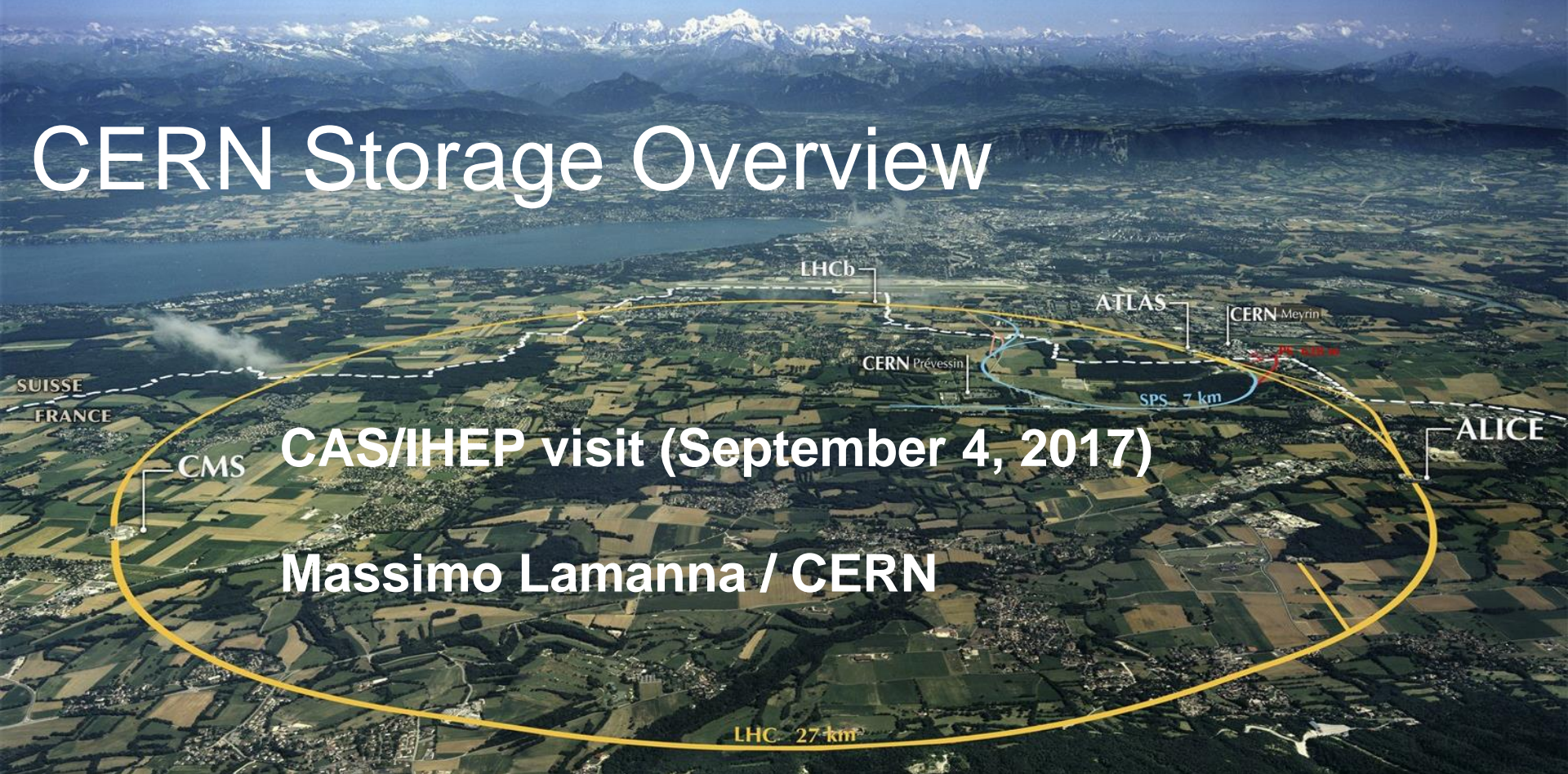




CERN Storage Overview



CAS/IHEP visit (September 4, 2017)

Massimo Lamanna / CERN



WLCG

Worldwide LHC Computing Grid

Make LHC computing possible

Worldwide infrastructure (collaboration) open to all LHC physicists
Computing/storage resources at CERN: ~ 20%; 80% across about 200 sites worldwide

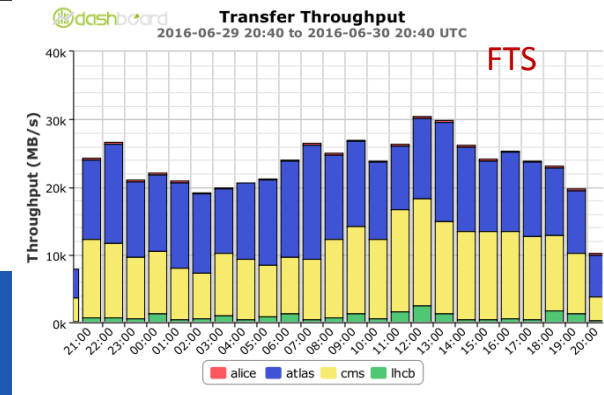
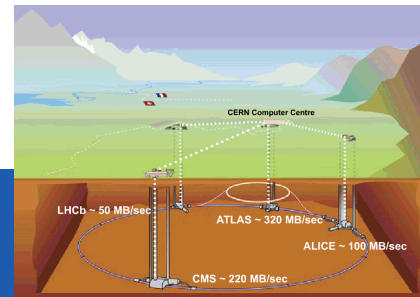
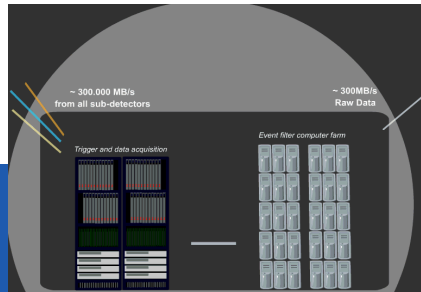
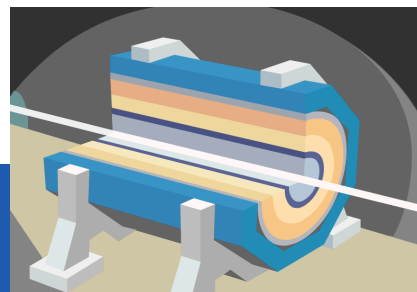
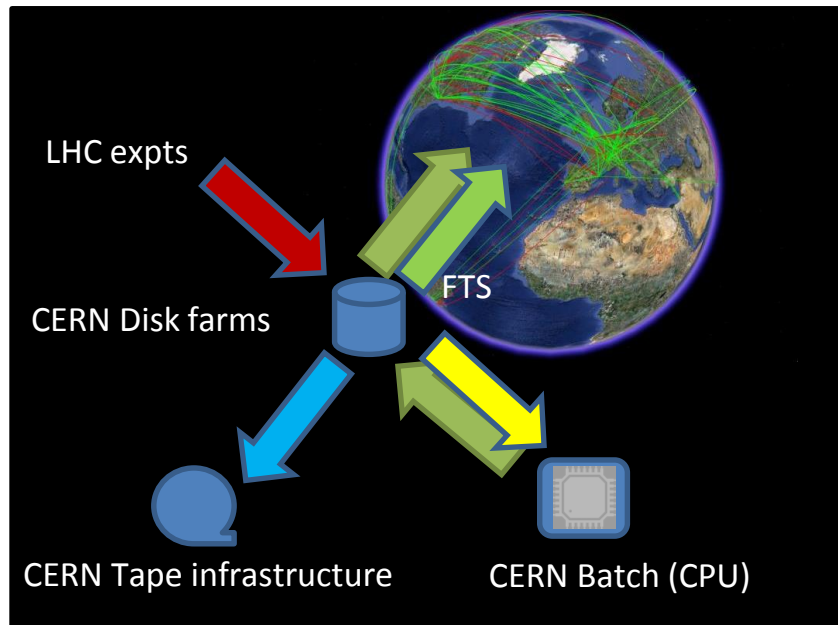
Data Reconstruction

Goals: data quality and immediate access for analysis
Organised activity dominated by heavy processing and replication (each expt: 1-8 GByte/s)

