CERN

CERN IT-Storage

EOS at CERN

Andreas Peters

Luca Mascetti

CERN - IT Storage

andreas.joachim.peters@cern.ch

luca.mascetti@cern.ch

www.cern.ch

CERN IT-Storage

Our goal: provide storage solution and tools for data management and data analysis to experiments and users and operate them.

COMPASS ALICE ATLAS CMS LHCb TOTEM AMS-02 ilc SHIVE NA62

Vision for CERN Storage

High capacity storage for experiments at EB scale

EOS Disk pools (on-demand reliability, on-demand performance)

currently ~340 PB deployed

CERN Tape Archive (high reliability, low cost, data preservation)

currently ~300 PB stored

A shared storage service across platforms

To cover the majority of the requirements for personal storage

Shared among all clients and services (access across different platform)

Fuse mounts, CIFS exports

Desktop Sync Client and Web/HTTP/DAV Access

EOS

ABOUT EOS

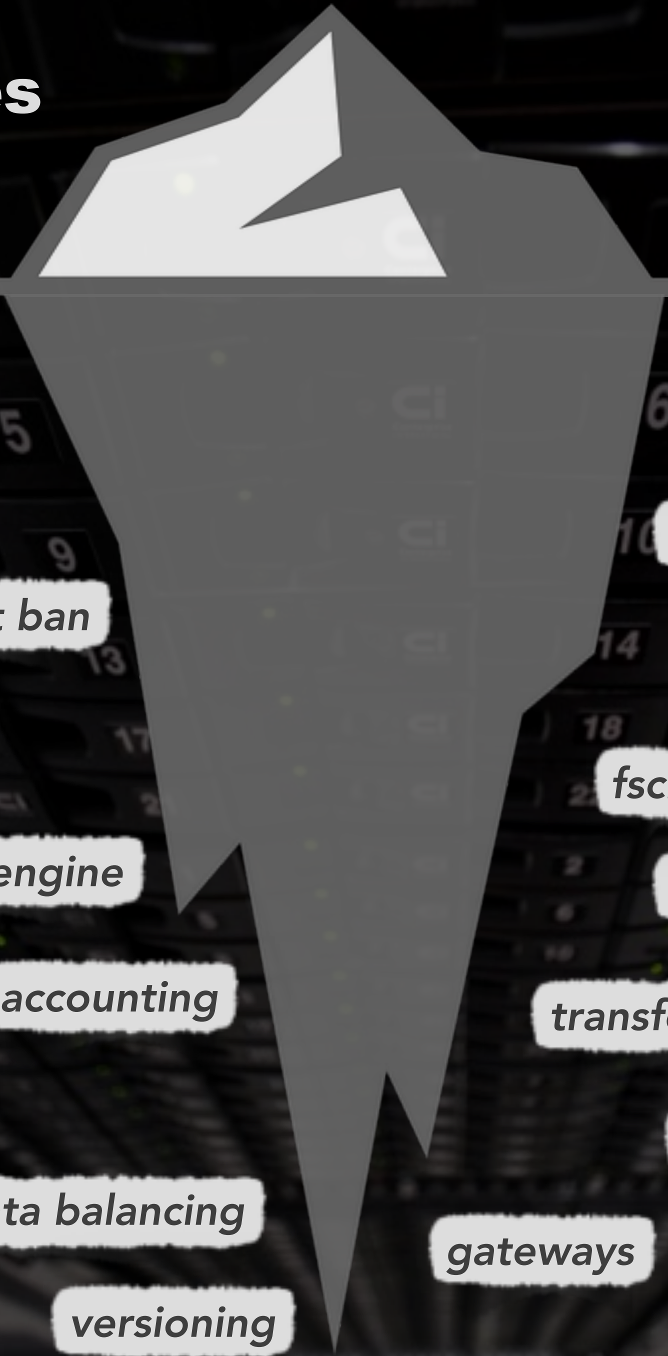
Elastic, Adaptable and Scalable

EOS is a simple and scalable open source software solution for central data recording, user analysis and data processing.

