EOS Open Storage

the CERN storage ecosystem for scientific data repositories

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

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

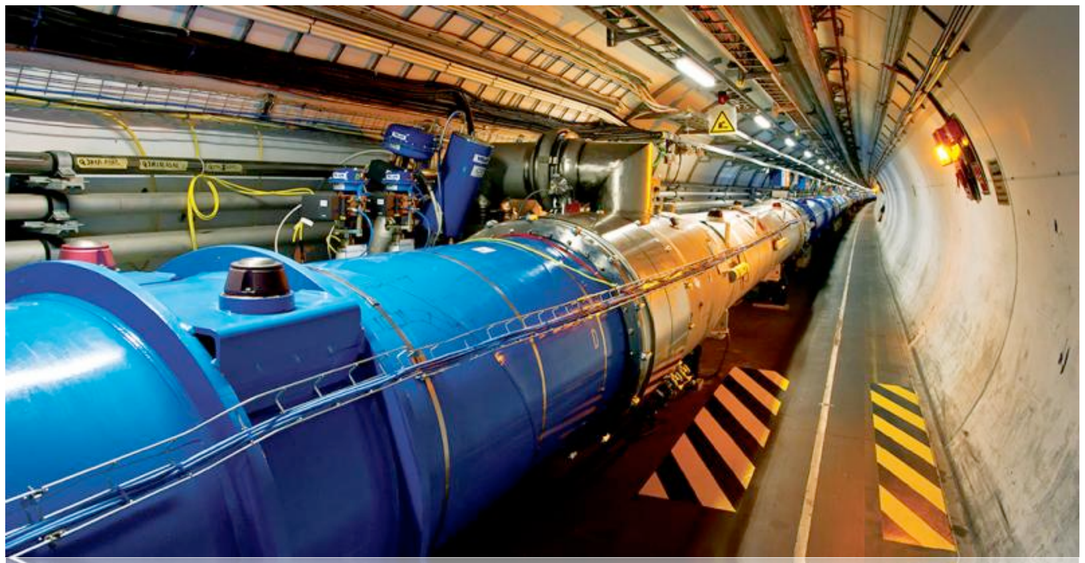
What is EOS?

EOS is a storage software solution for

- central data recording
- user analysis
- data processing

Introduction

EOS and the Large Hadron Collider LHC

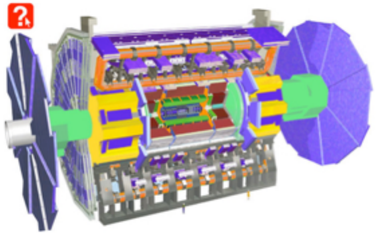


CERN mainstream use case

tape archive CASTOR

CERN Advanced STORage manager

LHC Detector



O(GB/s)

EOS Open Source Storage

5-10 GB/s

5-10 GB/s

peak 100 GB/s

local batch cluster

O(10^5 cores)

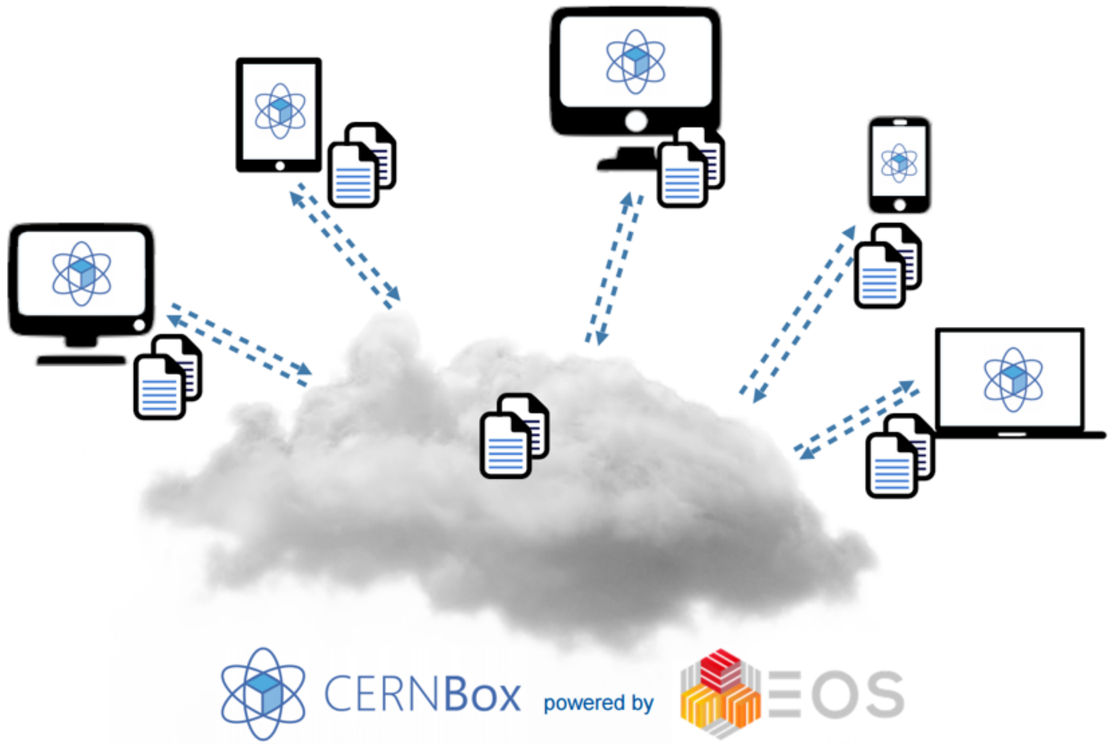
openstack™

Data Export to Worldwide Computing Grid

Introduction

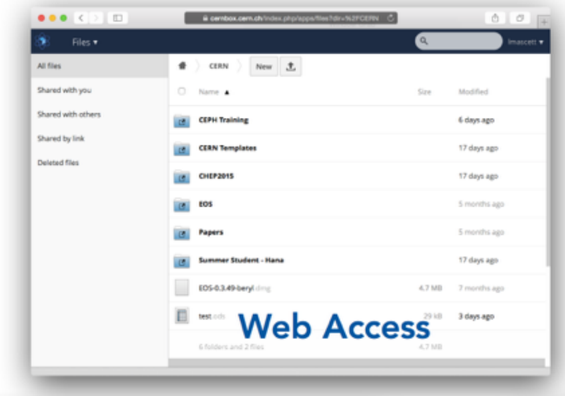
EOS and CERNBox

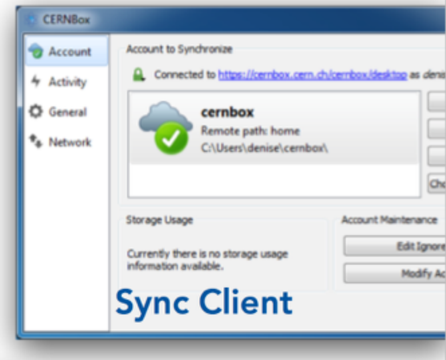
Sync & Share Platform
with collaborative editing

