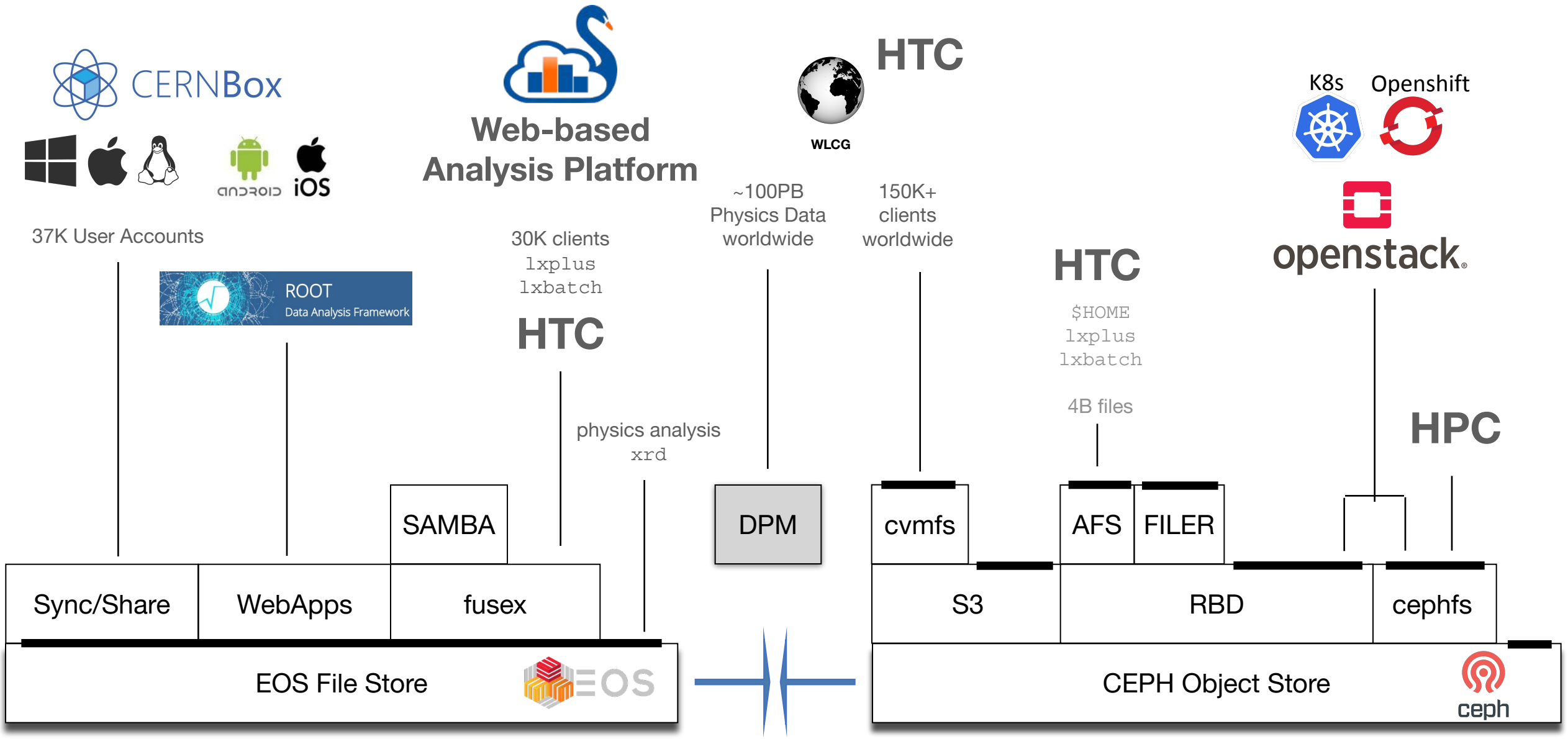
Sync and Share Access to HPC Resources at CERN

Dan van der Ster, Theofilos Mouratidis
CERN IT Storage Group

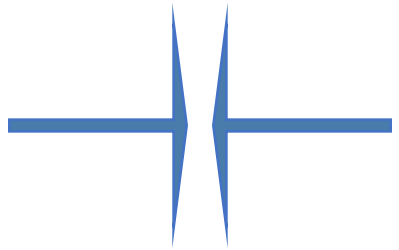
CS3 2022, 26 Jan 2022



15 production clusters, 18 PB general storage + 400PB physics

9 production clusters, 30 PB general storage

Grand unification of storage: theory and practice



Benefits of integrated service environment

…added value for users (obvious example: public clouds)