EOS supports thousands of clients with random remote I/O patters with multiprotocol support and tunable QoS.
HTTP, WebDAV, CIFS, FUSE, XRoot, gsiFTP

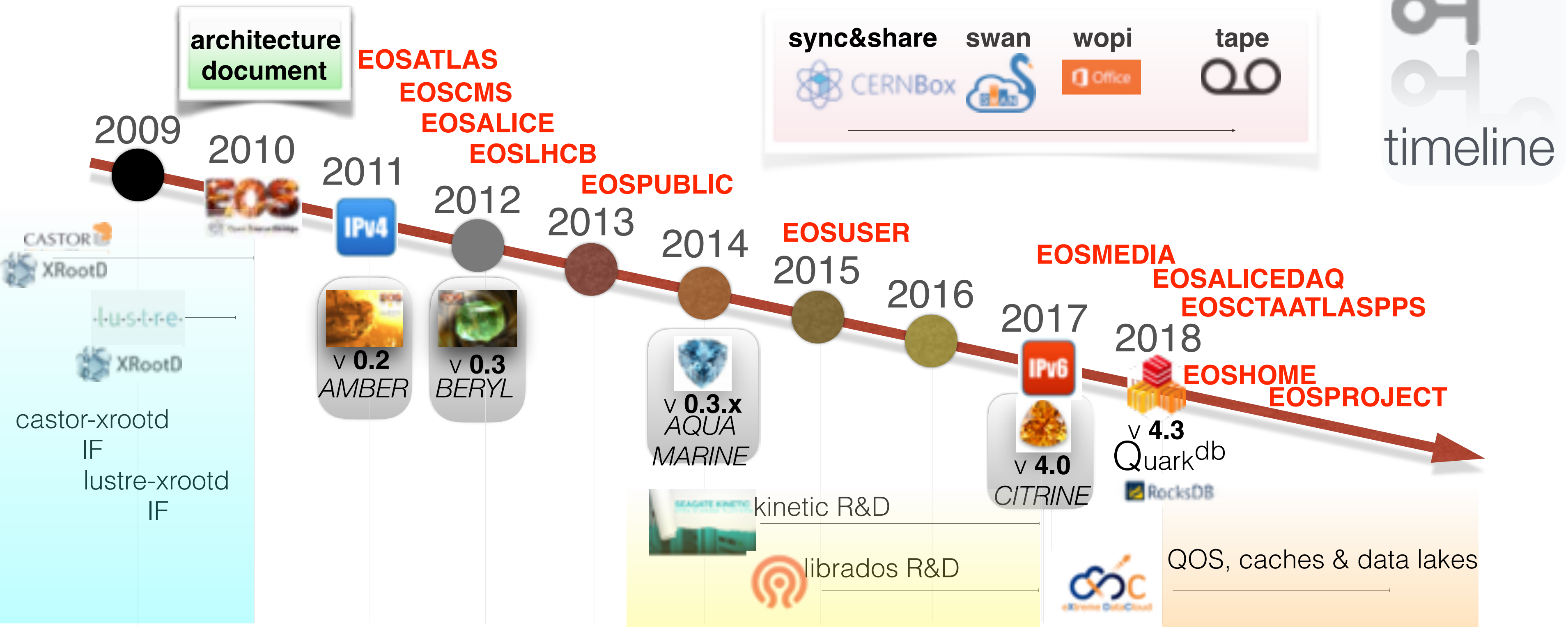
EOS offers a variety of authentication methods and user/project quotas.
KRB5, X509, Shared Secret and unix