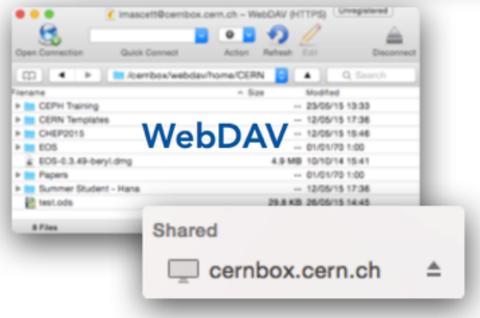


Available Access Methods

Office 365







EOS

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

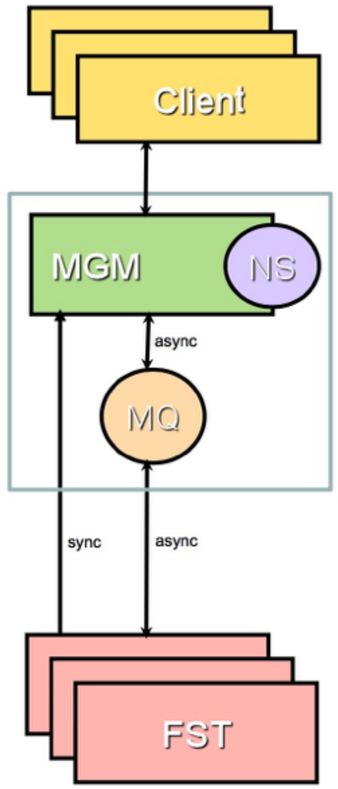
Architecture

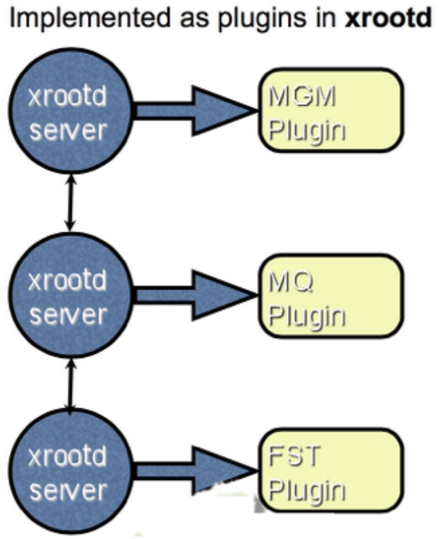
Storage Clients:
Browser, Appliations, Mounts

Meta Data Service / Namespace

Asynchronous Messaging Service

Data Storage





EOS is implemented in C++ using the XRootD framework

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

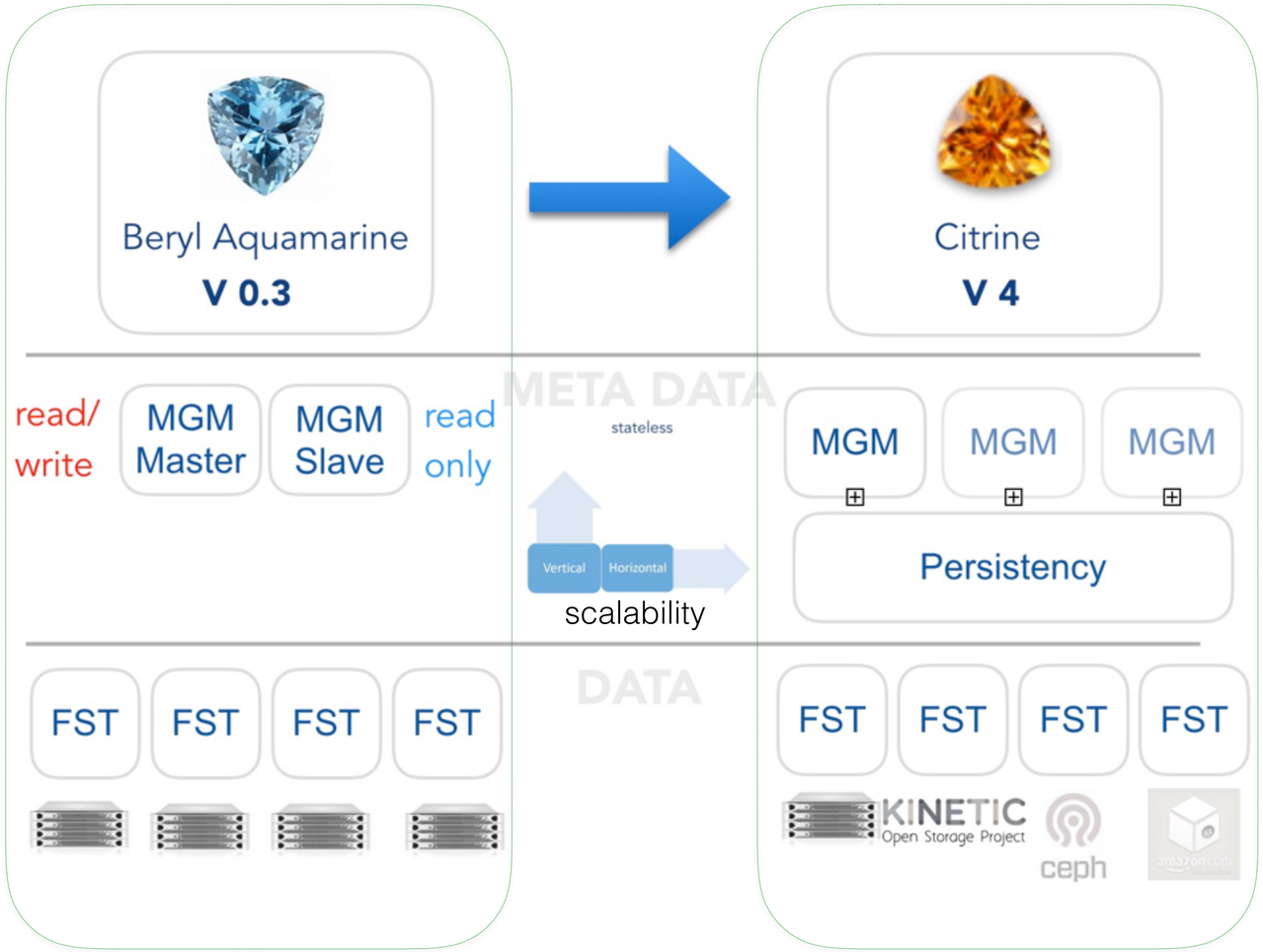
- third party transfer
- WAN latency compensation using vectored read requests
- pluggable authentication framework
- …

