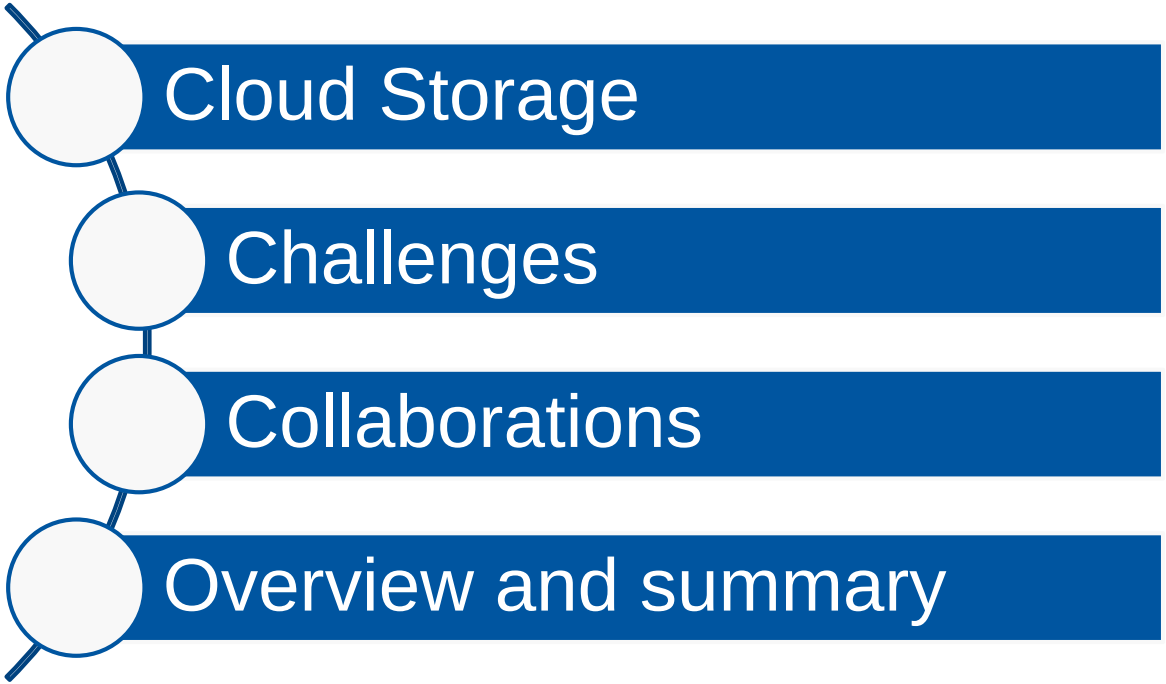
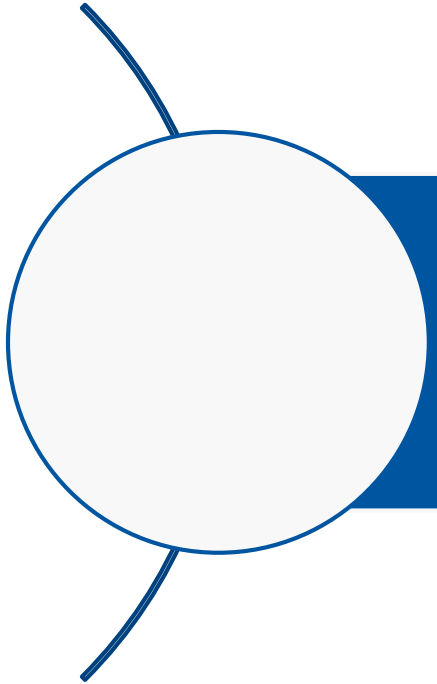

CERN

Cloud Storage for Data Intensive Sciences in Science and Industry

H. Labrador
E. Bocchi, D. Castro, M. Lamanna, L. Mascetti, K. Mosciscki, A. Peters, D. Piparo, E. Tejedor

Outline


Cloud Storage
Challenges
Collaborations
Overview and summary



Cloud Storage

About cloud storage

- From Cloud Storage to Cloud Sync & Share to Collaborative Platforms
- Commercial clouds like Dropbox are difficult to adapt to our environment for sync and share as they do not have direct access to our vast scientific data collection and to our existing integrations inside our infrastructure.
- Self-hosted clouds (ownCloud, Pydio, NextCloud, SeaFile, PowerFolder …)
  - More privacy, more control, Open Source, deploy on top of your system
- These platforms are part of the IT Service Portfolio of most of universities and research centers

