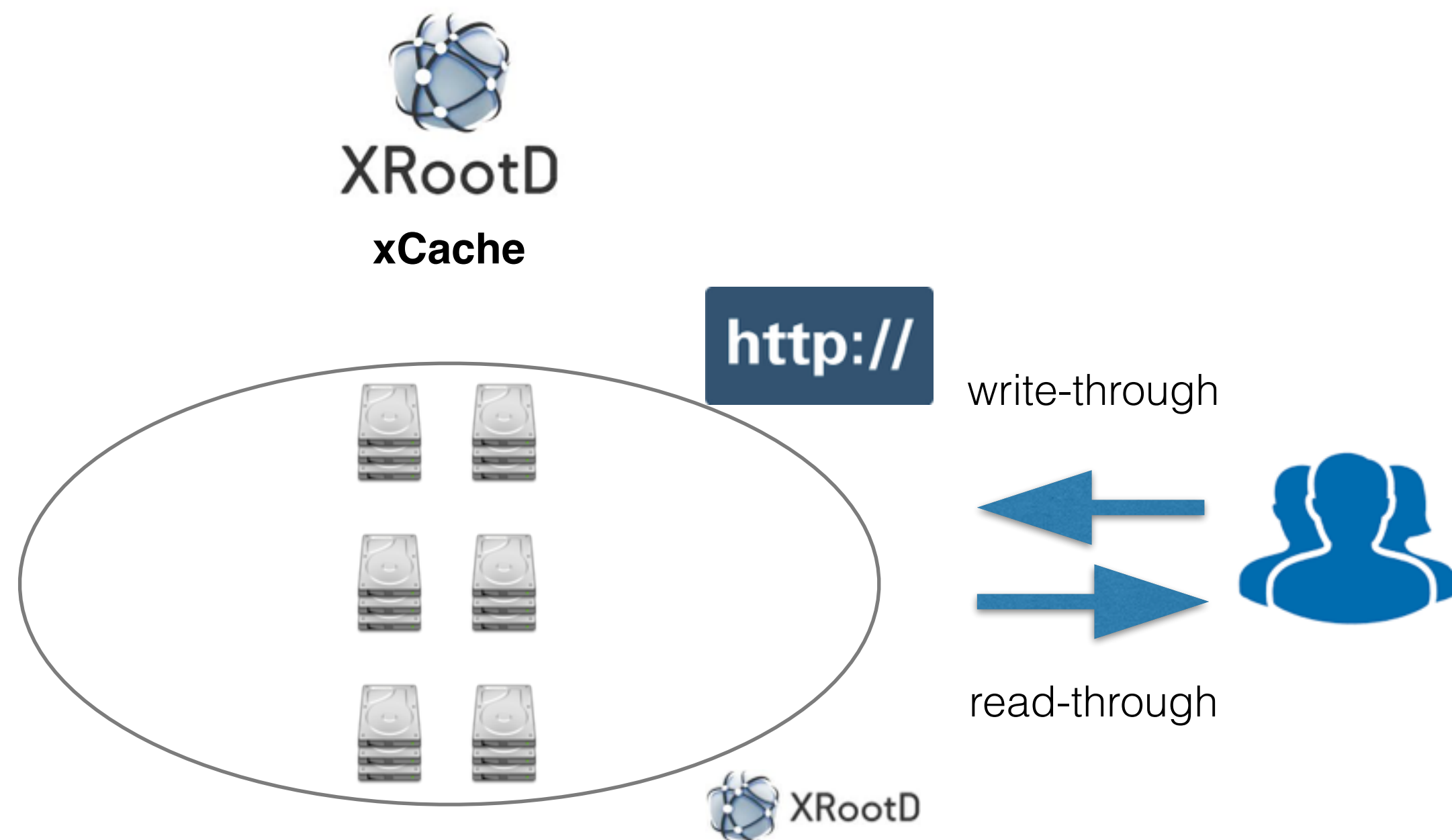
Adding clustered storage caches as dynamic resource.

Files can have replicas in static (EULAKE) and dynamic resources (CACHE-FOO).