Data Analysis

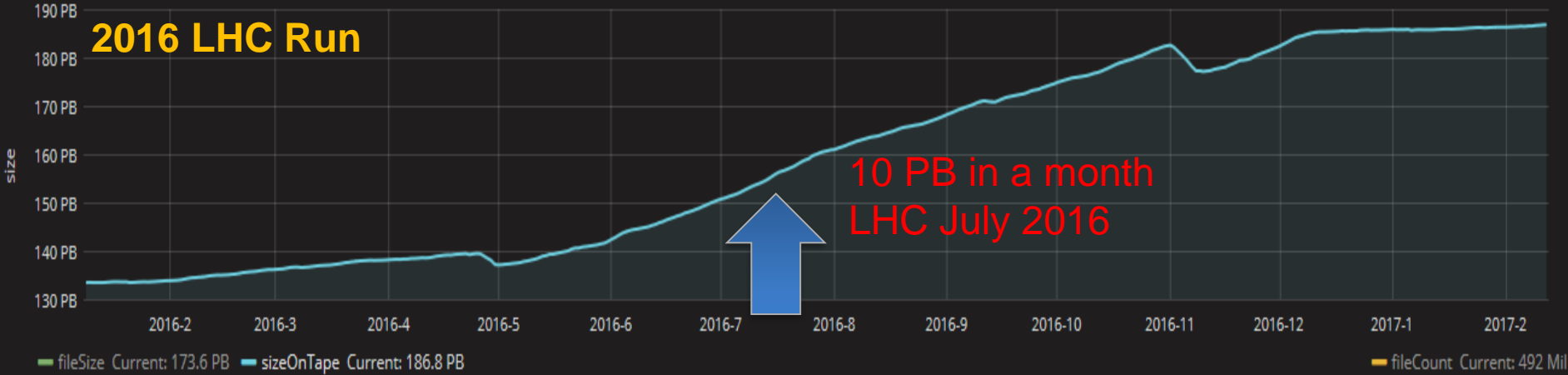
Goals: extract physics quantities (discovery)
Individual activities dominated by event selection and sharing (thousands of physicists)

(Detector) simulation

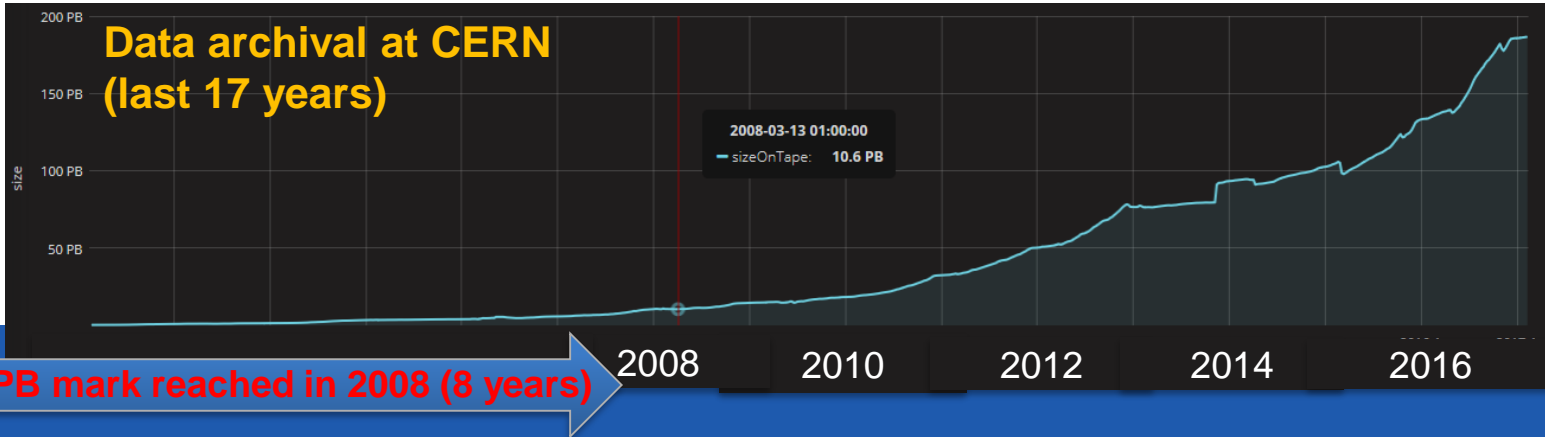


Physics Data in CASTOR

2016 LHC Run



Data archival at CERN (last 17 years)

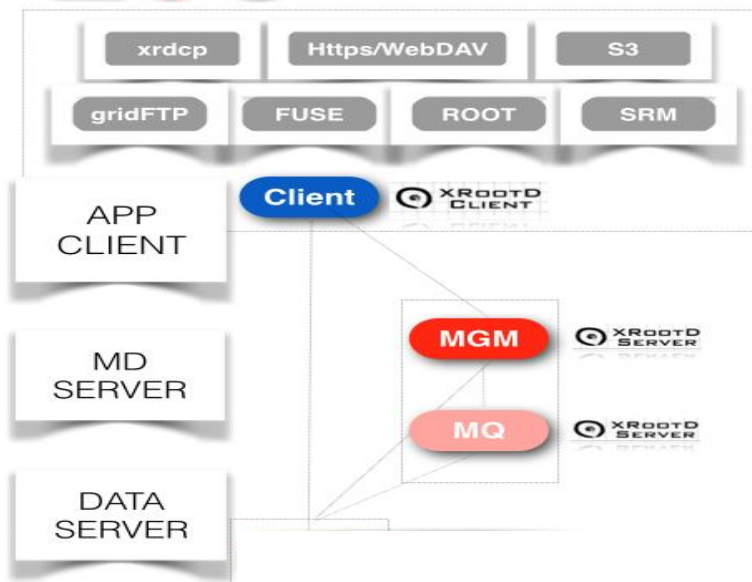


Content

- Storage (and environment) for physics (and for science at large)
 - EOS (storage) + CERNBox (sync and share) + SWAN (analysis)
- Storage for the CERN IT infrastructure
 - Ceph infrastructure



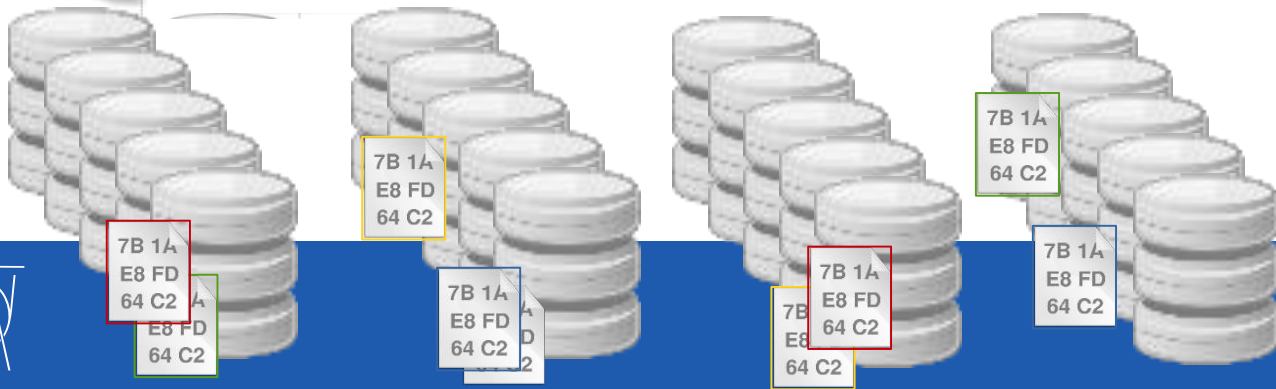
Open Source Storage



- › **EOS: Large disk farms for physics and beyond**
 - Developed at CERN
 - LHC: PBs for 100s/1000s independent scientists
 - 200 PB JBOD installed (CERN installations)

› Strategic points

- Distill **20+ years** of experience data management
- **Ultra-fast** name space
- Arbitrary level of data durability: cross-node file replication or RAIN on **commodity** hardware
- Large **protocol choice**