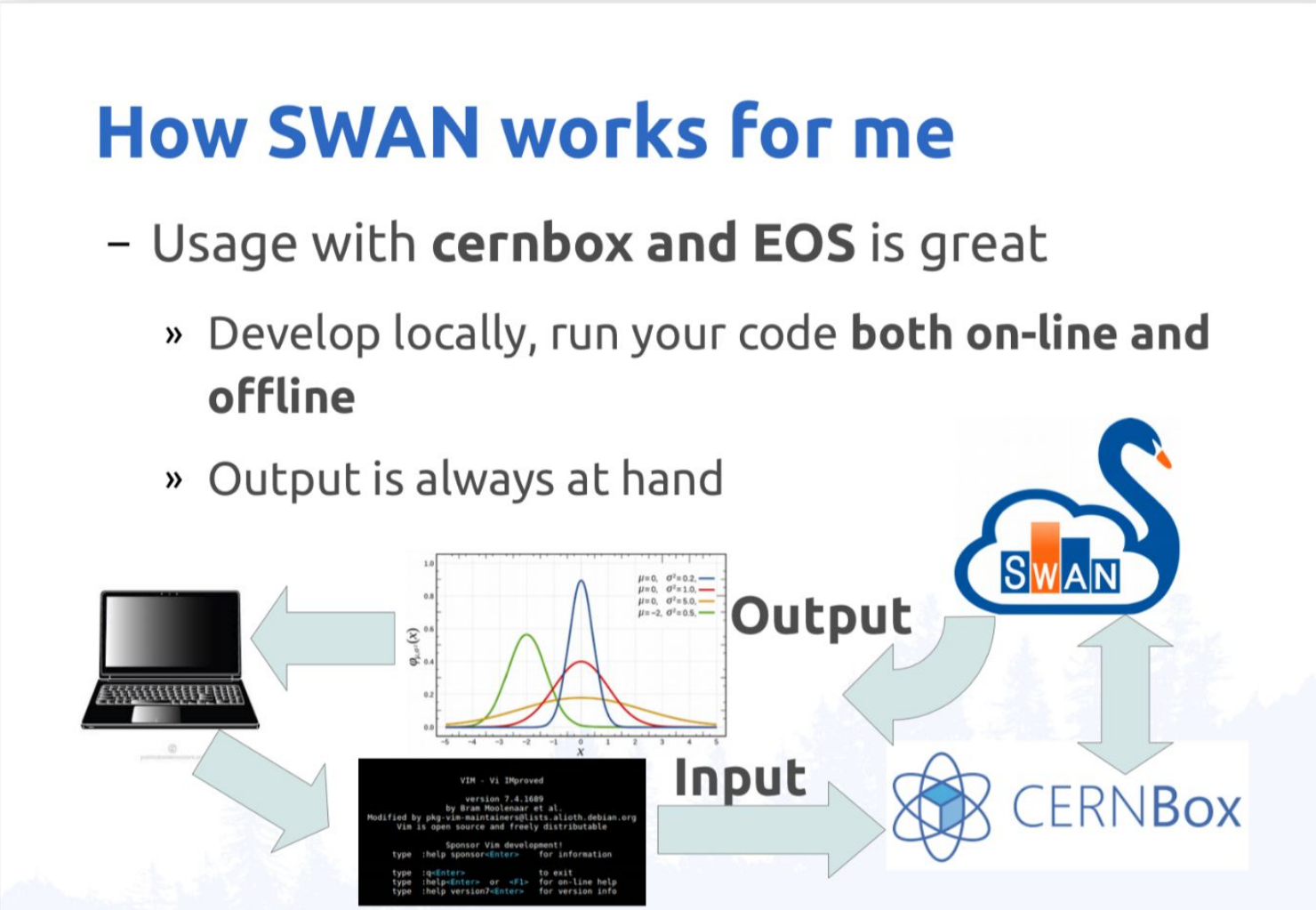
…combined storage tech to provide the most optimised service

…cost/benefit optimisation of available hardware resources at CERN DC

Some practical examples (prototyping and experimenting)