Architecture Transition 2017/18

EOS releases are named after **gemstones**

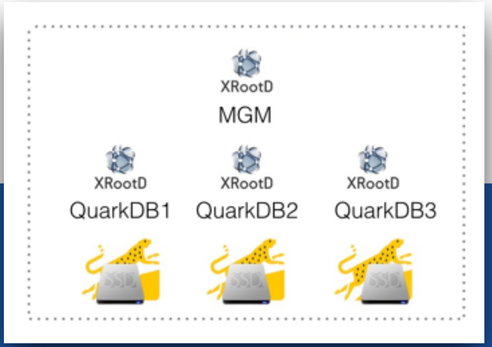
AQUAMARINE Version
- production <= 2017
- in-memory namespace

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

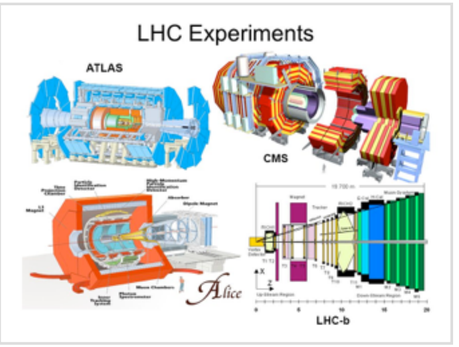


- during 2017 CERN services exceeded design limits - lower service availability

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



EOS at CERN

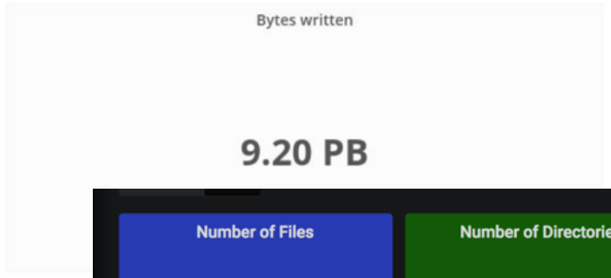


	2017	2018
Nodes	~1200	~1400
Disks	~40000	~50000
Raw capacity	~150PB	~250PB

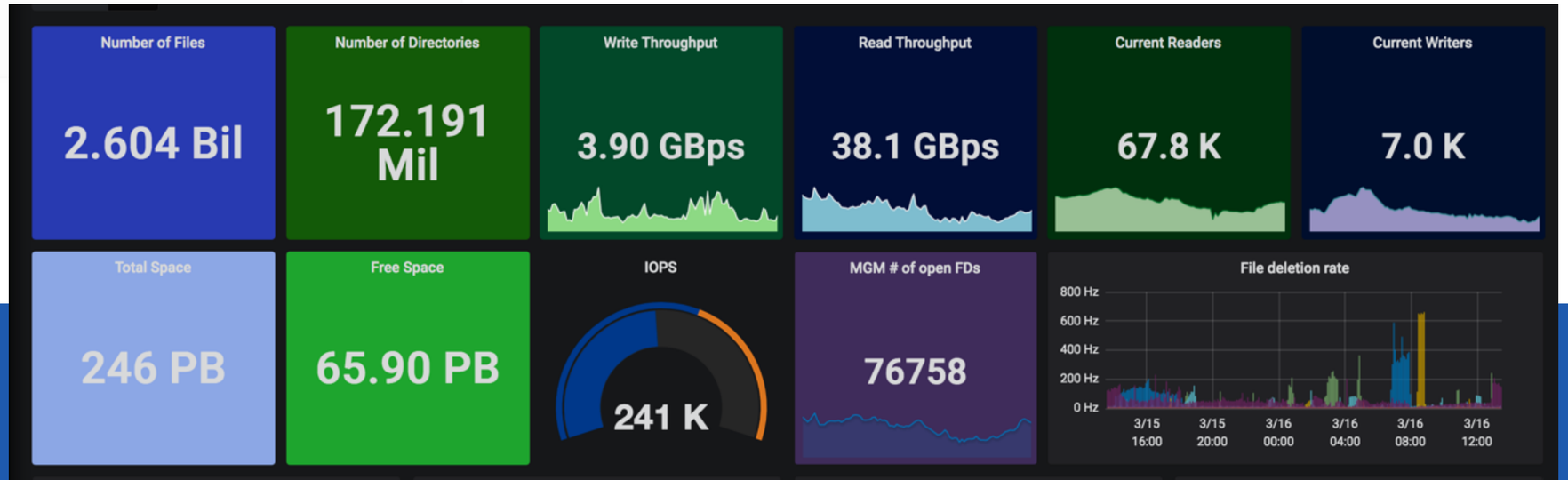
15 EOS instances

- 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

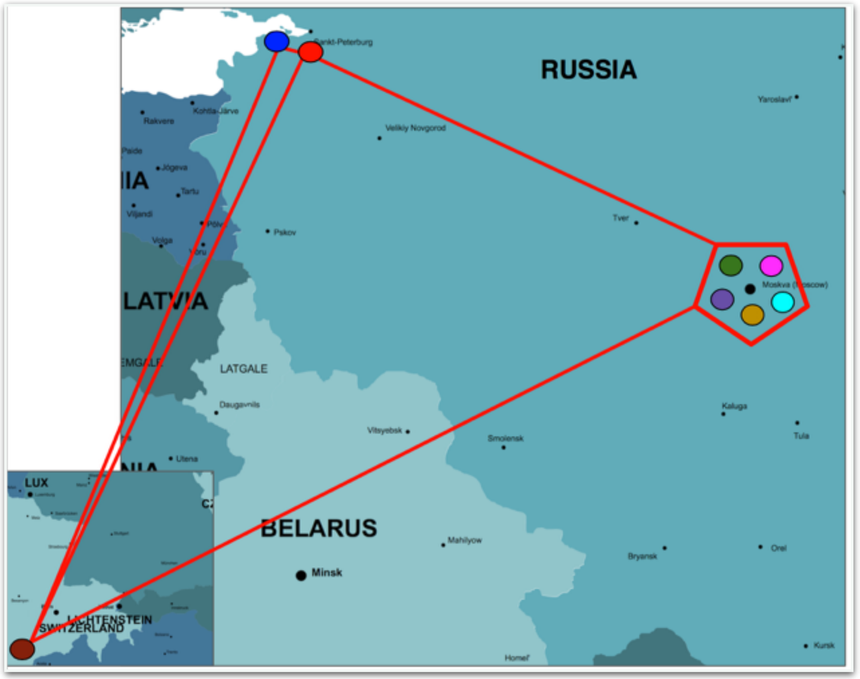


GRAFANA Dashboard 3/2018



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

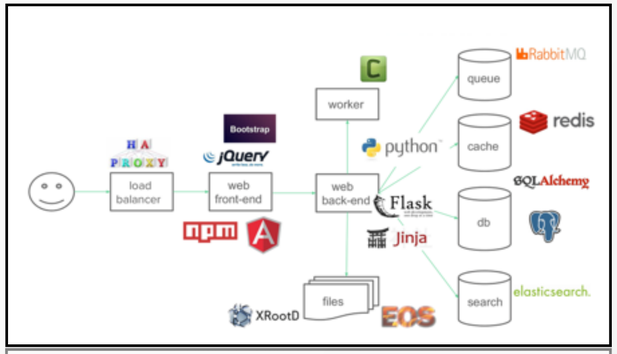
opendata.cern.ch

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

Start typing...