Number of Files

1390 M

Number of Directories

108 M

Write Throughput

2.390 GBps

Read Throughput

24.5 GBps

Current Readers

57.7 K

Current Writers

9.5 K

Total Space

166 PB

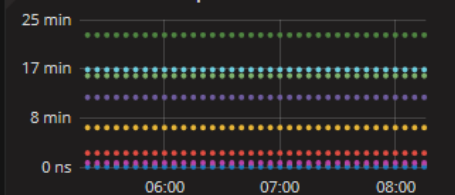
Free Space

43.81 PB

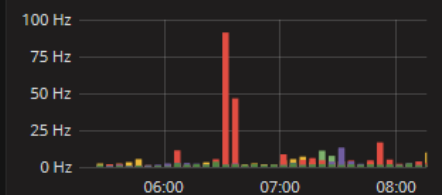
IOPS



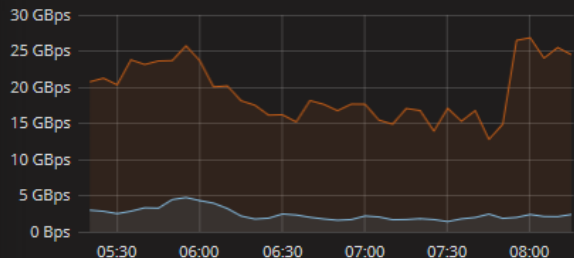
Namespace boot time



File deletion rate

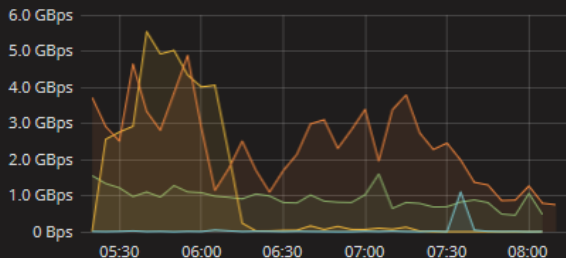


EOS Total IO



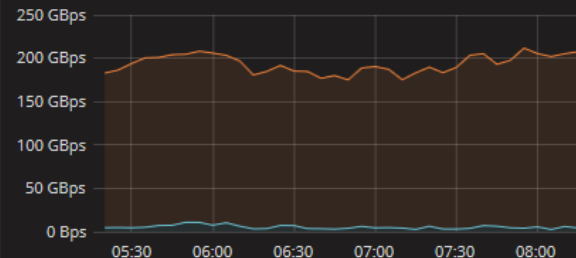
bytes_read Avg: 19.42 GBps bytes_written Avg: 2.43 GBps

EOS Total IO internal



balancing,0 draining,0 replication,0 gridftp,0

Aggregated Disk IO

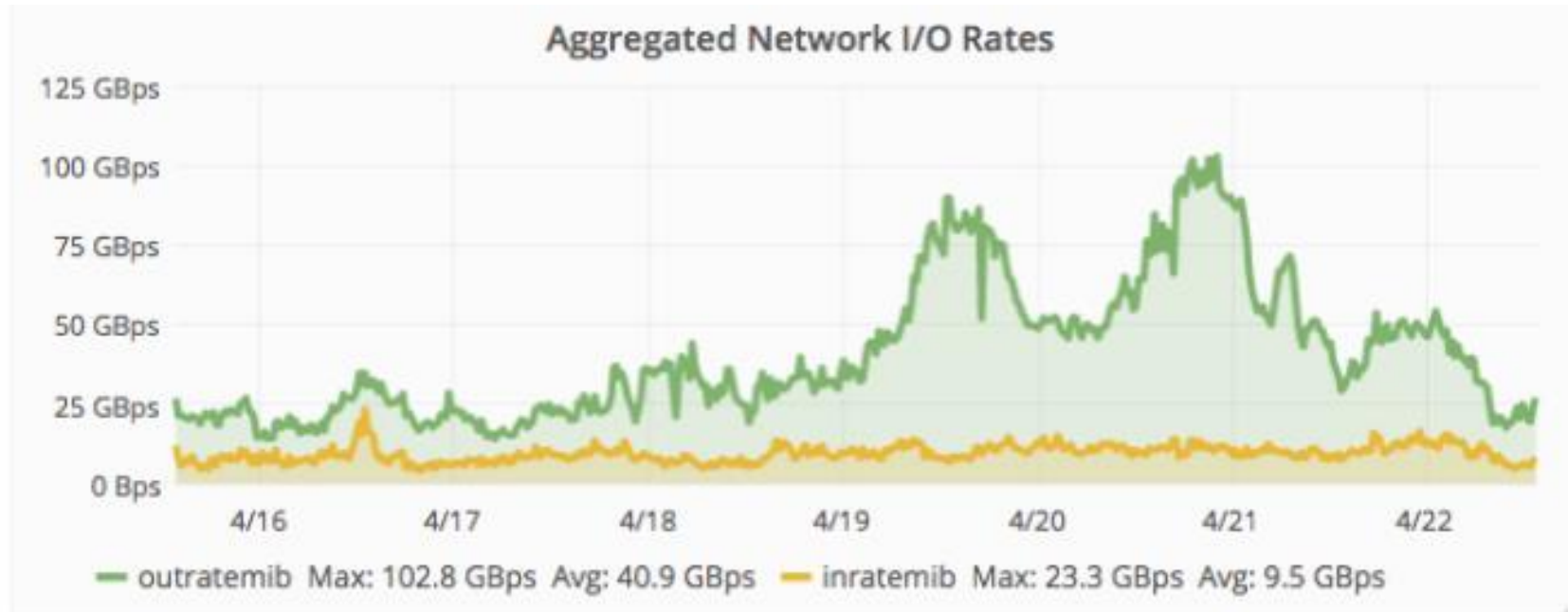


writerate Avg: 5.4 GBps readrate Avg: 188.1 GBps

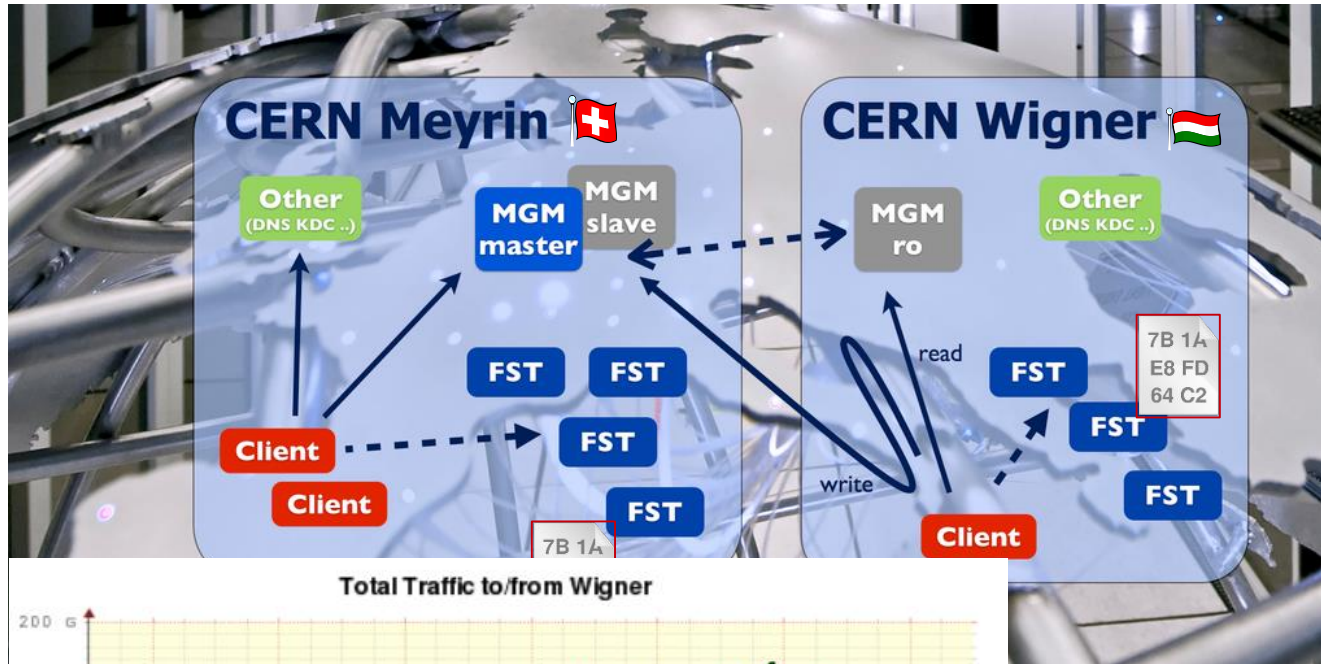


EOS Service

Storage for physics
And for general storage (CERNBox: see later)
Twin computer-centre deployment
3 · 100-Gb links (~22 ms latency)