distributed EOS setup
EULAKE

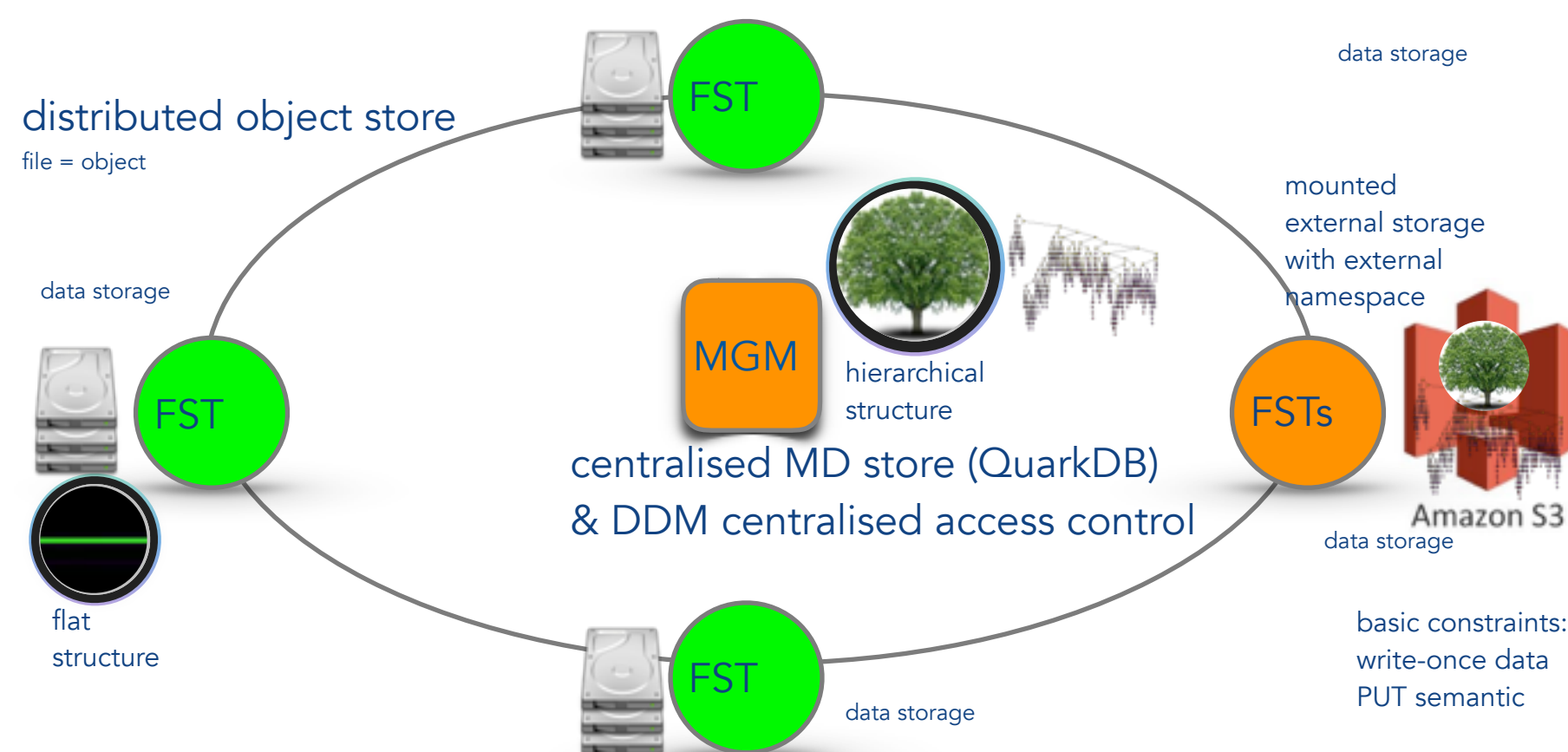
IO with credential tunnelling
temporary replica geotag creation
temporary replica geotag deletion



dynamic site cache resource
CACHE-FOO

EOS Hybrid Distributed Storage

- Attach external storage into datalake
 - external storage has not to be accessed via data lake - can be operated as is: better scalability
 - external storage connector uses a notification listener to publish creations and deletions and applies QOS (replication) policies to distribute data in the lake



Planned connectors

Amazon **S3**

CEPH **S3**

Shared **Filesystem** (with limitations)