Collaborative Platforms in to the Main Data Workflow

- We are currently expanding the original functionalities of cloud storage to create immersive collaborative platforms
- The challenge now if to closely integrate these platforms into the main data workflow to optimize the scientific analysis
- But this challenge is getting bigger and bigger than ever


Challenges

INTERNATIONAL JOURNAL OF HIGH-ENERGY PHYSICS

CERN COURIER

CERN COURIER APRIL 2018

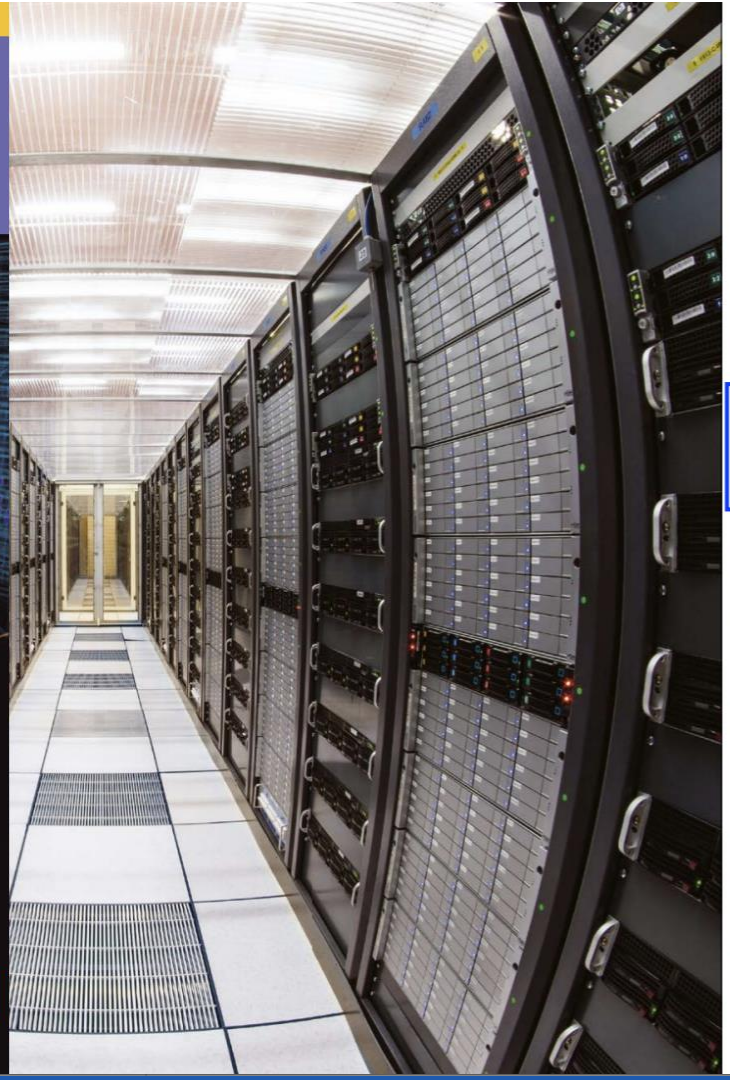
VOLUME 58 NUMBER 3 APRIL 2018

Tracking beauty decays

X-band technology spreads

Computing's radical future

Speaking up for the Higgs



Software and computing

Time to adapt for big data

Radical changes in computing and software are required to ensure the success of the LHC and other high-energy physics experiments into the 2020s, argues a new report.

It would be impossible for anyone to conceive of carrying out a particle-physics experiment today without the use of computers and software. Since the 1960s, high-energy physicists have pioneered the use of computers for data acquisition, simulation and analysis. This hasn't just accelerated progress in the field, but driven computing technology generally – from the development of the World Wide Web at CERN to the massive distributed resources of the Worldwide LHC Computing Grid (WLCG) that supports the LHC experiments. For many years these developments and the increasing complexity of data analysis rode a wave of hardware improvements that saw computers get faster every year. However, those blissful days of relying on Moore's law are now well behind us (see panel overleaf), and this has major ramifications for our field.