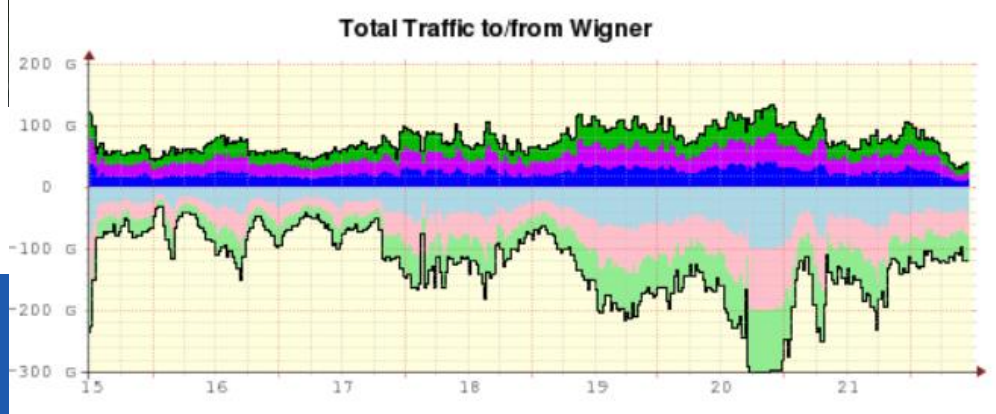
Our “20-ms-large” computer centre



Geneva – Budapest
 3 x 100 GB lines
 ~ 22 ms latency (diff routes)
 ~ 1000 km

Autonomic, Locality,
 Business continuity

Certainly more complex
 OK with 2 replicas, less interesting
 with other erasure codes



EOS evolution

- Resilient scalable catalogue well beyond 10B
Now 1.3 B entries
- High-performance POSIX access (Fuse)
- Archival capabilities (CTA)
- Extended usage in production of erasure code
 - Zero-operation mode
 - Cheaper hardware not impacting quality of service
- Collaboration with external sites
- HEP sites: Russian cloud, IHEP in Beijing, ...
- Other sciences/activities:
- JRC and AARNET best examples
- Evolution of the WLCG
 - Data federations

R&D

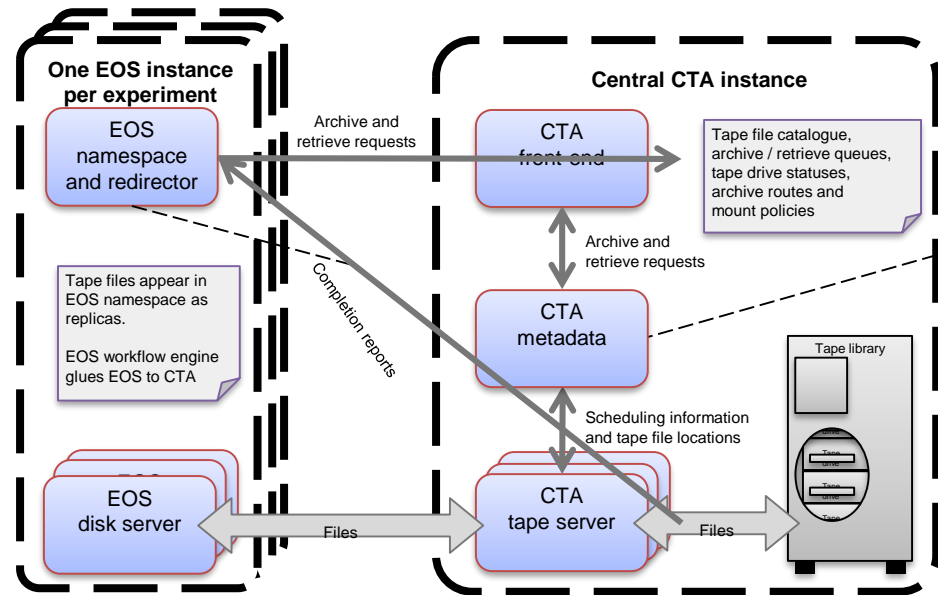
- CERN-IT extra-large disk server project
 - **8 x 24 x 6TB** disks connected to single front-end node [1.152 PB/node]
 - capacity/performance ratio ?
 - OS limitations handling 192 disks ?
 - RAID vs. ZRAID vs. Software EC
 - which network IF ?
 - which CPU type ?
 - TCO evaluation



CTA- CERN Tape Archive

A tape backend for EOS

- Removes duplication between current MSS (CASTOR) and EOS: namespace, file access and protocols, disk cache management
- Thin scheduling layer on top of existing CASTOR tape software
- EOS drives life cycle for archiving/restoring files from/to tape
- Same tape format as CASTOR – only need to migrate metadata
- Under development, aimed for LHC Run-3



Examples of collaboration

Joint Research Centre (JRC)

Science Service of the European Commission

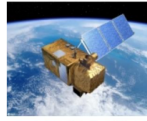


“Earth Observation & Social Sensing Big Data Pilot Project”

- The EU **Copernicus** Programme with the **Sentinel** fleet of satellites acts as a game changer by bringing EO in the Big Data era:
 - expected 10TB/day of **free and open** data
 - Requires new approaches for data management and processing
- Pilot project launched in January 2015
- Major goal: set up a central infrastructure for storing and processing of Earth Observation and Social Sensing data at JRC



Sentinel-1 (Credits: ESA/P. Carr)



Sentinel-2 (Credits: ESA/P. Carr)



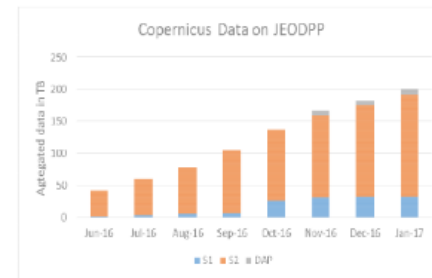
Sentinel-3 (Credits: ESA/J. Hua)

Joint
Research
Centre



EOS set-up at JRC

- Installation and configuration at JRC with strong support from CERN storage team
- Current set-up:
 - 1.4 PB gross capacity
 - 10 FST nodes, each with one JBOD of 24x6 TB disks
 - Using replica 2
- Further extension planned
 - 2017: extend to ~6 PB gross



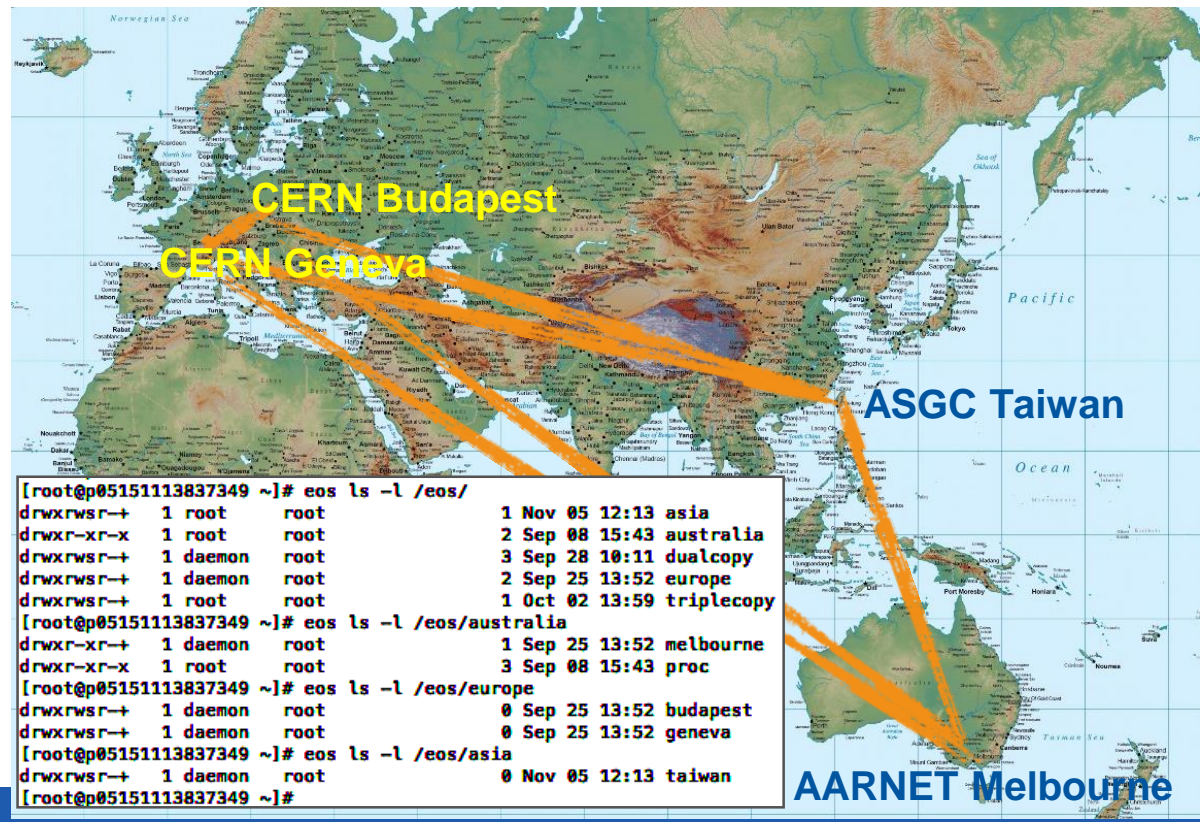
Joint
Research
Centre



A. Burger and P. Soille (JRC)

CERN,
AARNET,
ASGC
collaboration

Exploring the 300 ms region...