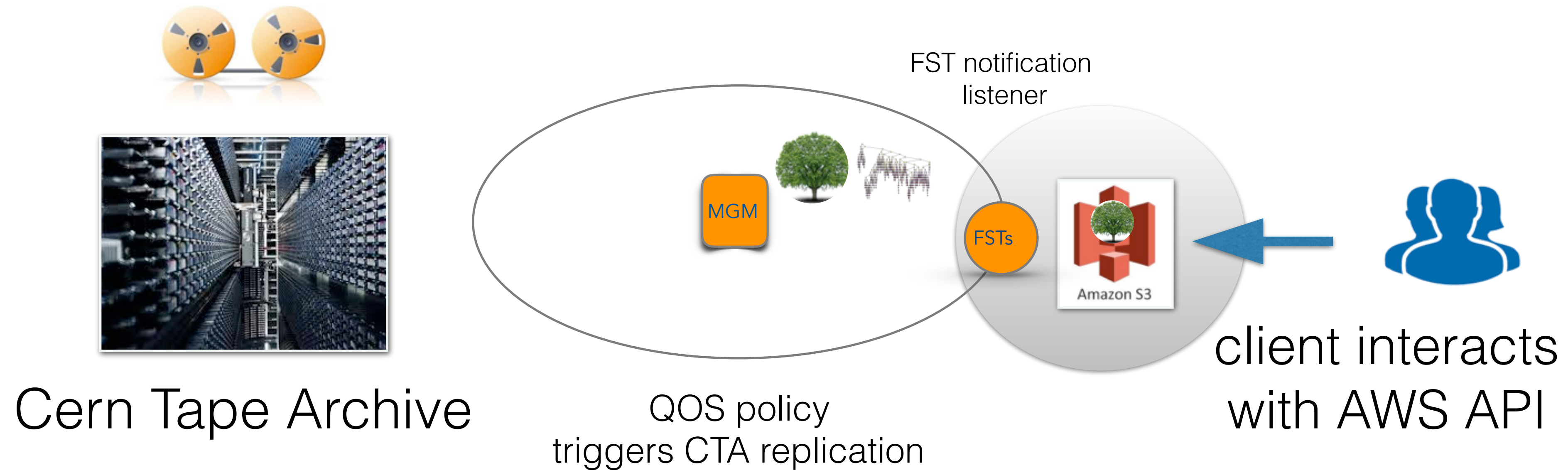
ExOS **Object Storage** (RADOS)