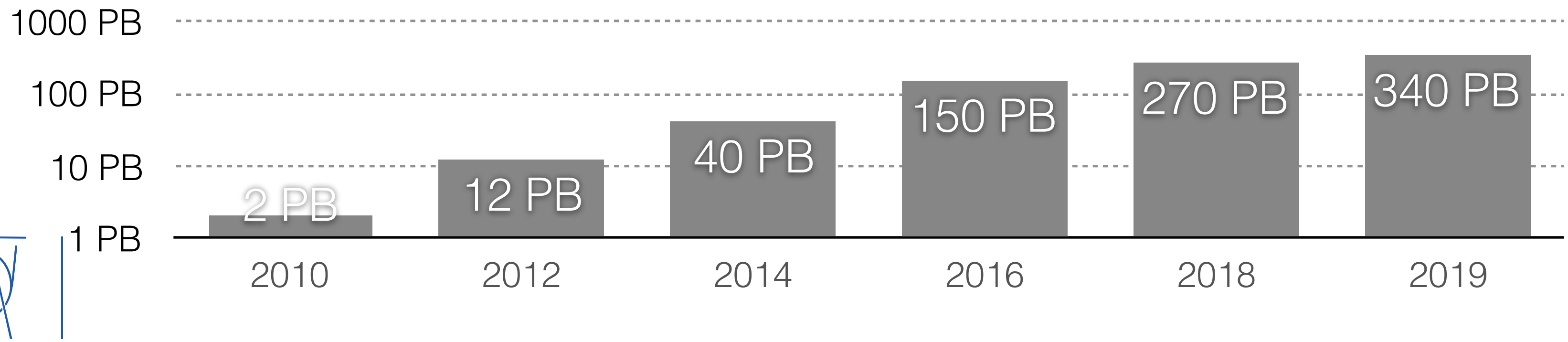
EOS Features



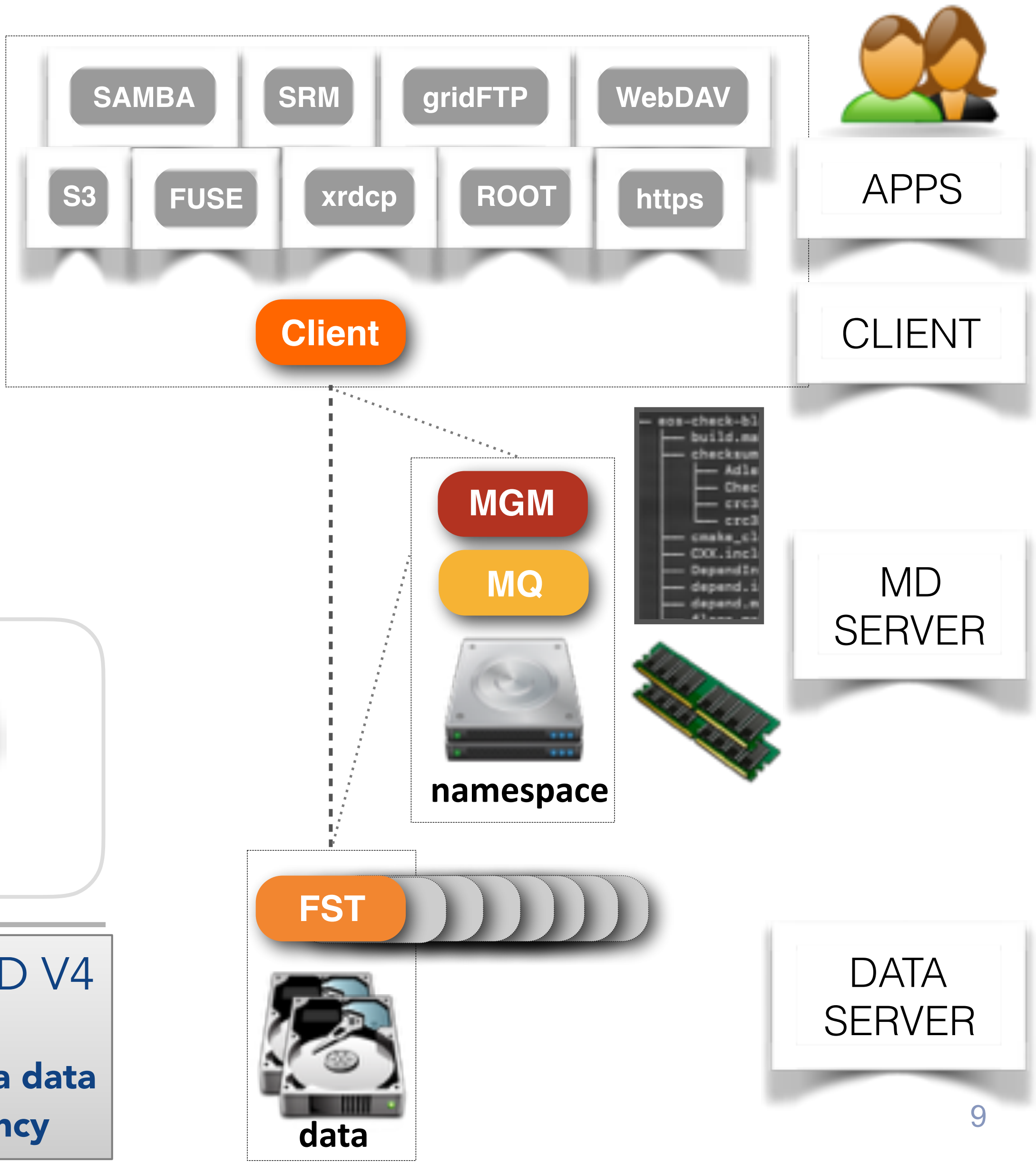
- rate limiting
- recycle bin
- user/host ban
- https/S3 extensions
- workflow engine
- accounting
- fuse optimisation
- intergroup data balancing
- versioning
- geo scheduling policies
- id mappings
- io monitoring
- fsck
- stat monitoring
- sticky ownership
- transfers engine
- rich ACLs
- sharing
- erasure encoding
- gateways
- and many more...