Search

search examples: collision datasets, keywords:education, energy:7



Get started

Dataset ×

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.

< 1 2 3 4 5 6 7 8

/MinimumBias/Run2011A-v1/RAW

A sample from MinimumBias primary dataset in RAW format from R

Dataset Collision CMS

Muons and electrons in PAT candidate format derived fr
12Oct2013-v1/AOD primary dataset

Preprocessed data for the two-lepton/four-lepton analysis example

Dataset Derived CMS

Muons and electrons in PAT candidate format derived fr
Apr21ReReco-v1/AOD primary dataset

Data preprocessed for the two-lepton/four-lepton analysis example

Dataset Derived CMS

LHC10h_PbPb_ESD_139038

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

opendata CERN

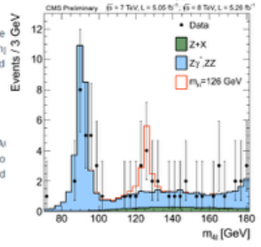
Higgs-to-four-lepton analysis example using 2011-2012 data

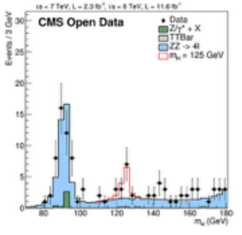
Description

Use with

/DoubleElectron/Run2011A-12Oct2013-v1/AOD

/DoubleMu/Run2011A-12Oct2013-v1/AOD





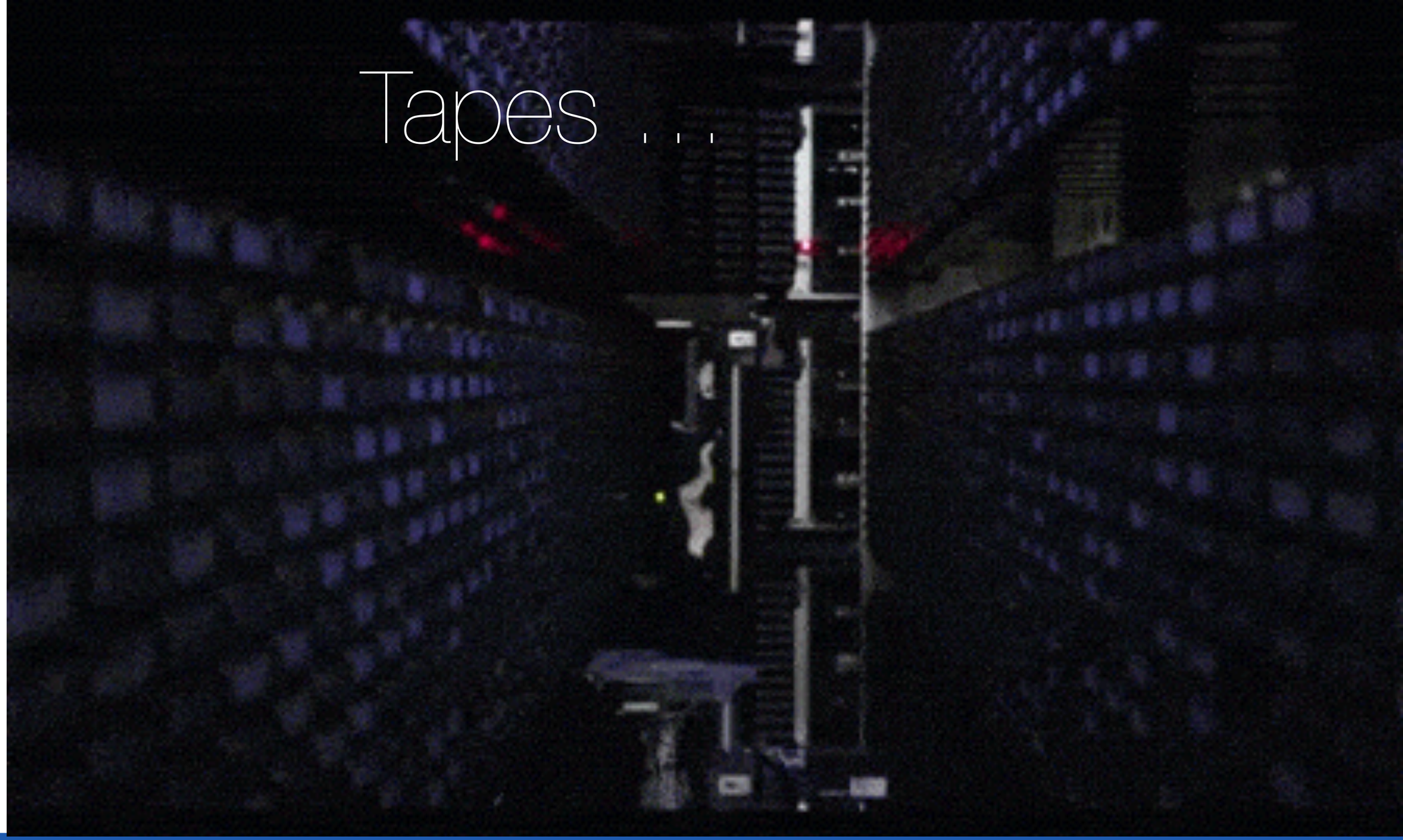
CERN Open Source for Open Data

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 DataCloud

Extreme (Data) Clouds

CERN

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

EOS

in

WLCG
Worldwide LHC Computing Grid

CERN

EOS as a datalake technology

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

CERN

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

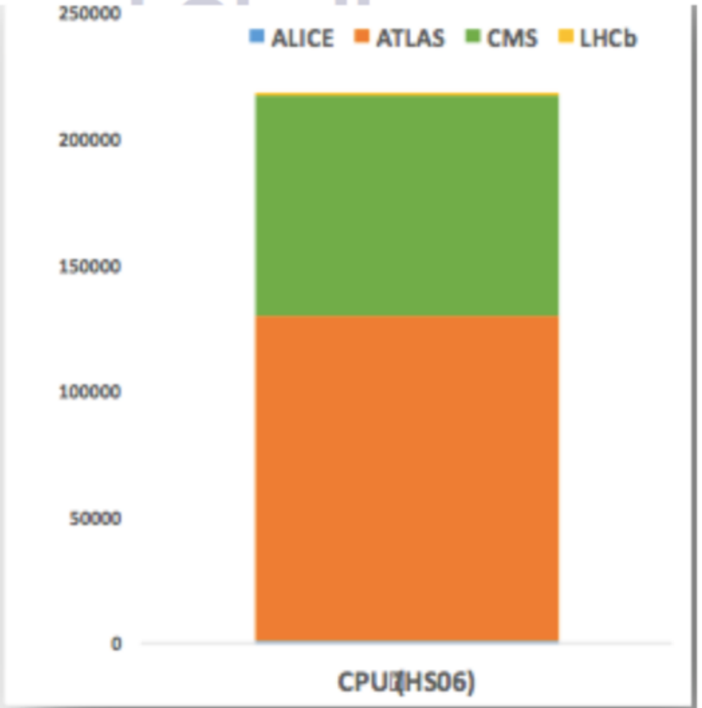
technology requirements: geo-awareness, storage tiering and automated file workflows fostered by fa(s)t - **QOS**