1. **Open Source Storage backend synergy: physics (EOS) and HPC (CephFS)**
   - **Previous talk "Converging Storage Layers with Virtual CephFS Drives for EOS/CERNBox"**
2. **Integration of HPC storage with the web-based analysis service environment**
   - **SWAN + Jupyter Notebooks**
3. **Easier access to user data in HPC storage (CephFS) via Sync/Share**

Online & offline access to user data



Credits: Nicolo Jacazio, INFN, Bologna, SWAN User Workshop, https://indico.cern.ch/event/834069/timetable/#20191011.detailed

High Performance Computing (HPC)

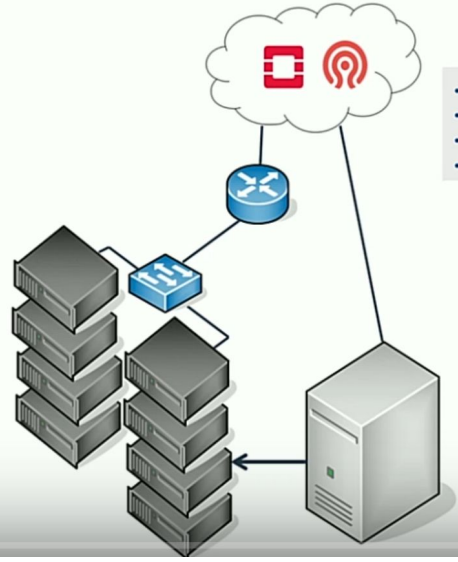
- **Applications and use cases that do not fit the standard batch HTC model. Typically parallel MPI applications**
  - Theory Lattice QCD studies (TH)
  - Accelerator physics, beam simulation, plasma simulations... (BE, TE)
  - Computation Fluid Dynamics, CFD (EN, EP, HSE)
    - Also structural analysis, field calculations (EN,PH,TE), currently mainly on Windows fat boxes (run by IT-CDA)
- **Job duration often very long, (e.g. seve CFD and QCD)**
  - Stability of OS and environment critical
  - MPI application performance require fast inte latency between nodes in a cluster. Some ap require fast access to shared storage

Open HPC Infrastructure

HPC Worker nodes

- Intel Xeon E5 2630 v3
- 128GB Memory 1600Mhz
- RAID 10 SATA HDDs
- Low-latency Chelsio T520-LL-CR
- Communication iWARP/RDMA CM