XRootD/WebDAV+REST

Hybrid Distributed Storage

Example: AWS Integration

- transparent S3 backup on tapes



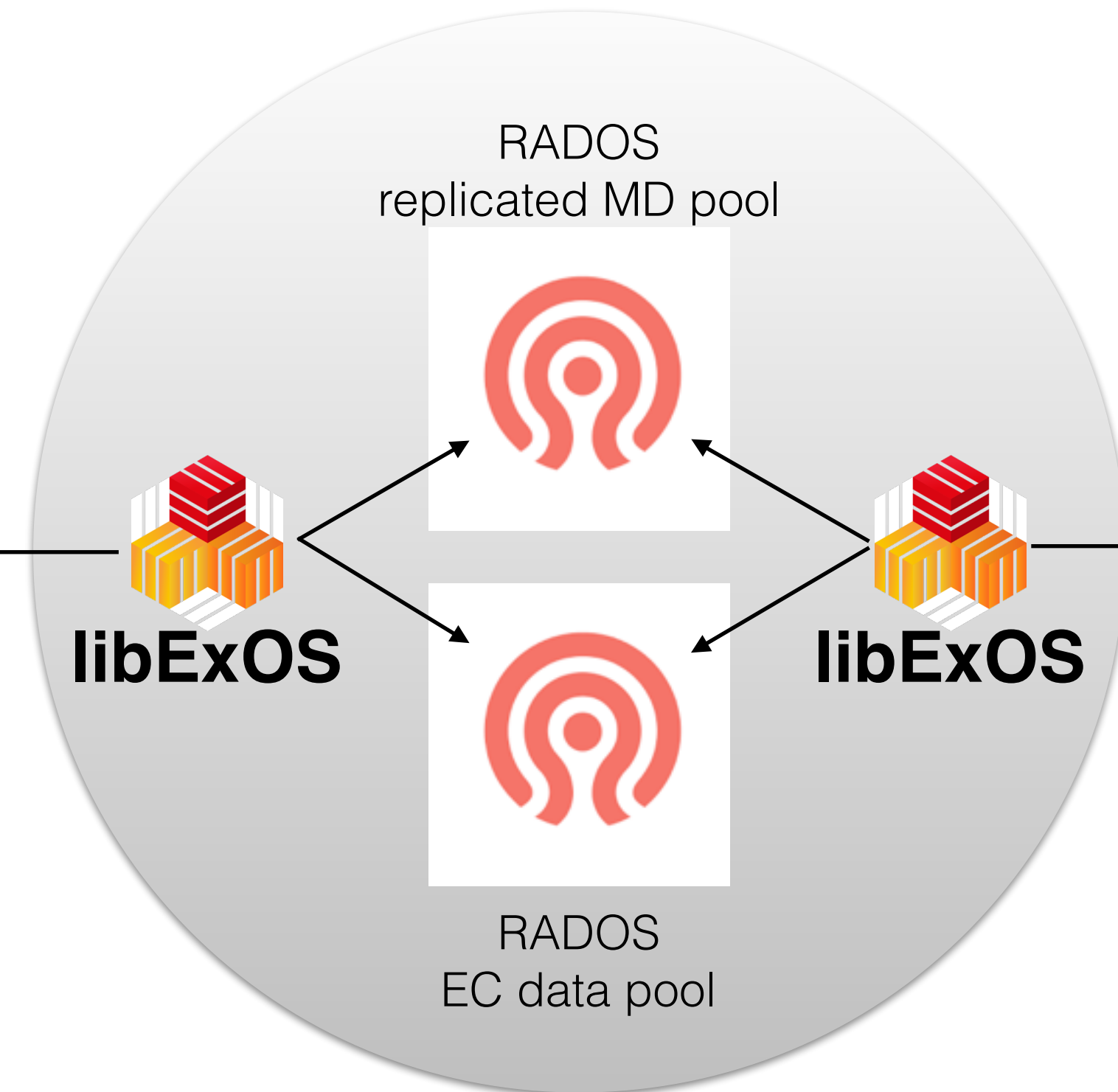
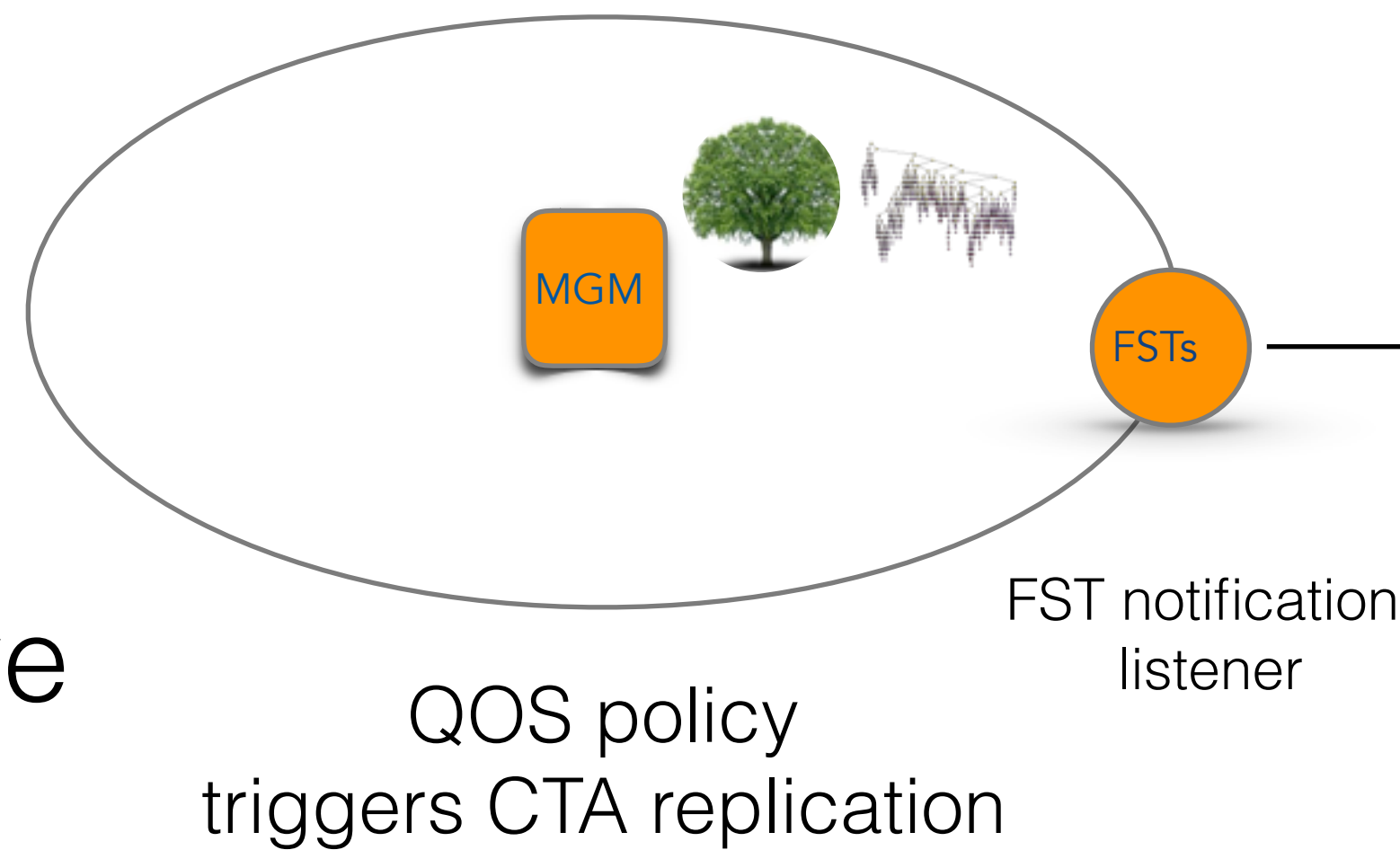