The high-luminosity upgrade of the LHC (HL-LHC), due to enter operation in the mid-2020s, will push the frontiers of accelerator and detector technology, bringing enormous challenges to software and computing (*CERN Courier* October 2017 p5). The scale of the HL-LHC data challenge is staggering: the machine will collect almost 25 times more data than the LHC has produced up to now, and the total LHC dataset (which already stands at almost 1 exabyte) will grow many times larger. If the LHC's ATLAS and CMS experiments project their current computing models to Run 4 of the LHC in 2026, the CPU and disk space required will jump by between a factor of 20 to 40 (figures 1 and 2).

Even with optimistic projections of technological improvements there would be a huge shortfall in computing resources. The WLCG hardware budget is already around 100 million Swiss francs per year and, given the changing nature of computing hardware and slowing technological gains, it is out of the question to simply throw

Inside the CERN computer centre in 2017. (Image credit: J Ordan/CERN.)

New computing (re) evolution

- Interesting new projects are coming:
  - High-Luminosity LHC, DUNE, SKA
- They bring another dimension to computing characterized by:
  - Data explosion
  - Computing explosion
  - More and more distributed user communities
- We are not alone! This characteristics apply also to other sciences:
  - Finance, Biology, Genomics ….

How to face it?

- New technologies!
  - Rethink the way HEP does data analysis
  - Natural sharing capabilities and seamless integration of large facilities with private resources
- Collaboration with industrial partners to get pieces for our solutions
  - Re-use as much as possible
- Why our solutions are not as easy as Dropbox?
  - Attract brilliant students to enhance these platforms
- We know how to build bigger data centers, we should optimize the human part of the process to make it more efficient

New HEP-developed technologies: the Science Box

- EOS: CERN Open Storage
  https://indico.cern.ch/event/587955/contributions/2936837/

- CERNBox: The CERN Cloud Storage
  https://indico.cern.ch/event/587955/contributions/2936817/

- SWAN: CERN Service For Web Data Analysis
  https://indico.cern.ch/event/587955/contributions/2935943/

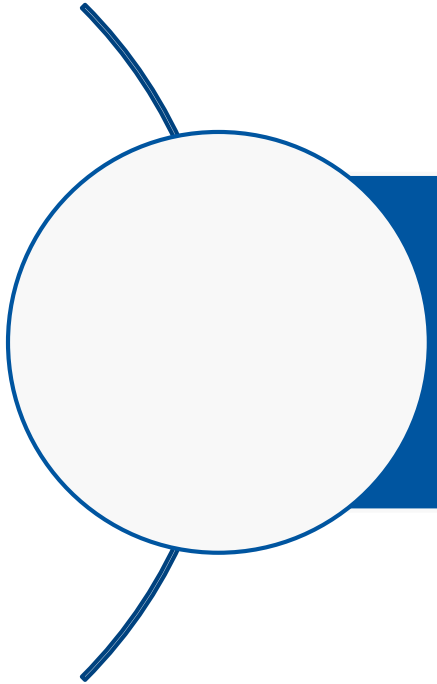
- CVMFS: The CernVM File System
  https://indico.cern.ch/event/587955/contributions/3012720/

New HEP-developed technologies: the Science Box

- Science-box deployment running on multiple clouds
  - Amazon Web Services
  - Helix Nebula Cloud (T-Systems & IBM)