Operated since mid-2016:
~300 client nodes
~1PB CephFS



CephFS on FileStore

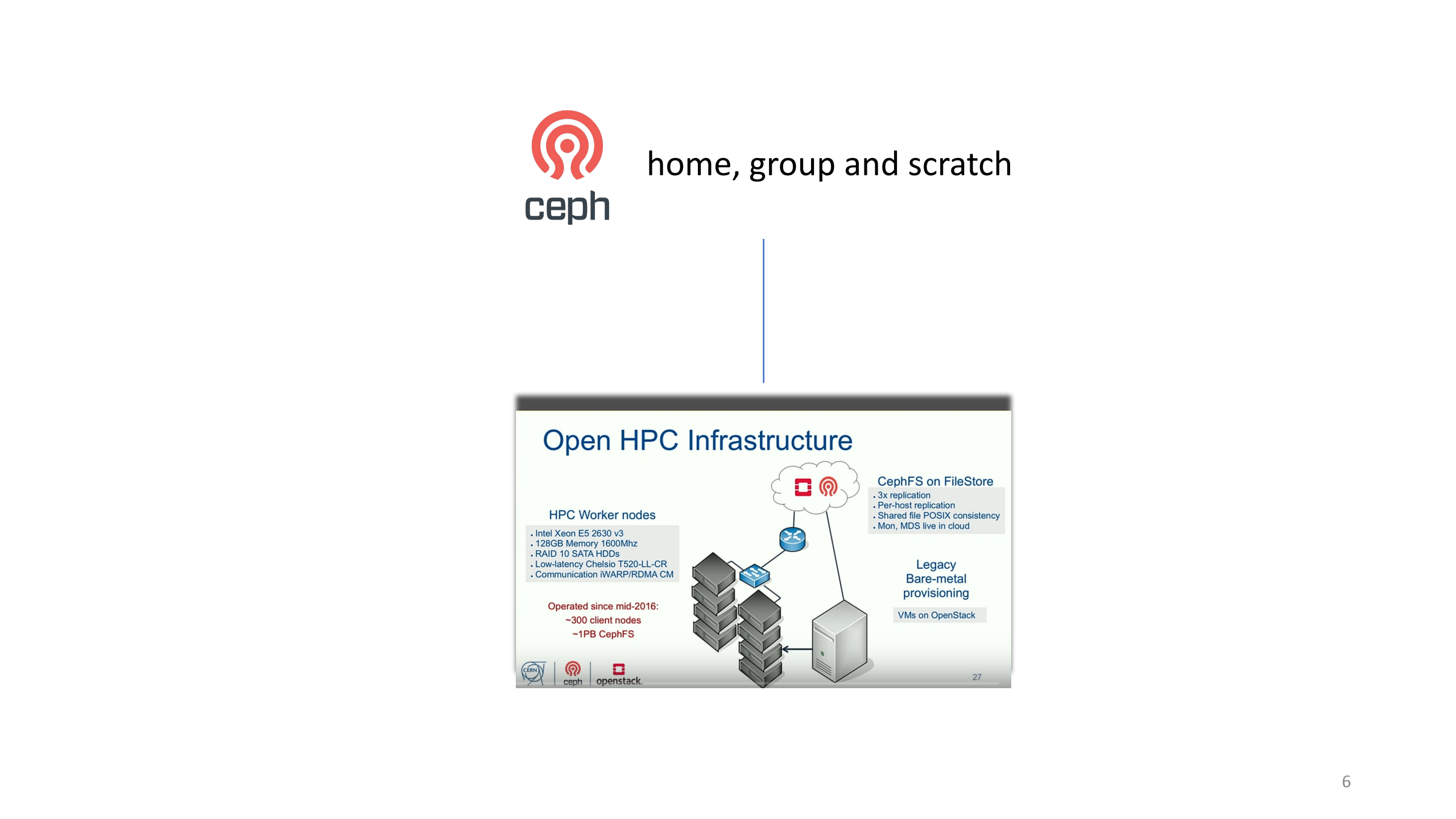
- 3x replication
- Per-host replication
- Shared file POSIX consistency
- Mon, MDS live in cloud

Legacy Bare-metal provisioning

VMs on OpenStack

ceph openstack.