Hybrid Distributed Storage

Example High Performance DAQ with Object Storage



Cern Tape Archive



DAQ Farm



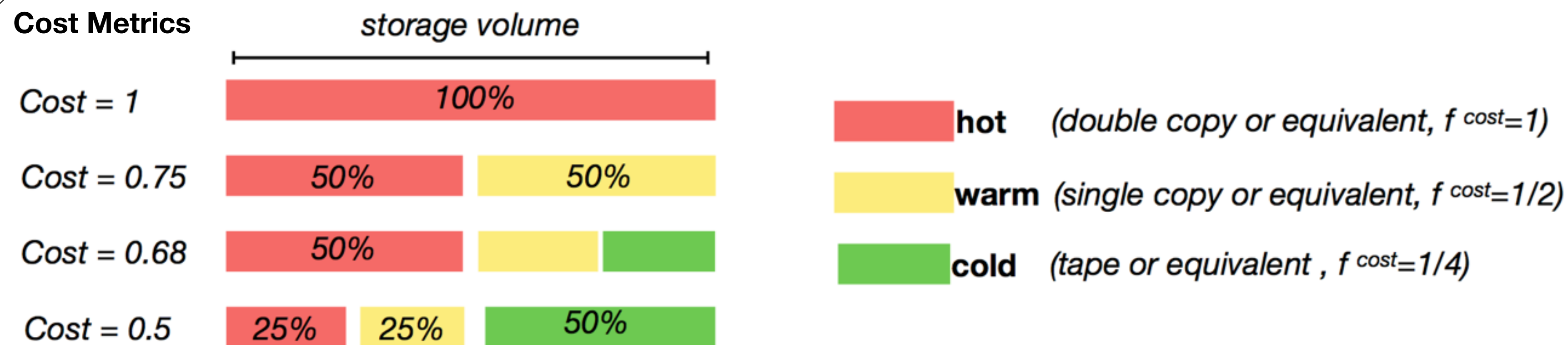
libExOS is a lock-free minimal implementation to store data in RADOS object stores optimised for erase encoding
 leverages CERN IT-ST experience as author of RADOS striping library & intel EC





QOS in EOS

How do we save?



EOS provides a **workflow engine** and **QOS transformations**

- event (put, delete) and time trigger (file age, last access) workflows
- file layout transformations [replica \leftrightarrow EC encode*] [e.g. save 70%]
- policies are expressed as external attributes and express structure and geographical placement [skipping a lot of details]

* can do erasure encoding over WAN resources/centers

used for CTA



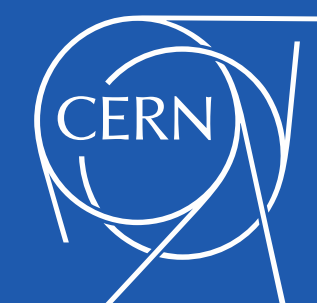


CERN Scientific Services Bundle

We have bundled a demonstration setup of four CERN developed cloud and analysis platform services called [UBoxed](#).

encapsulated four components

- [EOS](#) - scalable **storage** platform with data, metadata and messaging server components
- [CERNBox](#) - dropbox-like add-on for **sync-and-share** services on top of EOS
- [SWAN](#) - service for web based **interactive analysis** with jupyter notebook interface
- [CVMFS](#) - CernVM file system - a scalable **software distribution** service




Try dockerized Demo Setup on CentOS7 or Ubuntu:

eos-docs.web.cern.ch/eos-docs/quickstart/uboxed.html




CERN Scientific Services Bundle

CERN Accelerating science




EOS
Disk-based storage service.

[Docs](#) [More Info](#)



CERNBox
Cloud Storage with Sync&Share.

[Try it!](#) [More Info](#)



SWAN
Interactive Data Analysis in the Cloud.

[Try it!](#) [More Info](#)

Documentation

<http://cernbox.cern.ch/cernbox/doc/boxed>

User information

The following accounts have been pre-created (username:password)

- user0:test0
- user1:test1
- ...
- user9:test9


FEEDBACK

Get in touch with us at:
cernbox-talk (at) cern (dot) ch


Web Service Interface after UBoxed installation

EOS CITRINE documentation » Install » previous | next | index

Scientific Services Installation: EOS, CERNBox, SWAN and CVMFS

 We have bundled a demonstration setup of four CERN developed cloud and analysis platform services called **UBoxed**. It encapsulates four components:

- **EOS** - scalable storage platform with data, metadata and messaging server components
- **CERNBox** - dropbox-like add-on for sync-and-share services on top of EOS
- **SWAN** - service for web based interactive analysis with jupyter notebook interface
- **CVMFS** - CernVM file system - a scalable software distribution service



Documentation
<http://cernbox.cern.ch/cernbox/doc/boxed>

User Information

FEEDBACK

Get in touch with us at:
cernbox-talk (at) cern (dot) ch

Table Of Contents

Scientific Services Installation: EOS, CERNBox, SWAN and CVMFS

- Preparation
- Quick Setup
 - Install Services
 - Setup and Initialize Services
 - Run a Self Test
 - Connect to your services
- Stop Services
- Cleanup docker images and volumes