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

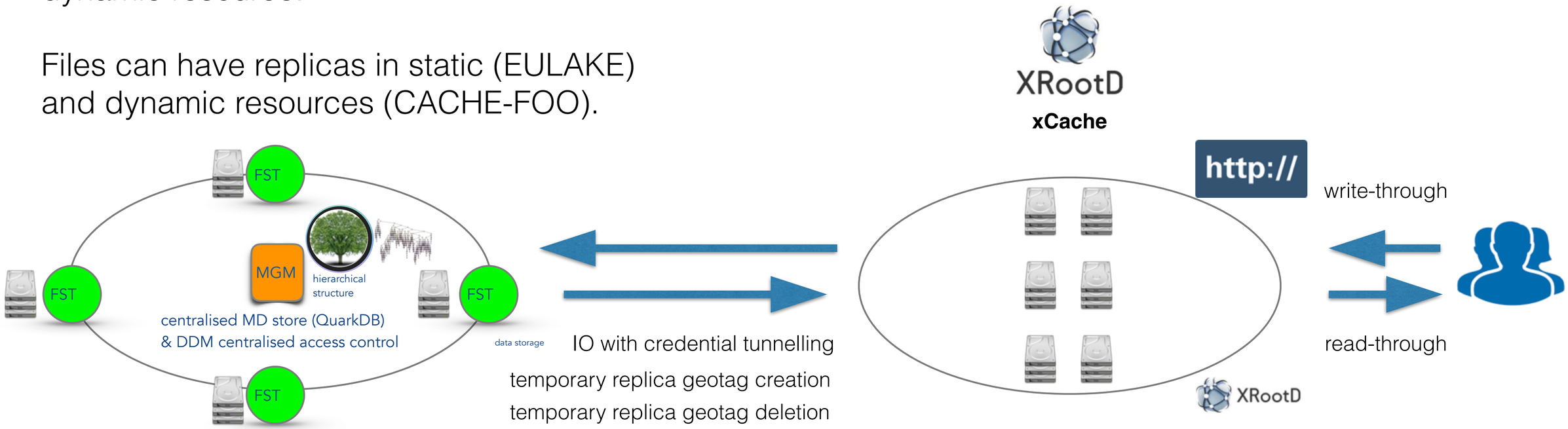
Enable Other Storage

- Scope of **EOS** in **XDC** & **WLCG** Datalake project
  - enable storage **caches**
  - 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

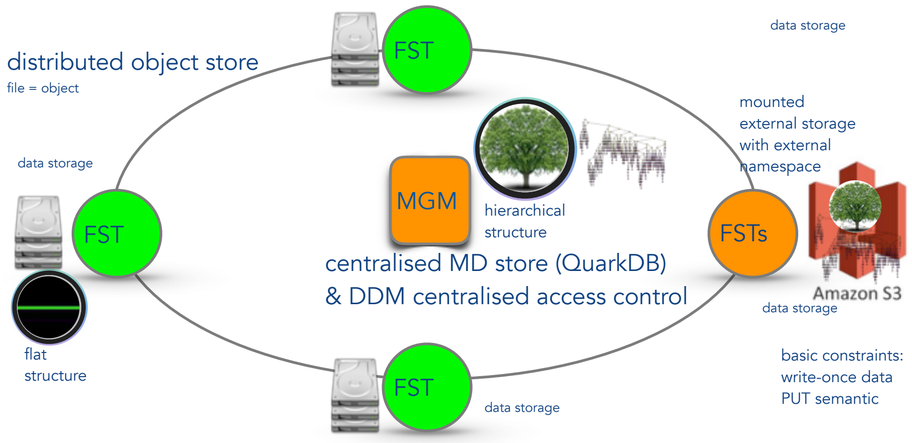
EULAKE

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

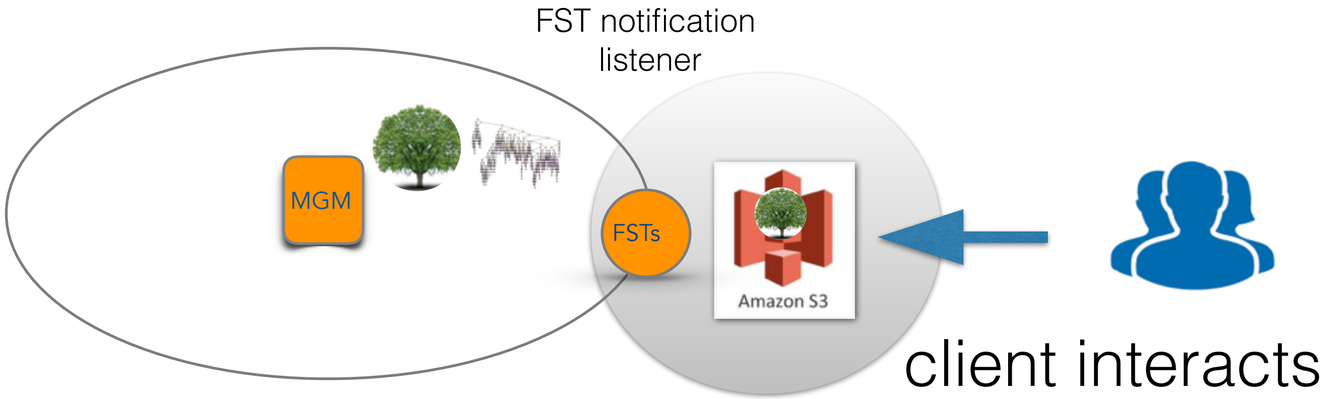
Amazon **S3**
CEPH **S3**
Shared **Filesystem** (with limitations)
ExOS **O**bject **S**torage (RADOS)
XRootD/Web**DAV+REST**