Credits: CERN IT-CM group



home, group and scratch

Easier access to user data in HPC storage (CEPHFS) via Sync/Share

Easily Accessible User Data

CERNBox ownCloud

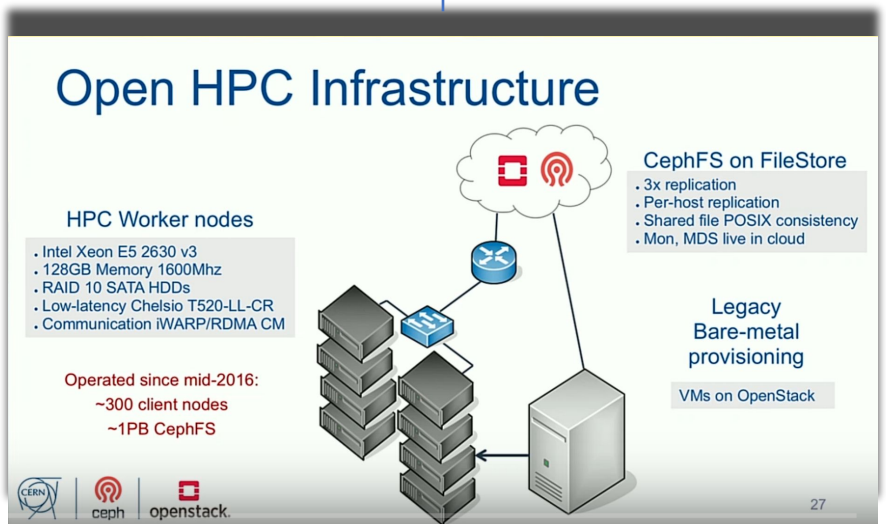
android iOS

Reva

reva.link

ceph

home, group, scratch



Easier access to user data in HPC storage (CEPHFS) via Sync/Share

Easily Accessible User Data

Reva

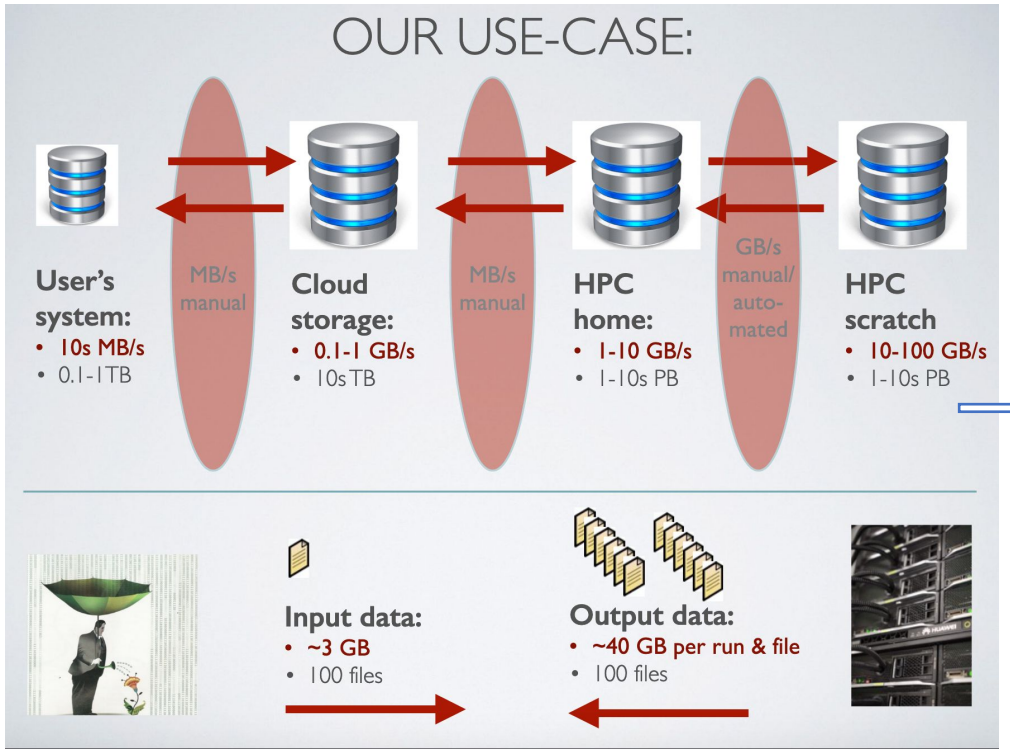
reva.link

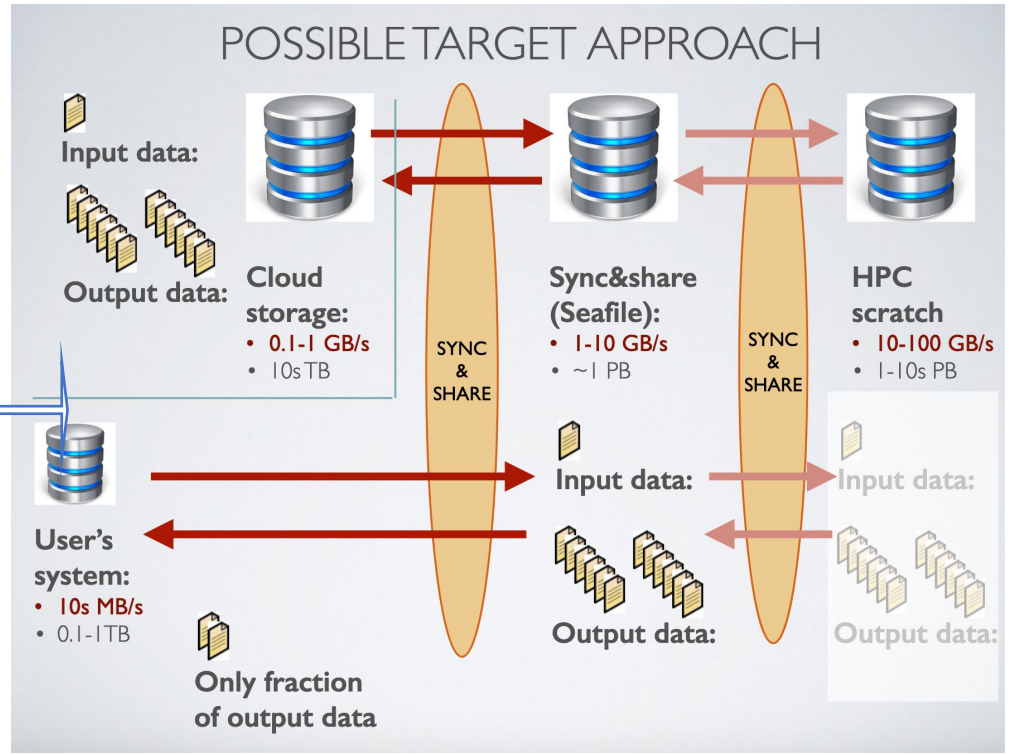
home, group, scratch

SYNC & SHARE TO REPLACE HPC HOMES?

HPC Department PSNC

M. Brzezniak et al., CS3 2018
https://indico.cern.ch/event/663264/contributions/2818149