Previous topic

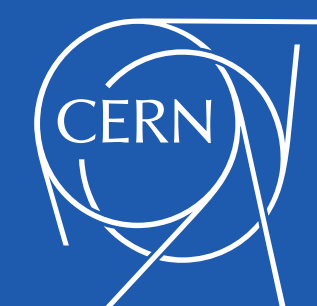
EOS Docker Installation

Next topic

RPM installation



Try dockerized Demo Setup on CentOS7 or Ubuntu:
eos-docs.web.cern.ch/eos-docs/quickstart/uboxed.html





CERN Scientific Services Bundle

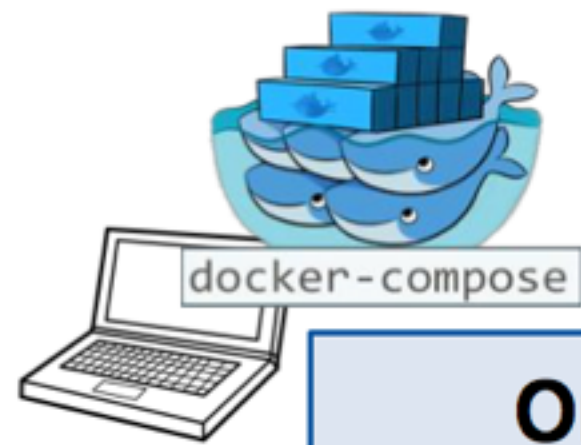


+



CERNBox

+



One-Click Demo Deployment

- Single-box installation via **docker-compose**
- No configuration required
- Download and run services in 15 minutes

<https://github.com/cernbox/uboxed>

Production-oriented Deployment

- Container orchestration with **Kubernetes**
- Scale-out storage and computing
- Tolerant to node failure for high-availability

<https://github.com/cernbox/kuboxed>



EOS as a filesystem /eos

background to /eos

- a filesystem mount is **standard API supported by every application**
 - not always the most efficient for physics analysis
- a filesystem mount is very **delicate interface**
 - any failure translates into applications failures, job inefficiencies etc.
- FUSE is a simple (not always) but not the most efficient way to implement a filesystem
- implementing a filesystem in general is challenging, currently deployed implementation has many POSIX problems
- we implemented **3rd generation of a FUSE based client** for EOS



EOS as a filesystem /eos

features

- more **POSIX** - better **performance** - cross client md/data consistency
- strong security: **krb5** & **certificate** authentication - **oauth2** under consideration
- distributed byte range **locking** - small file caching
- **hard links** (starting with version 4.2.19)
- **rich ACLs** support on the way