EOS Project History





EOS Architecture



EOS Production Releases

Aquamarine
V 0.3.X

Citrine
V 4.X

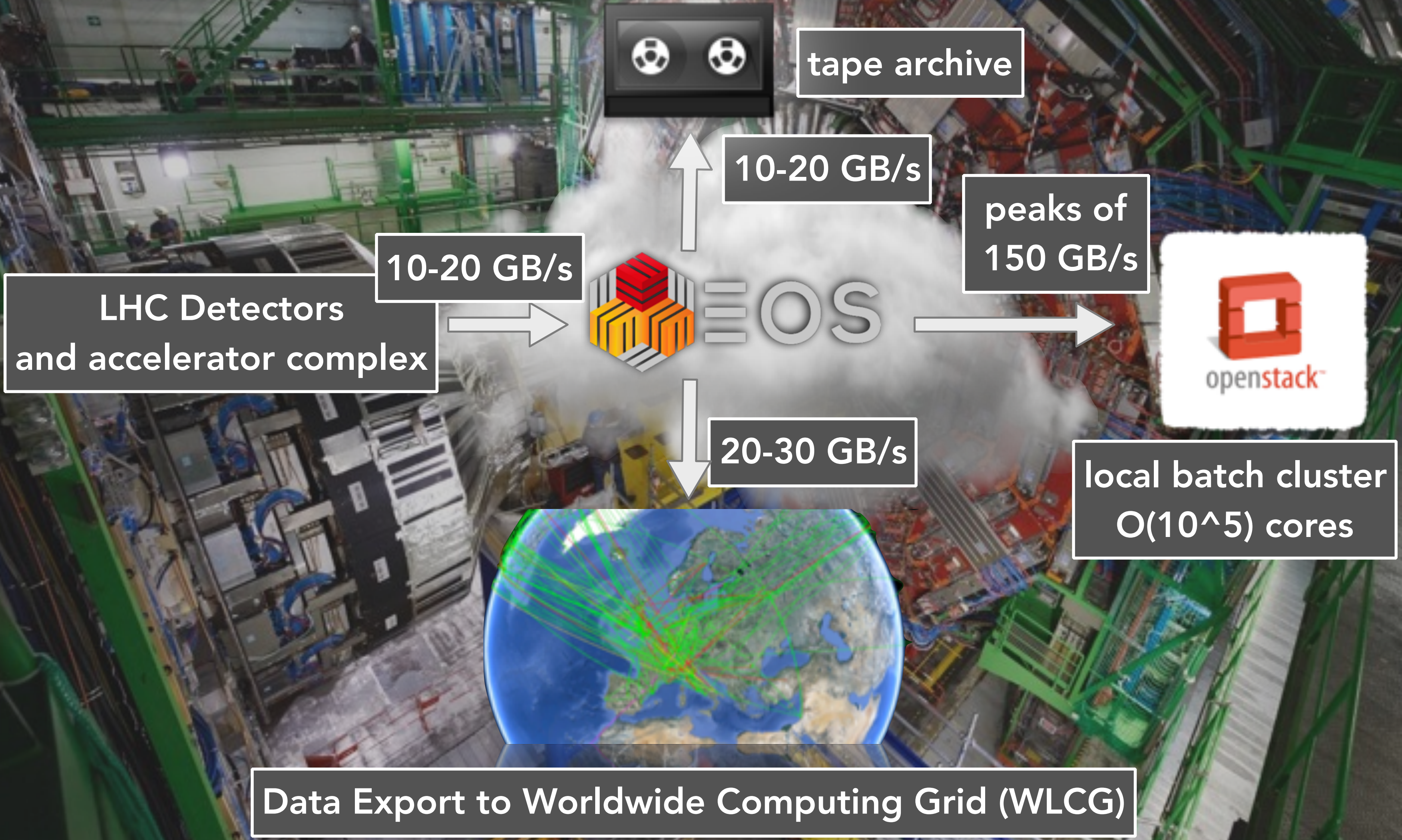
XRootD V3

IPV4
namespace in-memory
data on attached disks

XRootD V4

IPV6
plugins for meta data
& data persistency

How is it used? CERN’s mainstream usecase