Integration of HPC storage with the web-based analysis service environment

Easily Accessible User Data

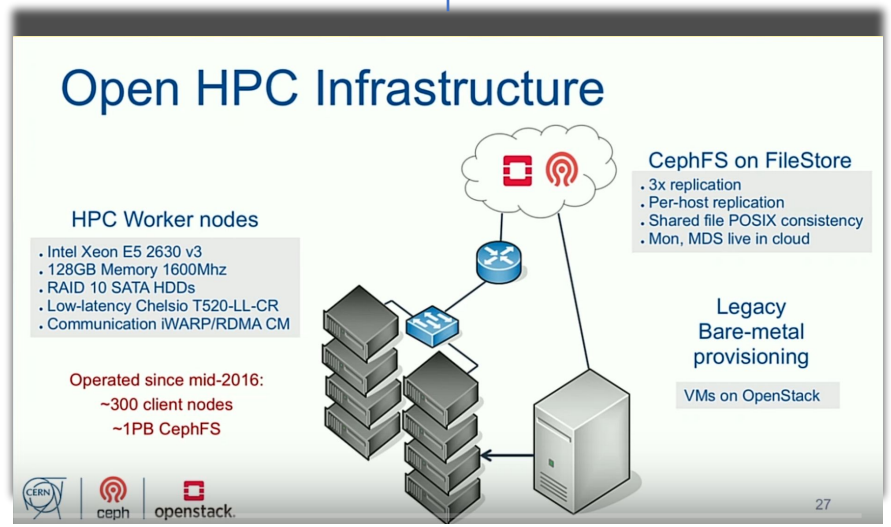
CERNBox ownCloud

android iOS

ceph

home, group, scratch

CEPHFS mounts
Krb5 auth



EOSXD mounts
Krb5 auth

EOS

Long-term storage

CephFS+Reva: Introduction

- Reva is an interoperability platform that CERNBox is based on
  - It provides Sync&Share functionality
  - It integrates different services together, such as storage
- The HPC users want to synchronise their home directories
- Is CephFS a good fit for a CERNBox backend?
- Can CERNBox support CephFS directly without EOS/NFS layers?
- https://indico.cern.ch/event/970232/contributions/4158391/