Hybrid Distributed Storage

Example: AWS Integration

- transparent S3 backup on tapes

Cern Tape Archive



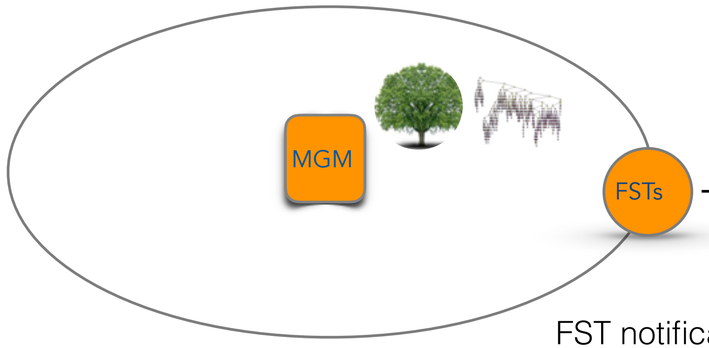
QOS policy
triggers CTA replication

client interacts
with AWS API

Hybrid Distributed Storage

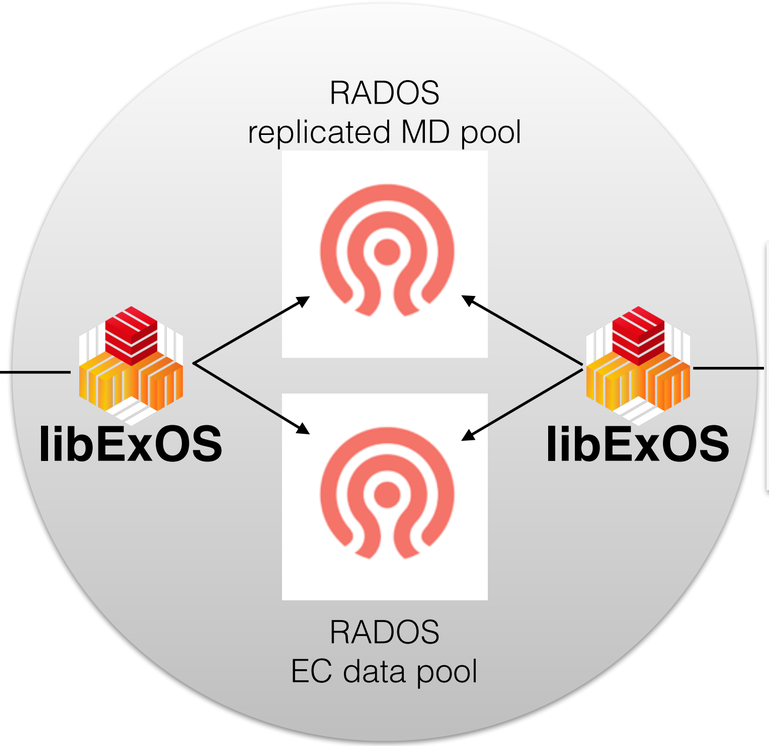
Example High Performance DAQ with Object Storage

Cern Tape Archive



QOS policy
triggers CTA replication

FST notification
listener



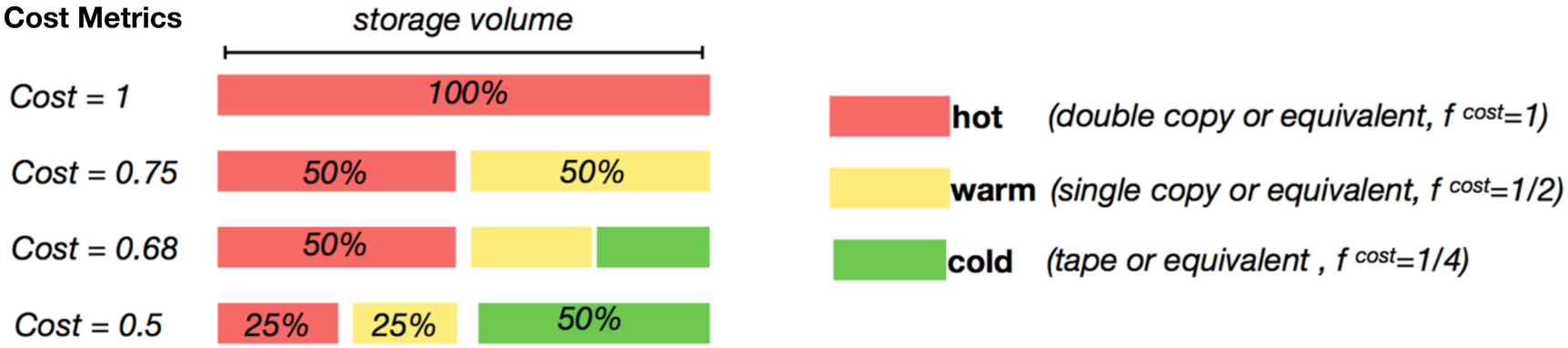
DAQ Farm

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 <=> 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.