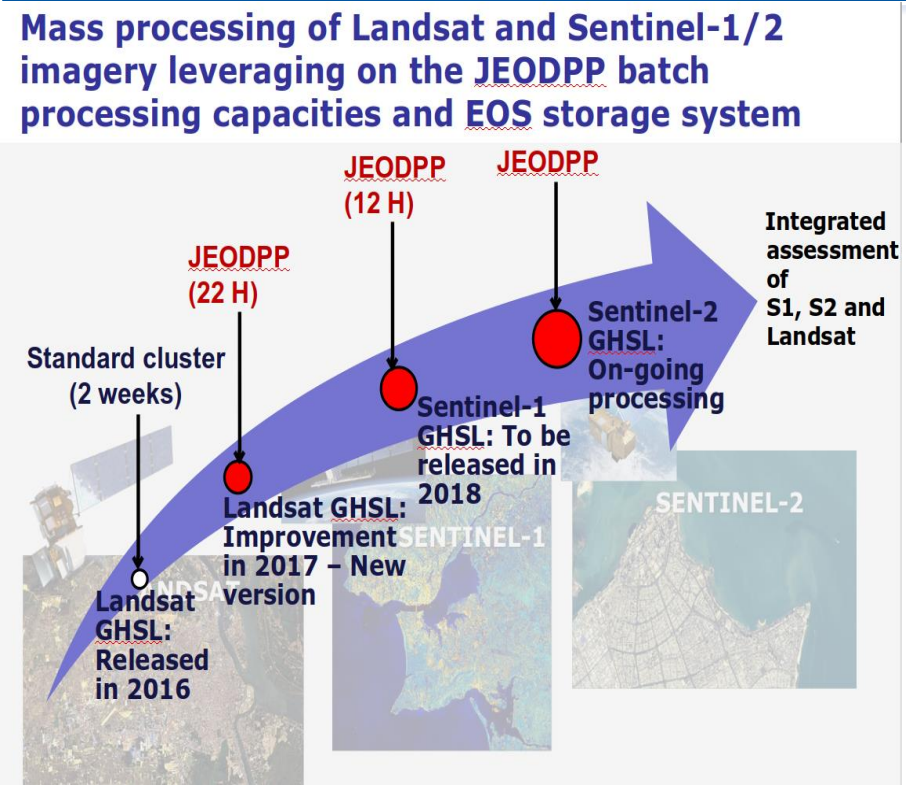
- Production-oriented deployment with Kubernetes on
  - OpenStack at CERN
  - CERN Container Service (on-going effort)

- SWAN connection to Spark clusters in the TOTEM experiment for online massive data processing