D. Jericho (AARNET), L. Mascetti (CERN),
Asa Hsu (ASGC Taipei)





National Research Centre (NRC)
"Kurchatov Institute"



Big Data Technologies Laboratory
<http://bigdatalab.nrcki.ru/>



Russian Federated Data Storage System Prototype

Andrey Kiryanov, Alexei Klimentov, Andrey Zarochentsev

on behalf of BigData lab @ NRC "KI" and
Russian Federated Data Storage Project

- HEP communities
 - Collaboration
 - Complementarity

- Federation

- Moscow area
- St Petersburg area (CERN)
- Sites from Russian Data Intensive Grid And WLCG site

- EOS workshop

A.Kyrianov et al.





National Research Centre (NRC)
"Kurchatov institute"



Big Data Technologies Laboratory
<http://bigdatalab.nrcki.ru/>




Russian Federated Data Storage System Prototype

Andrey Kiryanov, Alexei Klimentov, Andrey Zarochentsev


on behalf of BigData lab @ NRC "KI" and Russian Federated Data Storage Project

EOS Workshop, 2-3 Feb 2017

1

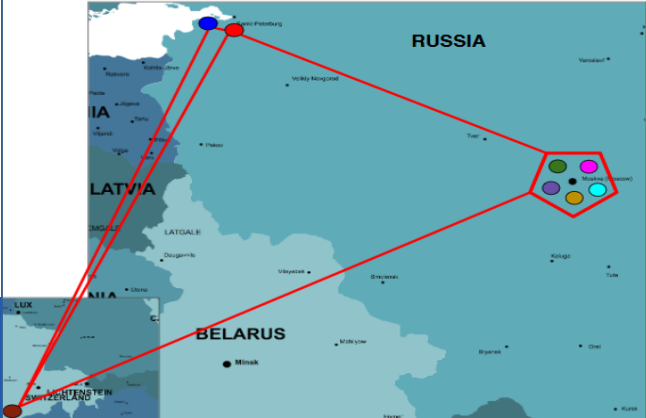


National Research Centre (NRC)
"Kurchatov institute"



Big Data Technologies Laboratory
<http://bigdatalab.nrcki.ru/>

Federation topology



SPb Region

- SPbSU
- PNPI

Moscow Region

- JINR
- NRC "KI"
- MEFPh
- SINP
- ITEP

and

- CERN

4

EOS AT 6,500 KILOMETRES WIDE

An Australian Experience
David Jericho – Solutions Architect, AARNET

SOLUTIONS WE HAVE TRIED



- Hadoop
 - MapR, Hortonworks, Apache official
- XtreamFS
- Ceph
- GlusterFS
- pNFS
- OrangeFS

... and others



SUCCESSES WE'VE HAD

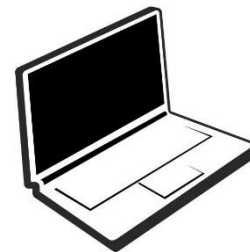
- IT WORKS!
 - Stable, server issues have been almost exclusively container related
 - Fast
- Obvious write latency penalty
 - Users don't notice
- Hello all, I know it's Monday...
 - CERN have been very responsive, THANKYOU!



EOS in DOCKER - 1 minute

- currently the docker scripts are only to get a **one machine instance** for testing
- the CERNBOX team is finishing a complete **dockerized CERNBOX**-like service package bundling EOS + OwnCloud
- prototyped a single host **ALICE docker** storage container with a pre-configured EOSALICE instance using the physical network inside the container
- interesting option to combine with **kubernetes** to simplify deployment in a storage federation integration is on the work plan ...
- if there is a broader interest, we can integrate the work of AARNET which is deploying EOS only via docker containers and add ALICE specifics

ALICE Tier 1/2/ Workshop 2017



1st EOS workshop (February 2-3 2017)

Participants



Open Source Storage



Integration and support

Disk technology

External Collaborators



Open Source Storage

Open Source Storage





CERNBox



CERNBox



- Starting point: Dropbox-like service
 - Cloud synchronisation service
 - Just the starting point!
- Innovative way to offer storage
 - Sync and share from ownCloud GmbH
 - EOS as a back-end (all LHC data!)
 - New way to interact with your data
 - HEP and beyond Broader scientific/university community

3rd Cloud Services for Synchronisation and Sharing (CS3)

Novel applications, cloud storage technology, collaborations

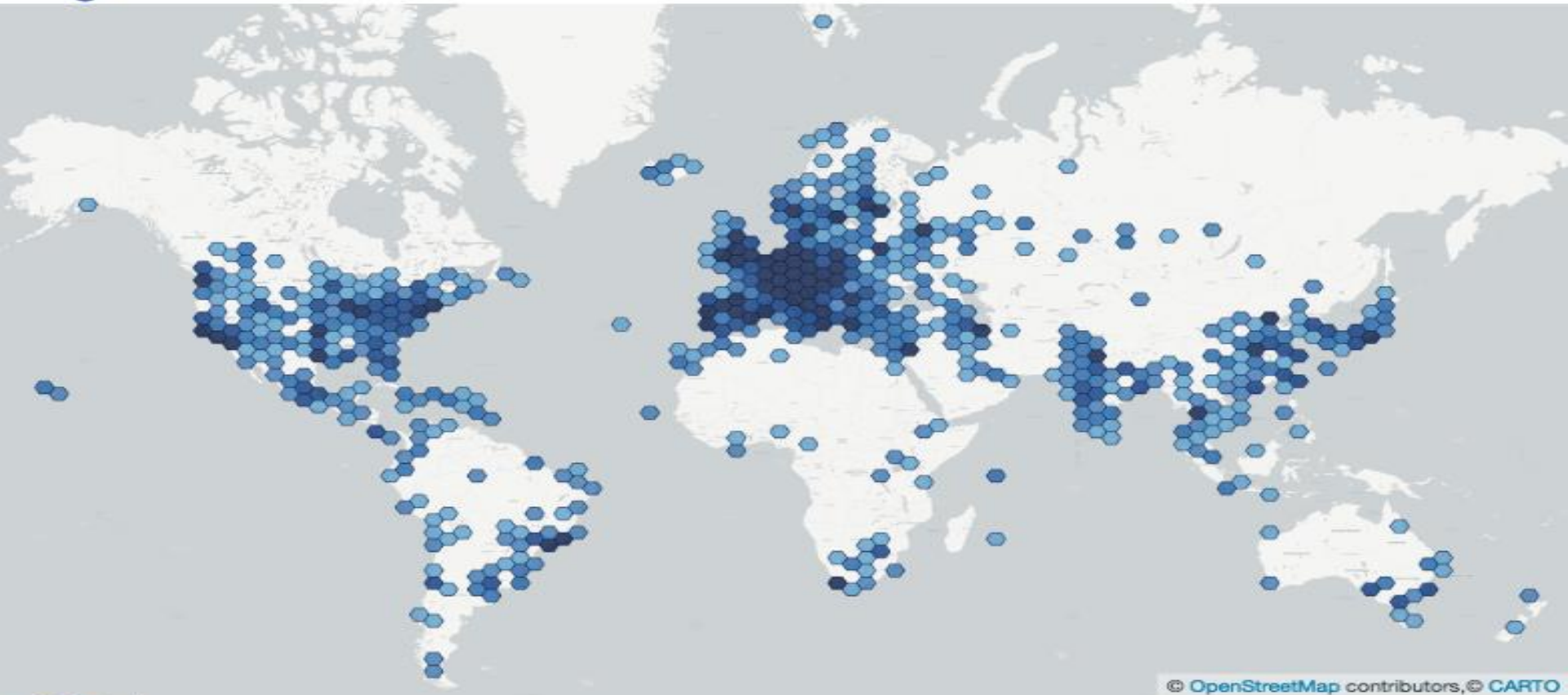
Amsterdam January 2017 *** 120+ participants *** 15+ companies