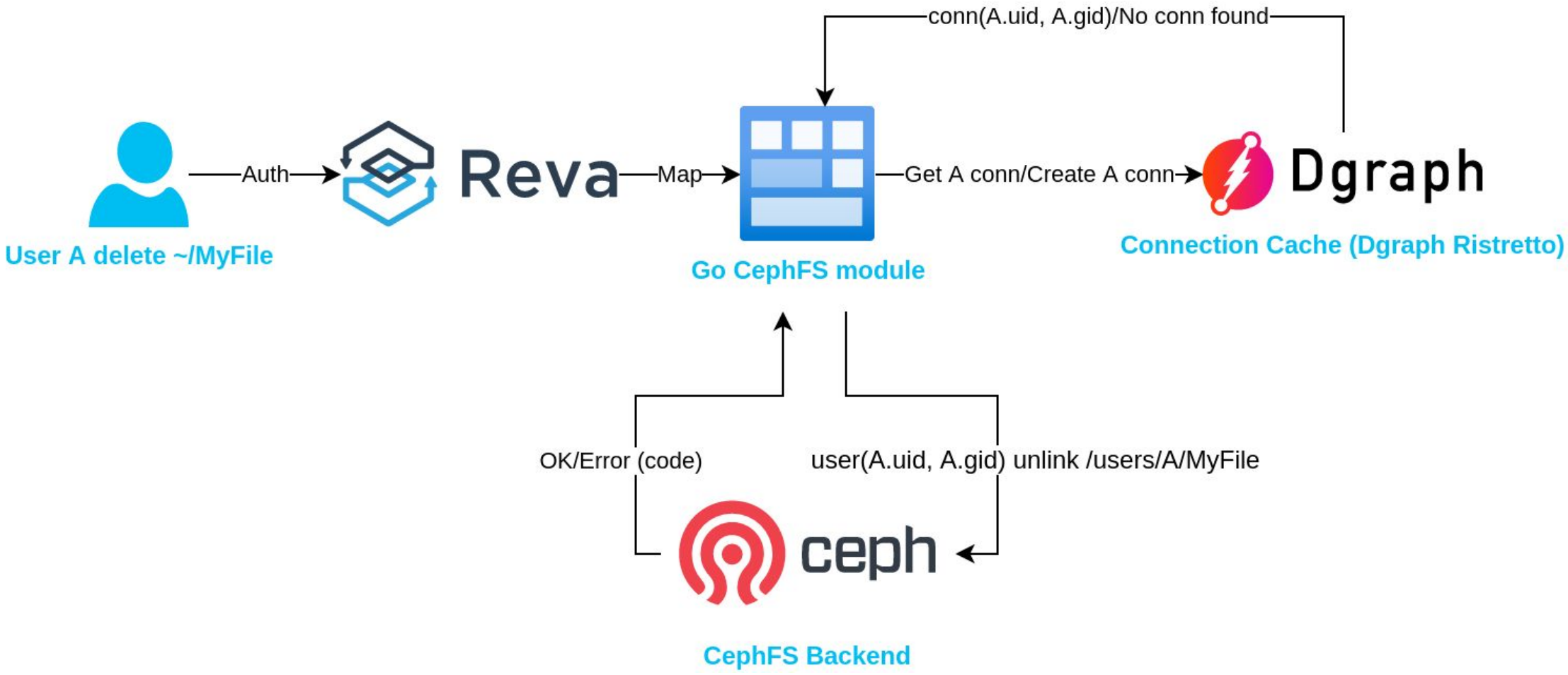
CephFS+Reva: Features (1)

- A full featured Reva requires the following features:
  - Recursive mtimes → optimised file sync
  - File IDs → Shares, File versions
  - Uploads/Downloads
  - File Versions
  - Recycle Bin
  - Home directories
  - Shares
  - Permissions

CephFS+Reva: Features (2)

- Current POC supports the below features:
  - Recursive mtime propagation with CephFS built-in recursive accounting
  - File IDs are partly supported
    - Without direct access to backend
    - We need to track changes in real time
  - Home directories with REVA Api
  - Implemented TUS downloads
  - POSIX ACL permissions by mapping to CERNBox ones
    - Cache user connections, one connection per user
    - Need parameterised permissions in CephFS API
- File Versions and Recycle Bin not yet supported
  - Better use CephFS snapshots

CephFS+Reva: Features (3)



CephFS+Reva: Comments

- Reva is a new platform and dev docs are missing
  - It made the development a bit challenging
- The API interface seems trivial but:
  - Each method does more than it describes
  - Features support is hidden in the plain methods
- To start, I copied the current localFS module
  - I studied the purpose of each method
  - I slowly adapted the code to use libcephfs
- Each storage should provide its own features, not everything
- Reva is built with Owncloud and EOS in mind
  - No support for POSIX ACLs/Permissions by default
  - No snapshot support

CephFS+Reva: Discussion

- The POC is currently running in CERNBox QA
- Basic functionalities for HPC use-case work well
- CephFS seems like a good fit for a CERNBox backend
- Some changes on both sides will make the module better
  - CERNBox to support snapshots, POSIX ACLs
  - Ceph to support file ids
- The implementation is new
  - Needs some benchmarks
  - Didn’t pass the test suite yet

Summary

- HPC plays an increasingly important role in High Energy Physics
- The CephFS REVA plugin is merged
  - CERN is in the process of enabling it so HPC users can access the files via CERNBox
- Sounds interesting?
  - jakub.moscicki@cern.ch
  - daniel.vanderster@cern.ch
  - theofilos.mouratidis@cern.ch


CERN