Collaborations

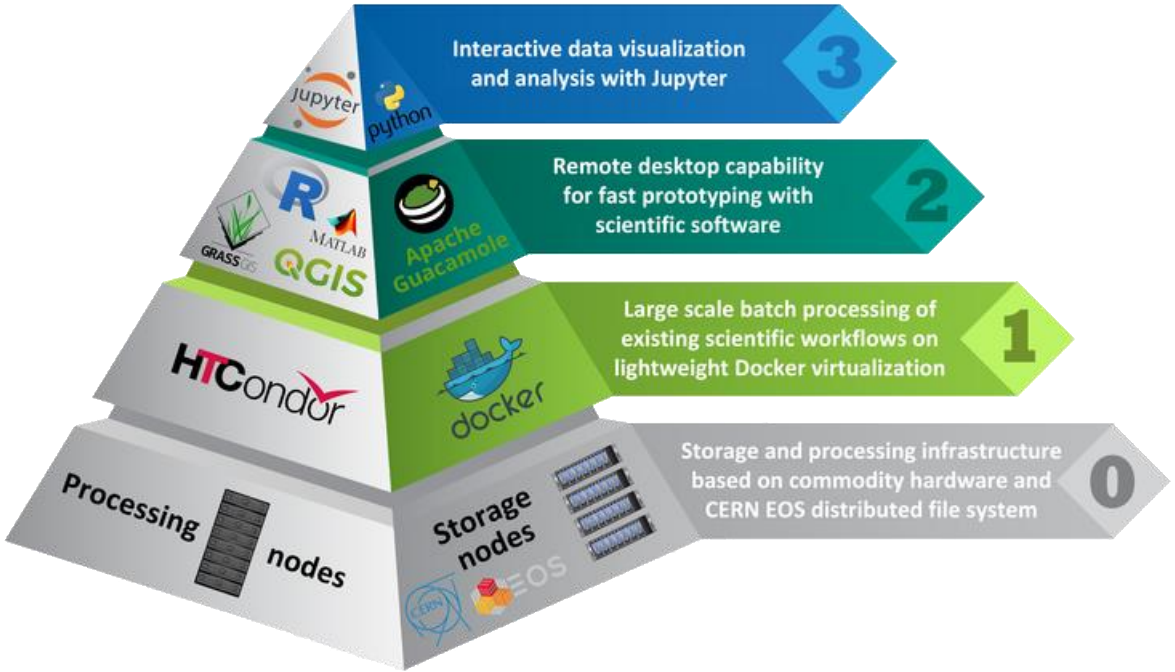
JRC: Earth Observation



- More/better satellites => more data
- Turnaround time
  - Importance of satellite data for everyday life
    - Pollution
    - Flood
    - Climate changes

- Slides from C. Macmillan (JRC):
  - opening session at
  - "Big data in Space",Oct 2017, Toulouse

JRC: Earth Observation Data and Processing Platform (JEODPP)



Similar tools as SWAN

Application specific

Commonalities with WLCG sites (Docker, HTCondor)

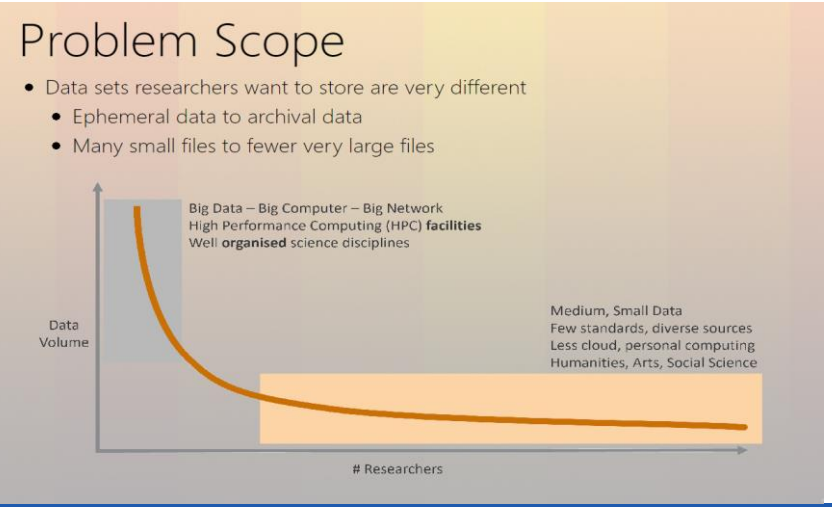
Commonalities with WLCG sites (EOS, OpenStack)

P. Hasenohr and A. BurgerJRC presentation at CS3 2018 (Krakow)

P. Soille et al., FGCS, 2017, DOI: 10.1016/j.future.2017.11.007

ARRNet: large content-deliver-networks

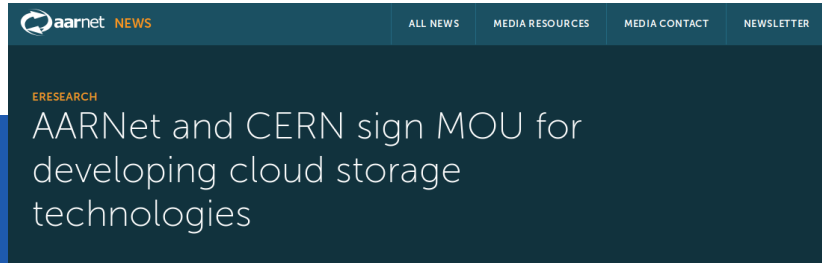
- **CloudStor** is a securely sync, share and storage service using the high-speed AARNet network.



cloudstor

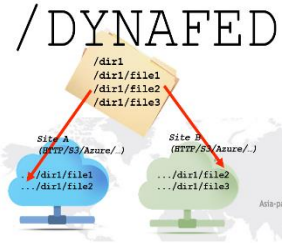
- Cloudstor uses EOS, the scalable back-end storage developed at CERN.