Krakow January 2018 ←





CERNBox Clients



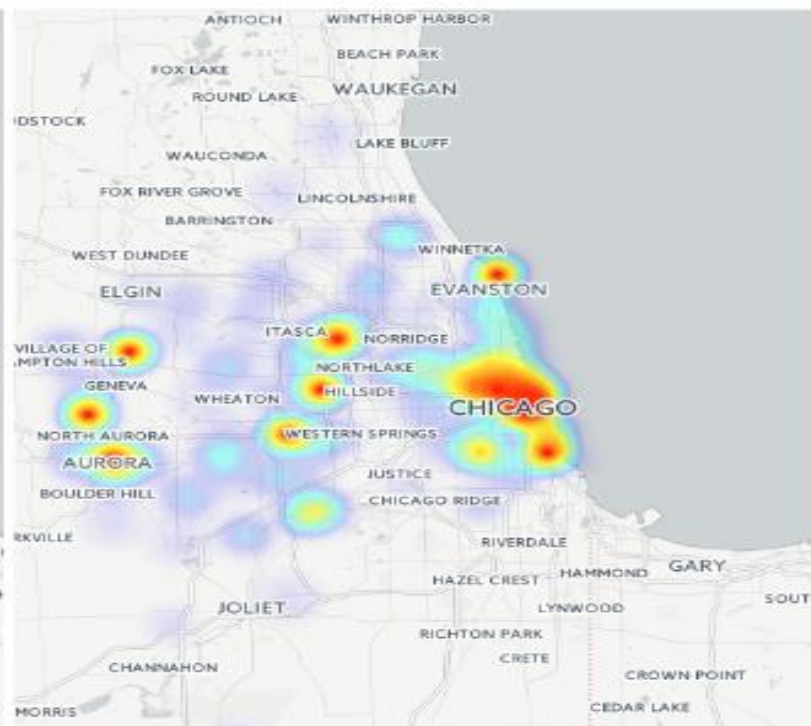
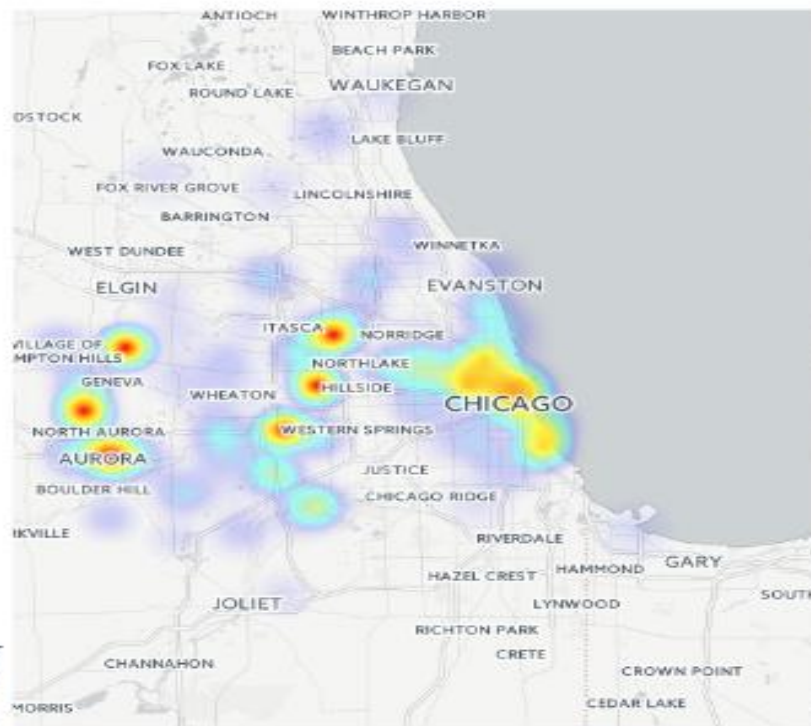
© OpenStreetMap contributors, © CARTO





38th INTERNATIONAL CONFERENCE ON HIGH ENERGY PHYSICS

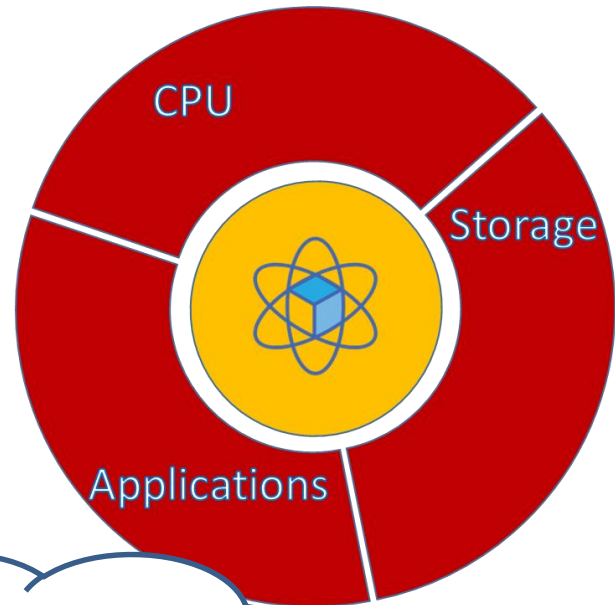
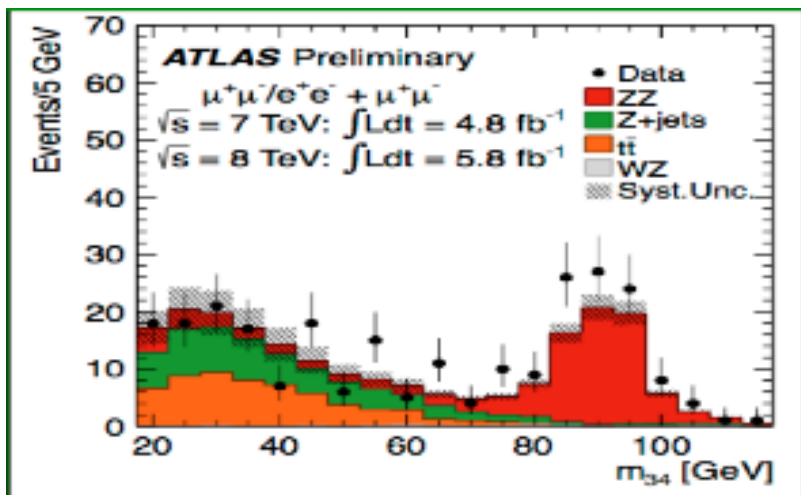
AUGUST 3 - 10, 2016
CHICAGO





Cloud analysis: SWAN project

with CERN Physics Department



Lots of activity in previous projects with several Russian groups, notably with V. Korenkov (JINR Dubna)

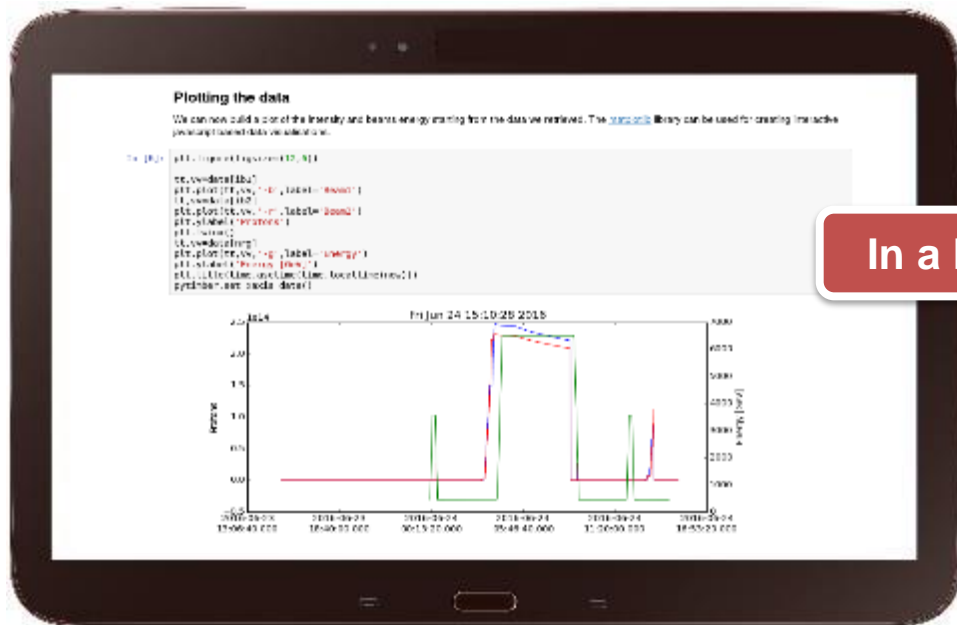


ROOT is the CERN data analysis framework: <http://root.cern.ch>



Interface: The Notebook

Jupyter Notebook: A web-based **interactive computing** interface and platform that combines **code**, **equations**, **text** and **visualisations**