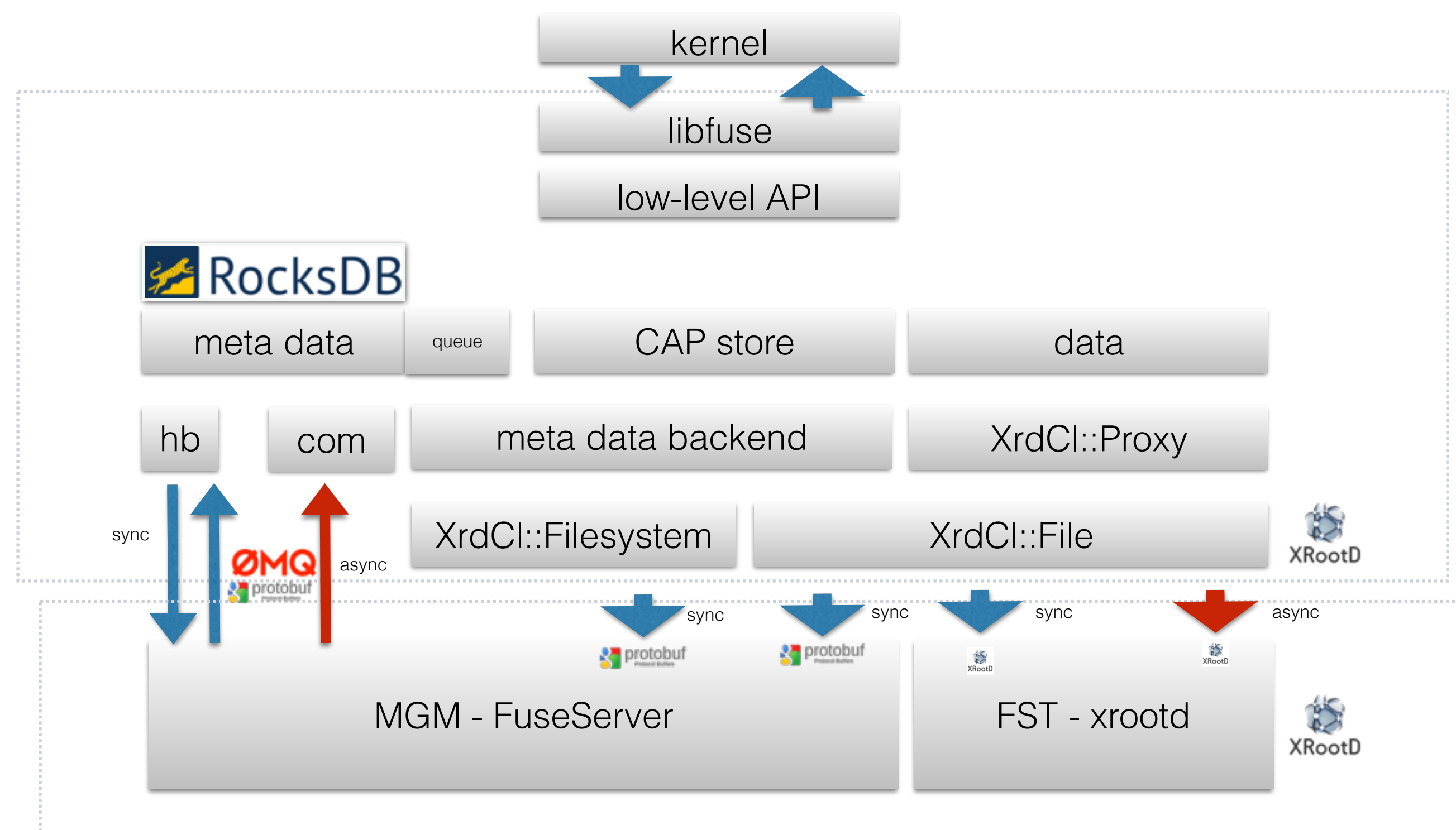


eosxd

FUSE filesystem daemon

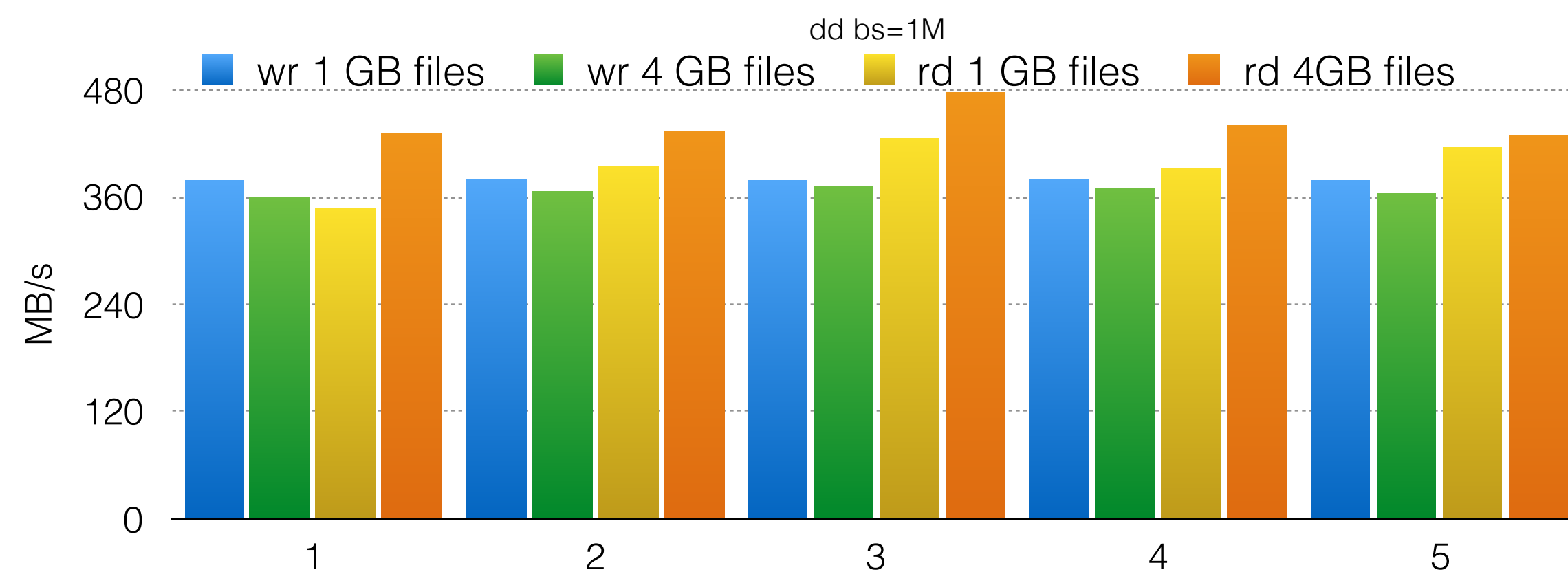


Architecture



Example Performance Metrics

```
1000x mkdir = 870/s  
1000x rmdir = 2800/s  
1000x touch = 310/s  
untar (1000 dirs) = 1.8s  
untar (1000 files) = 2.8s
```