tape archive

10-20 GB/s

LHC Detectors
and accelerator complex

10-20 GB/s

peaks of
150 GB/s

local batch cluster
O(10^5) cores

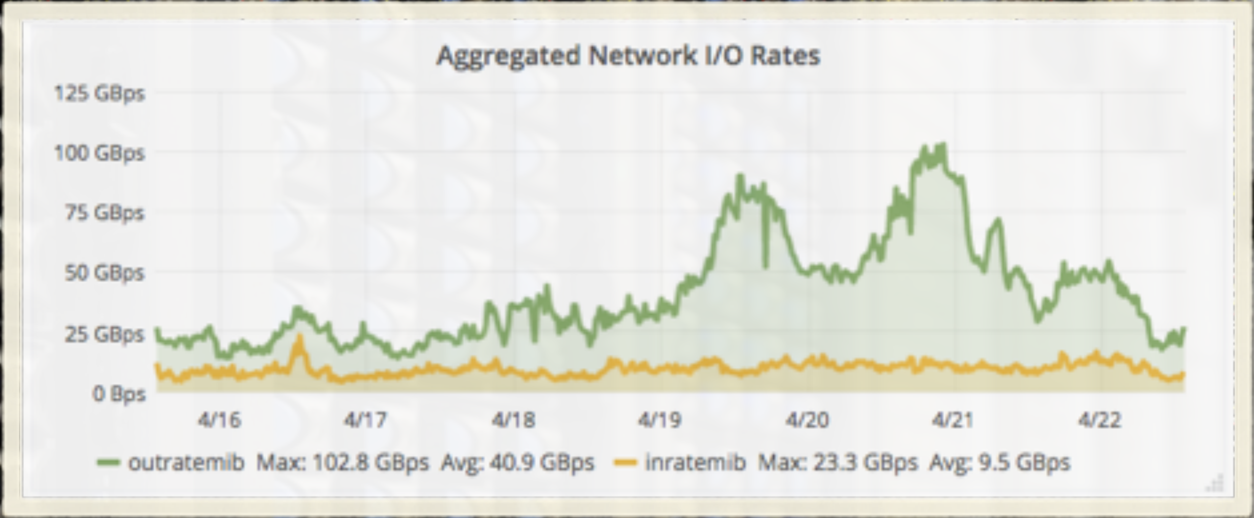
20-30 GB/s

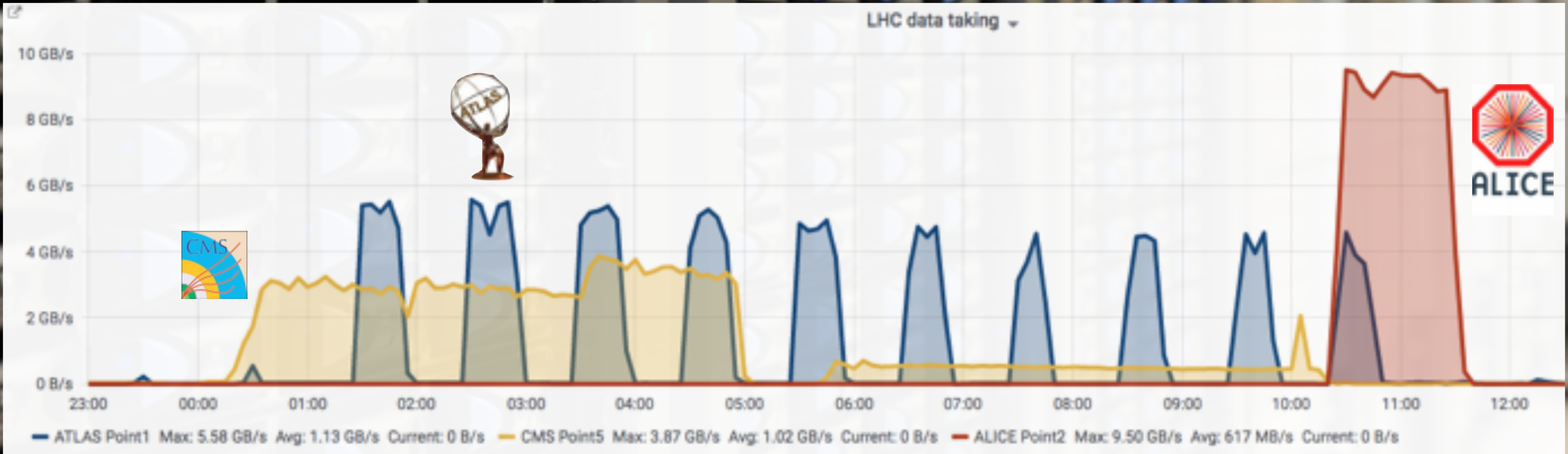
Data Export to Worldwide Computing Grid (WLCG)