In a Browser



Interface: The Notebook

Text

Code

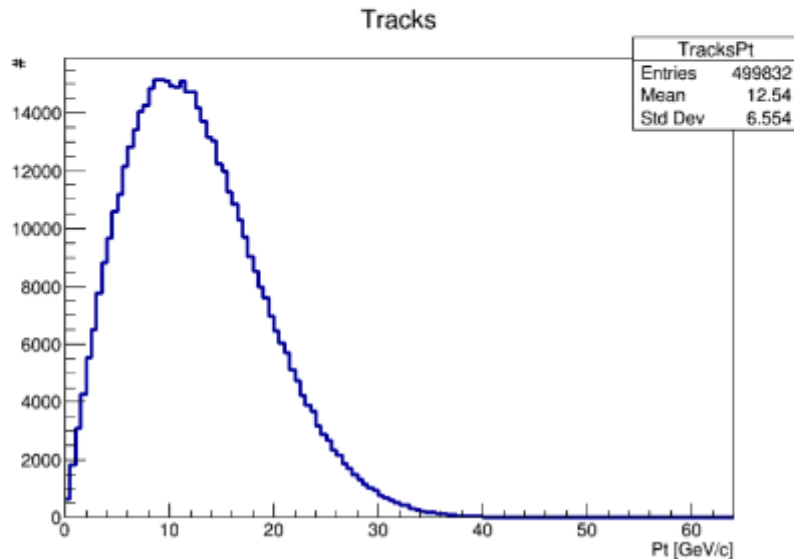
Graphics

Access TTree in Python using PyROOT and fill a histogram

Loop over the TTree called "events" in a file located on the web. The tree is accessed with the dot operator. Same holds for the access to the branches: no need to set them up - they are just accessed by name, again with the dot operator.

```
In [1]: import ROOT

f = ROOT.TFile.Open("http://indico.cern.ch/event/395198/material/0/0.root");
h = ROOT.TH1F("TracksPt", "Tracks;Pt [GeV/c];#", 128, 0, 64)
for event in f.events:
    for track in event.tracks:
        h.Fill(track.Pt())
c = ROOT.TCanvas()
h.Draw()
c.Draw()
```





CERNBox as Home



Control Panel

Logout

Files

Running

Clusters

Select items to perform actions on them.

Upload

New ▾



<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	ACAT 2016	
<input type="checkbox"/>	CHEP 2016	
<input type="checkbox"/>	cmsdata	
<input type="checkbox"/>	CSC	
<input type="checkbox"/>	ExampleDir	
<input type="checkbox"/>	IMLmeeting	
<input type="checkbox"/>	mylibs	
<input type="checkbox"/>	node_modules	
<input type="checkbox"/>	other	
<input type="checkbox"/>	ROOT-Primer	

Same content as in cernbox.cern.ch



SWAN Use Cases

```
title = { "model": "Signal" , "pdfBkg" : "Partially reconstructed" , "cmbBkg": "Combinatorial background"}

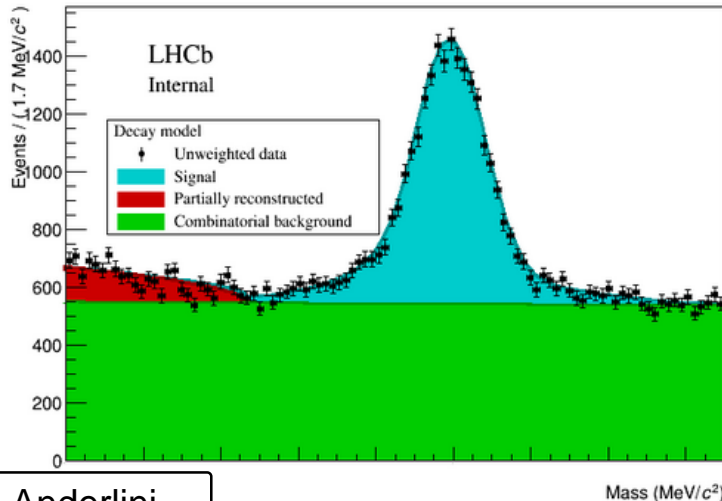
for (component, color) in [ ("model",kCyan), {"pdfBkg",kRed}, {"cmbBkg",kGreen}]:
  model.plotOn (frame, LineColor(color+2) , DrawOption('L') , Components(component) , LineWidth(5))
  model.plotOn (frame, FillColor(color+1) , DrawOption('F') , Components(component) , LineWidth(0) , Name("P"+component)
  ))
  leg.AddEntry ( frame.findObject ("P"+component), title[component] , "P" )

data.plotOn ( frame, MarkerColor ( ROOT.kBlack ) )
frame.Draw()
Graphics().lhcbMarker(0.2,0.8, "Internal")

leg.Draw()

ROOT.gPad.Draw()
```

Results coming
from real data!
(published now)



Physics Analysis

Rare B meson decay in LHCb

- Read data from EOS
- Setup complex fit
- Document and inspect results



- SWAN as platform for outreach
 - Introductory course about experimental HEP for future high school teachers

Particle open data teaching (Hiukkasfysiikan avoin data opetuksessa)

The screenshot shows a Jupyter Notebook window titled "Esim-pseudorapiditeetti-mittatarkkuus (autosaved)". The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Help), a toolbar with icons for file operations and execution, and a "Python 2" kernel indicator. The main content area contains a text cell with the following text:

Lähdetäänpä tutkimaan!

Lähdetään seuraavaksi tarkastelemaan, miten pseudorapiditeetin vaikutus mittatarkkuuteen voidaan havaita CMS-ilmäisimen keräämän oikean datan avulla. Käytetään CMS:n vuodelta 2011 kerättyä dataa [1], josta on valittu 10851 törmäystapahtumaa (events) tiedostoon "Zmumu_Run2011A_massoilla.csv". (Karsinta on suoritettu koodilla, joka on avoimesti saatavilla osoitteessa <https://github.com/tpmccauley/dimuon-filter>.)

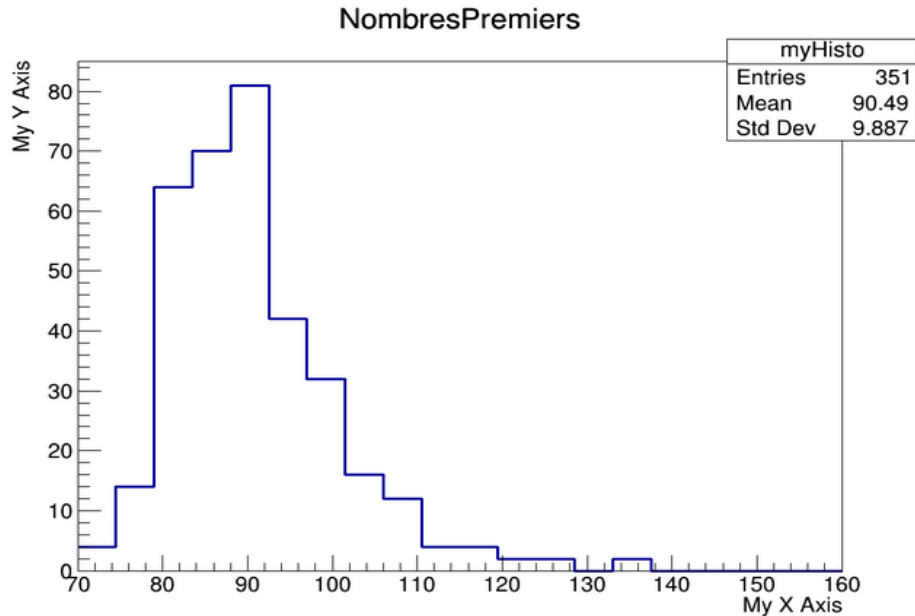
Tiedostoon on valittu niitä törmäystapahtumia, joissa syntynyt Z-bosoni on hajonnut myoniksi μ^- ja antimyoniksi μ^+ . Ilmaisimien on havainnut nämä myonit ja mitannut niiden liikemäärät.