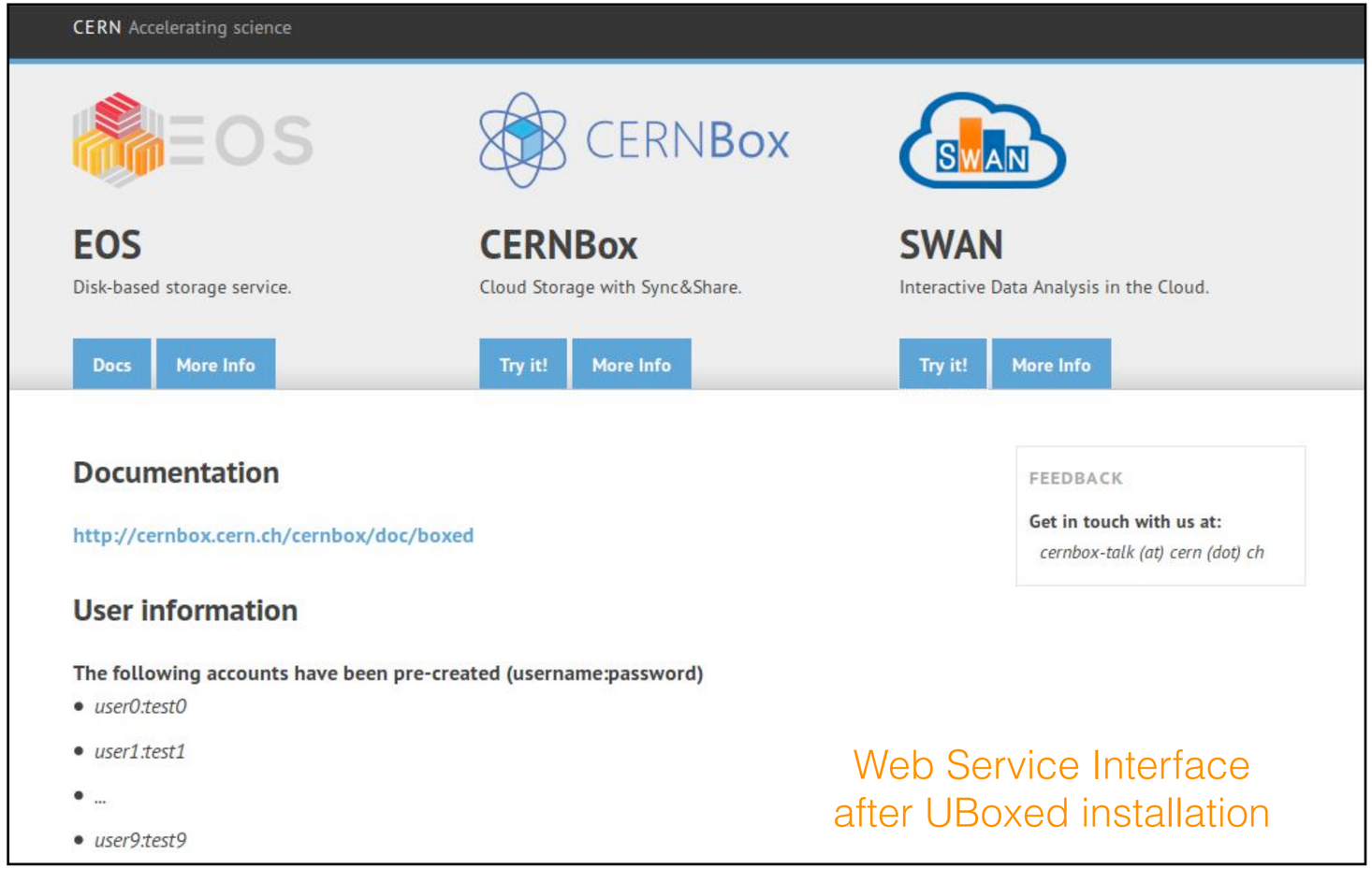
encapsulated four components

- 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

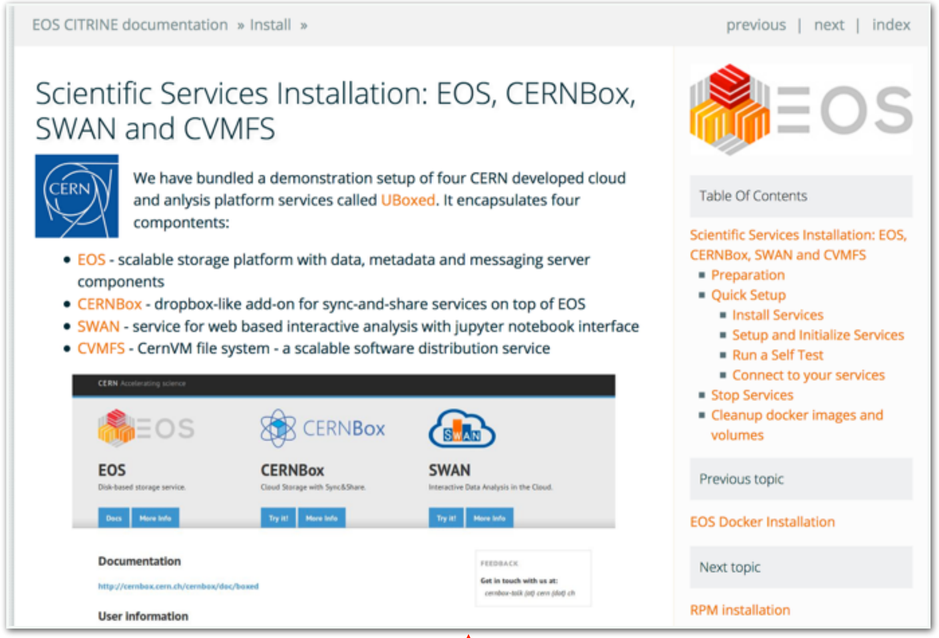
Try dockerized Demo Setup on CentOS7 or Ubuntu:

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

CERN Scientific Services Bundle



Web Service Interface
after UBoxed installation

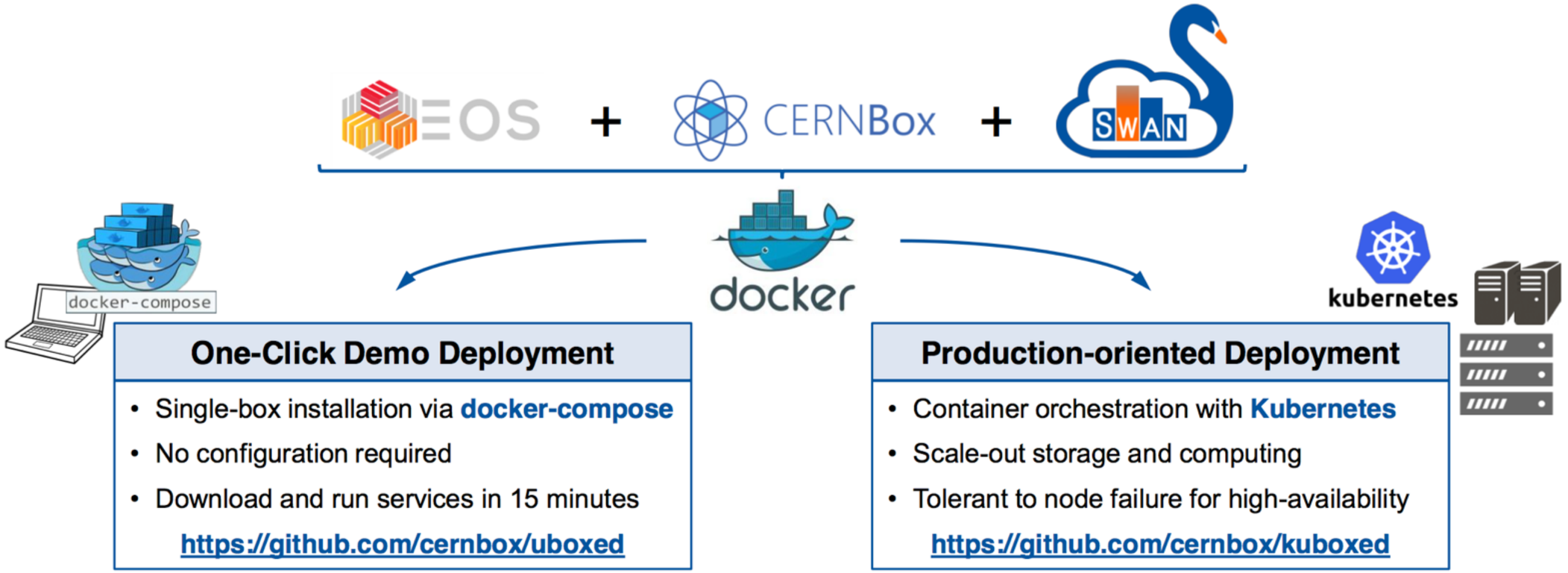


10 min

Try dockerized Demo Setup on CentOS7 or Ubuntu:

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

CERN Scientific Services Bundle



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 implemenation has many POSIX problems
- we implemented **3rd generation of a FUSE based client** for EOS

EOS as a filesystem /eos

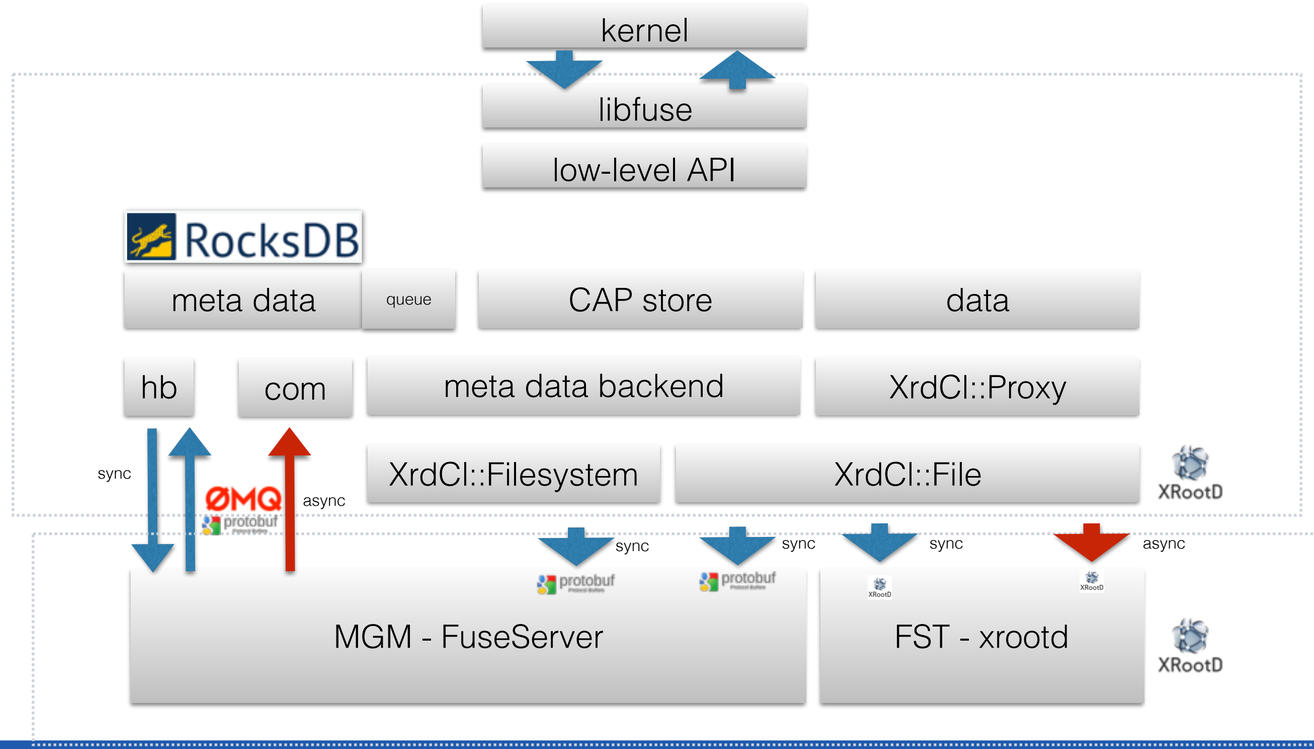
features

- 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

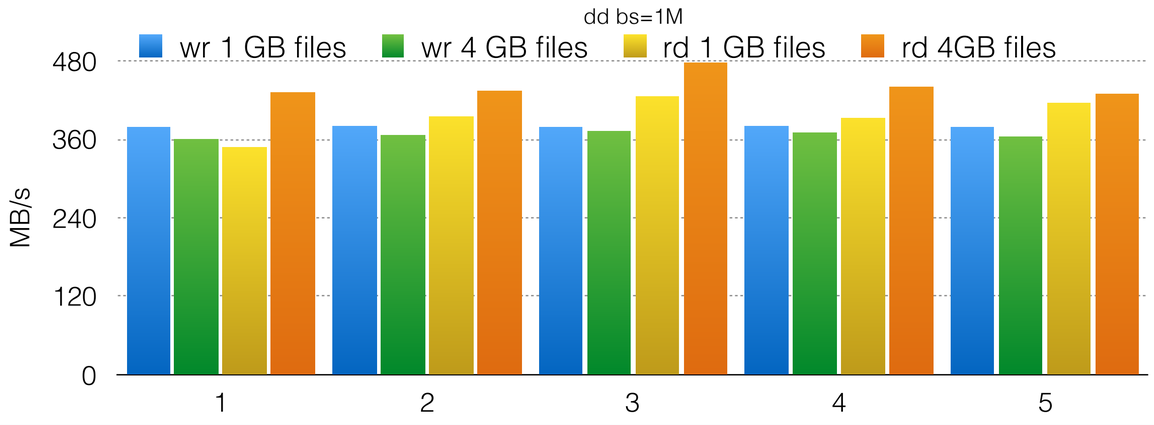
FUSE filesystem daemon

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
```



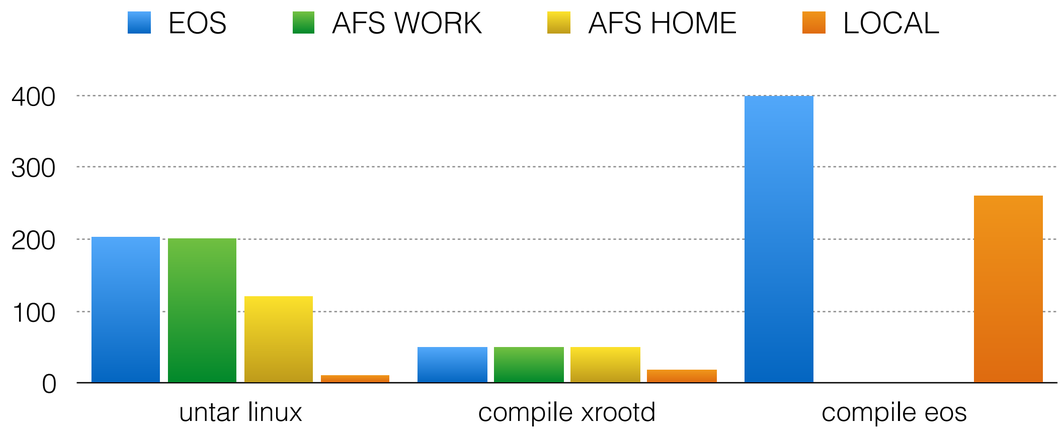
eosxd

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



commissioned to production at CERN during Q2/2018

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

EOS CITRINE latest release
20 of march - version 4.2.18

THANK YOU

QUESTIONS ?