- Similarities with some projects in HEP, Cloudstor (AARNET) is:
  - Distributed from the start
  - Multi-science from the start
  - Multi-platform from the start



eXtreme Data Cloud Data Lake Concept

- The Horizon2020 eXtreme DataCloud – XDC project aims at developing scalable technologies for federating storage resources and managing data in highly distributed computing environments, as required by the most demanding, data intensive research experiments in Europe and worldwide.

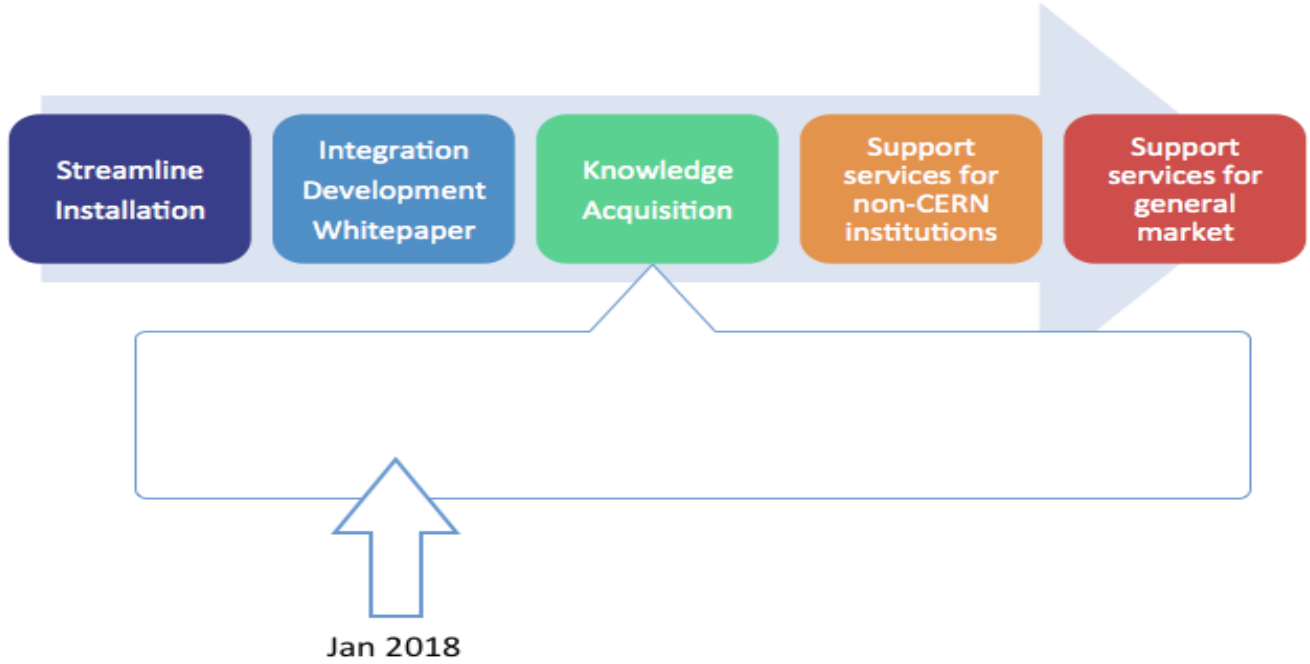


COMTRADE: Productionisation of CERN EOS Open Storage

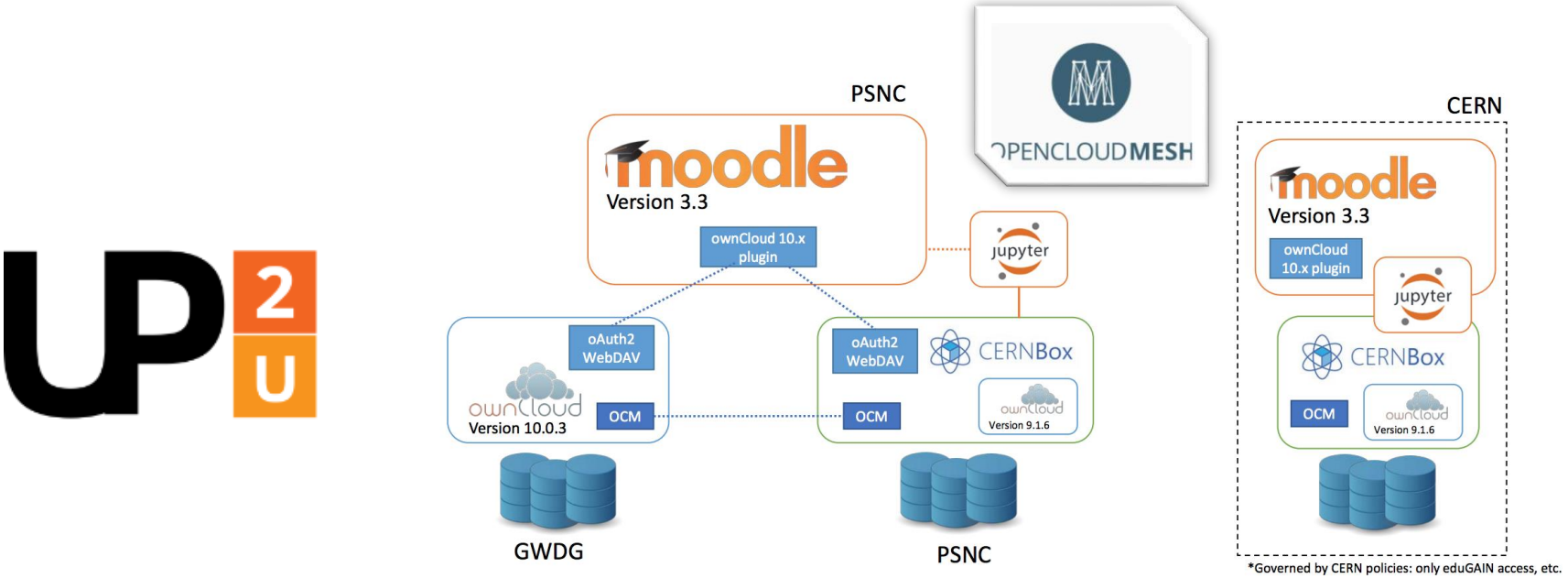
COMTRADE DIGITAL SERVICES

CERN openlab

EOS



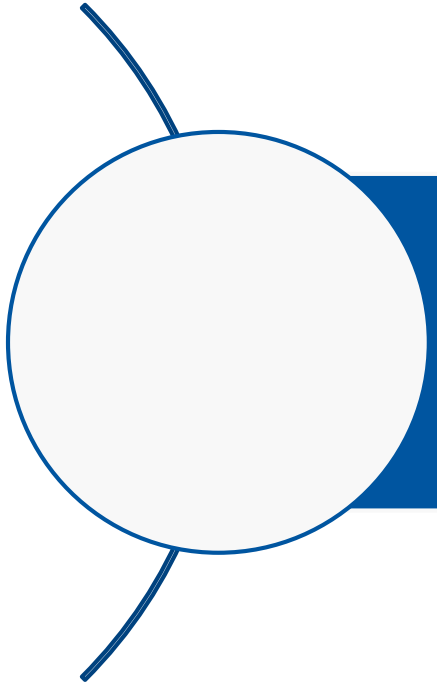
Up2You: bridge the gap between secondary schools and higher education



And the journey has just started: CS3 Conference

- Started as a workshop to learn
- We believe the drive of our community is an important factor of progress:
  - Which is needed (HL LHC)
  - Which can achieved by taking on board interesting technologies developed outside (ownCloud, Jupyter, Docker)
- Right participant base
  - HEP and non-HEP
  - WLCG sites, HPC sites, University sites
  - Academics, start-ups, established companies
- Need to join forces with non-HEP initiative (as JRC and AARNET)
- University (UP2U) important as outreach but also as source of input (expectation of usage of our tools)

And the journey has just started: CS3 Conference



Overview and Summary

Conclusions

- HEP-developed software (CERNBox, EOS, SWAN) can boost the use cases of other sciences (JRC, AARNet) and constitute the backend for a new generation of services.

- The challenge is to adopt cloud storage solutions into the main data analysis workflow and be able to cope with the data explosion we are facing.

- Focus on human efficiency rather than only on machine performance, changes in the way people perform their analysis.

See you in Rome!