Below the text is a Feynman diagram showing a Z boson (Z^0) decaying into a muon (μ^-) and an antimuon (μ^+).

In [138]:

```
import ROOT
htemp = ROOT.TH1F("myHisto", "NombresPremiers;My X Axis;My Y Axis", 20, 70, 160)
for i in range(len(data)):
    d = data[i][0]
    htemp.Fill(float(d))
c = ROOT.TCanvas("myCanvas", "myCanvasTitle", 1024, 768)
htemp.Draw()
c.Draw()

TROOT::Append(0: RuntimeError: Replacing existing TH1: myHisto (Potential memory leak).
TCanvas::Constructor:0: RuntimeError: Deleting canvas with same name: myCanvas
```



Up to University (Up2U)

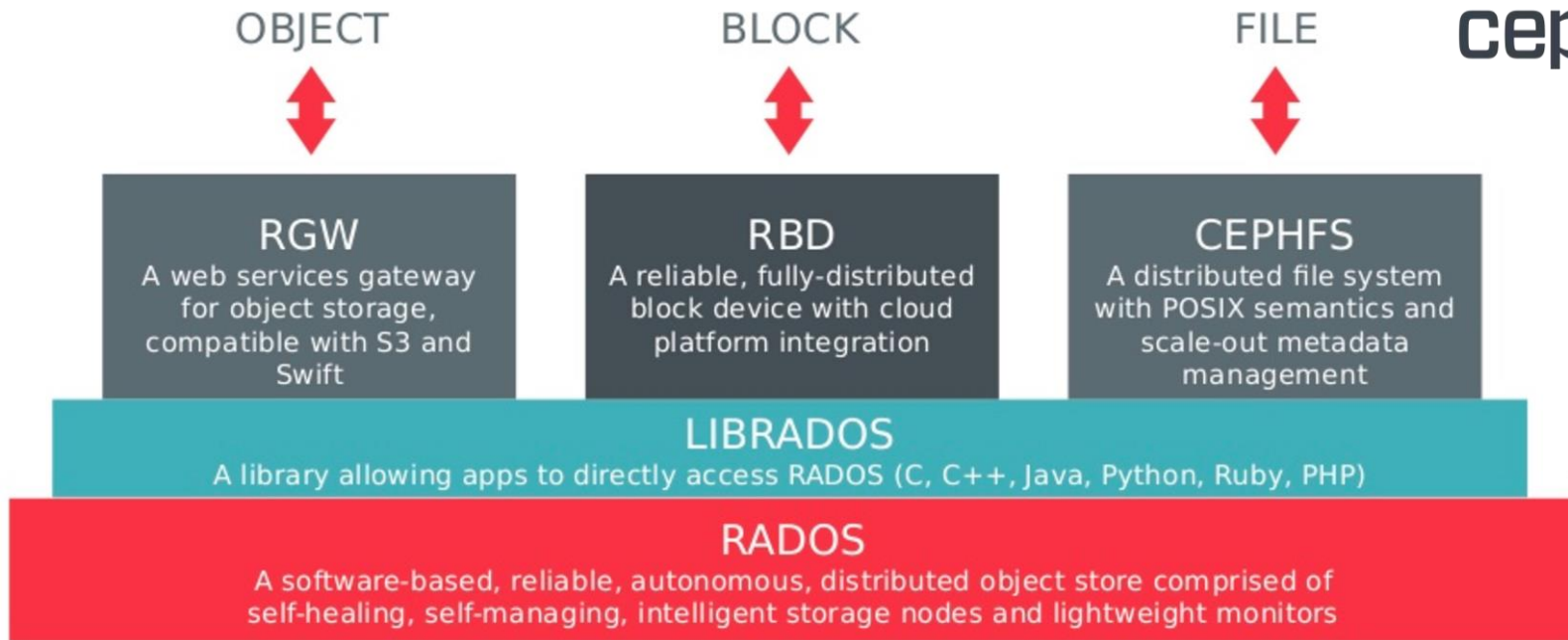
Bridging the gap between schools and universities through informal education

Mano S. (14 years old), K12 student

- Approaches programming for the first time
- Verifies numerically what he learned at school
- Shares results with his supervisor and classmates

Ceph for CERN IT infrastructure

Ceph at a Glance



Ceph/CERN Timeline

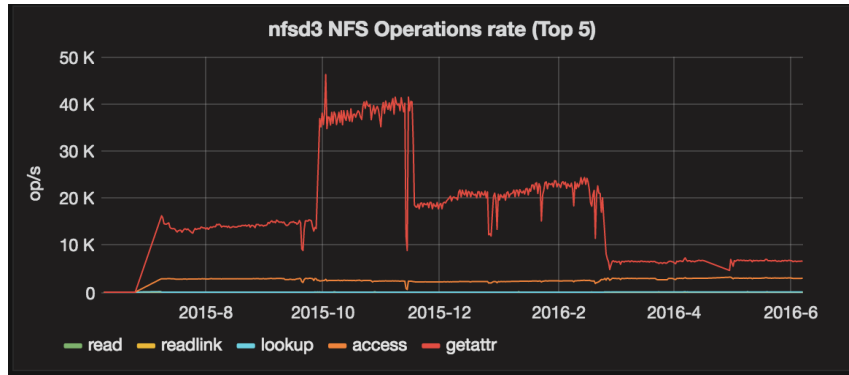
- Jan 2013: CERN IT needs block storage for OpenStack
 - 1H 2013: testing Ceph on old hardware.
 - 2H 2013: Installed 3PB new Cephcluster. Using 200 SSDs for OSD journaling.
 - Dec: OpenStack Cinder/Ceph in production.
- 2014-2015: stable operations.
 - Steady growth in used space and IOPS.
 - Ceph Advisory Board formed in Oct 2015 – CERN is a founding member.
- 2016: Replaced original 3PB Ceph with ~6PB new hardware.
 - CephFS starting with HPC use-cases.
- 2017: CephFS scale testing, testing OpenStack Manila.

RADOS Block Devices

- RBD: a virtual block device (similar to iSCSI) usable by VMs or recent Linux kernels.
 - Thinly provisioned, highly available, extendable, snapshottable.
 - QoS throttling: 100 IOPS for “standard” volumes, 500 IOPS for “io” volumes.
 - rbd-mirror for disaster recovery, iSCSI gateways available.
- Two clusters:
 - Meyrin: 5.5PB raw, 2.5PB used. 3 replicas, 2 rooms.
 - Wigner: 422TB, 146TB used. 3 replicas, used mostly for DR replication.

NFS on RBD

- ~60TB across 28 servers:
- OpenStack VM + RBD
- CentOS 7 with ZFS for DR
- *Not highly-available, but...*
- cheap, thinly provisioned, resizable, trivial to add new filers



Example: ~25 puppet masters reading node configurations at up to 40kHz



Ceph File System

- CephFS is stable since mid-2016
 - “stable” means that recovery tools exist to help if things go wrong.
- Horizontally scalable metadata now possible in Ceph 12.2.0.
- Snapshots not yet stable.

- 100 HPC nodes accessing CephFS since mid-2016. Few bugs found, quite stable.
 - Parallel IO, range locking, needs some dev work. Best to write to unique files.
- Planning to migrate our RBD/NFS solution to CephFS + Manila (see Arne’s slides).

- NFS and Samba gateways provided by the community.

Ceph Board

- In October 2015, the Ceph Advisory Board was formed to assist the community in driving the direction of open source software-defined storage technology.



Leonardo Vaz	Red Hat	Chair
Sage Weil	Red Hat	Architect
Dan van der Ster	CERN	Academic Rep
Wido den Hollander	42on	User Chair
Christian Reis	Canonical	Commercial
Seth Mason	Cisco	Commercial
Paul von Stamwitz	Fujitsu	Commercial
Anjaneya Chagram	Intel	Commercial
Allen Samuels	Sandisk	Commercial
Lars Marowsky-Brée	SUSE	Commercial

Recap

LHC challenging problems

Scale (price) and performances as a driver

Push the envelope of technology

In house (EOS) or from outside (Ceph, OwnCloud, Jupiter)

Important role exploring/prototyping

Attracting large communities around us (collaborations)



www.cern.ch