eosxd

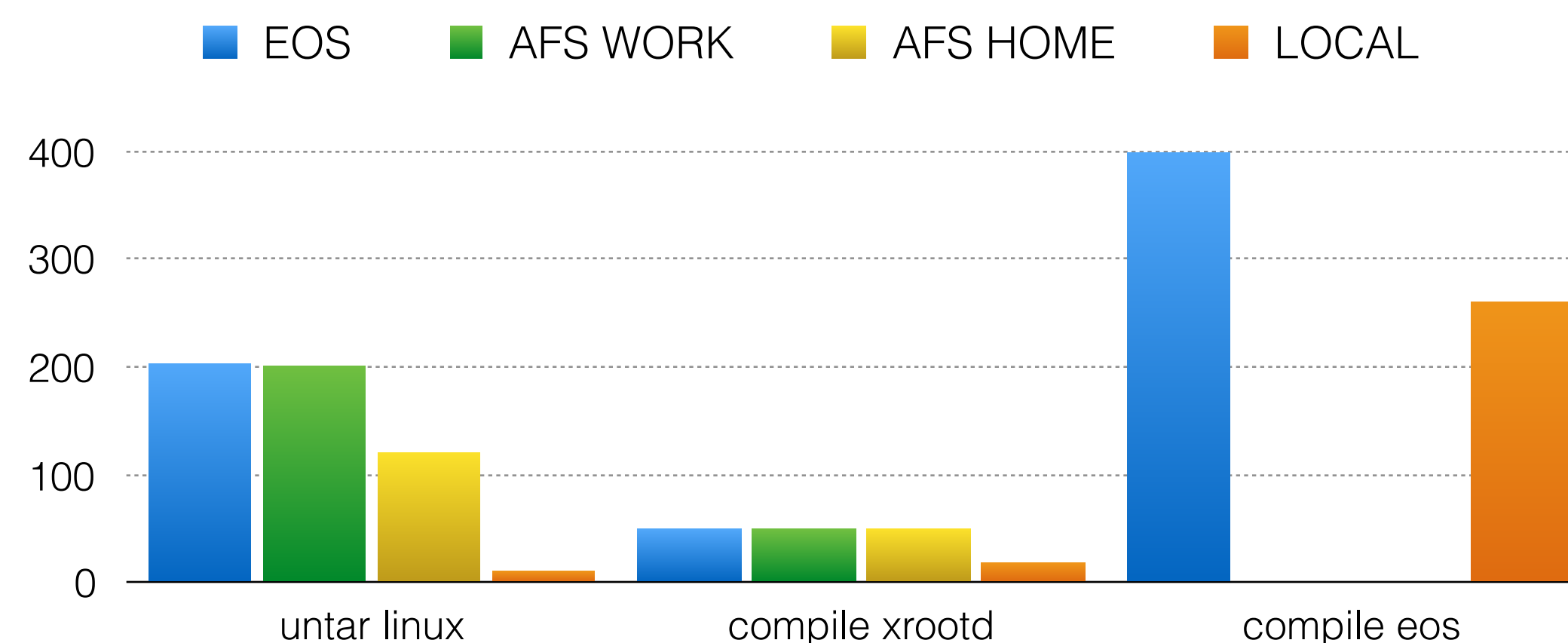
FUSE filesystem daemon



- aim to take over some AFS use cases
- related to AFS phaseout project at CERN (longterm)
- provide at least POSIX features of AFS

Example Performance metrics

- untar linux source (65k files/directories)
- compile xrootd
- compile eos





EOS Vision

- evolve from CERN Open Source to **Community Open Source** project - outcome of 2nd EOS workshop
- leverage power of community storage Open Source
- **embedded** technologies
(object storage & filesystem hybrids)
- **slim-down** storage customisation layers



Summary & Outlook

- **EOS** design undergoes a significant architectural evolution to prepare for current and future storage scale - 2018 is a year of big changes
single CITRINE instance with 3 billion files and 1kHz 24h average creation rate in pre-production
- **EOS** & CERN scientific services offer a rich portfolio for scientific data repositories
- **EOS** project is actively working on an evolution of distributed storage
- **EOS** is very actively developed open source storage software (including up and downs) shifting focus to higher-level storage abstractions

THANK YOU

QUESTIONS ?