EOS Production instances @ CERN

Total Space
340 PB

Files Stored
4.92 Bil

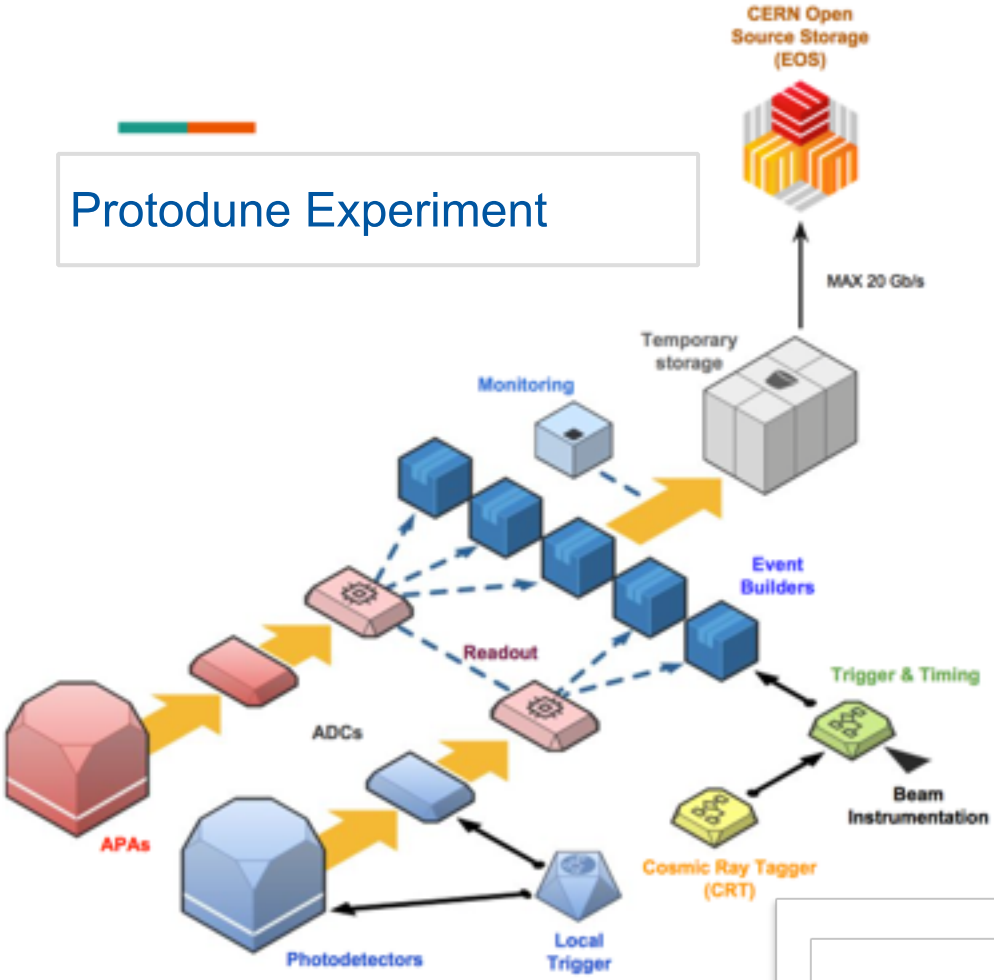


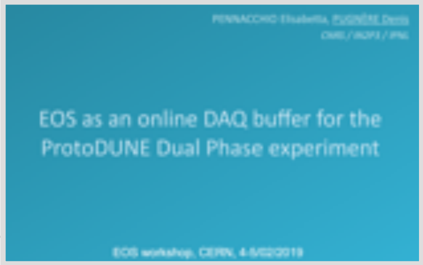


EOS as Online Storage

EOS Workshop 2019

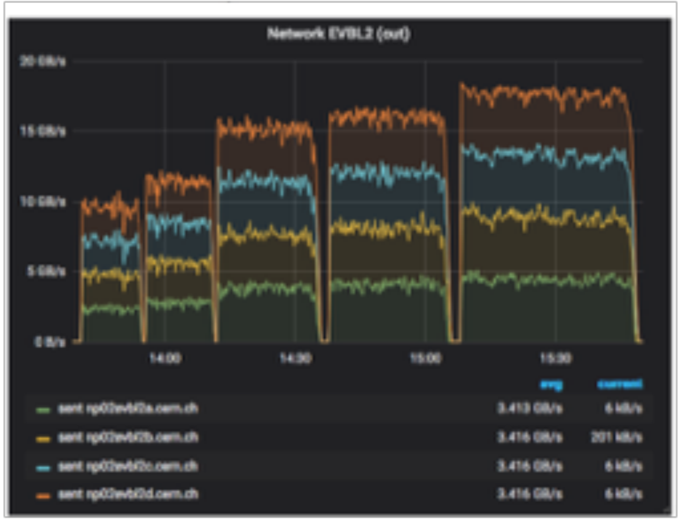
Protodune Experiment




PENNACCHIO Elisabetta, PUGNÈRE Denis
CNRS / IN2P3 / IPNL

EOS as an online DAQ buffer for the ProtoDUNE Dual Phase experiment

EOS workshop, CERN, 4-5/02/2019


Network EVBL2 (out)

	avg	current
🟩 sent np02evb02a.cern.ch	3.413 GB/s	6 kB/s
🟨 sent np02evb02b.cern.ch	3.416 GB/s	201 kB/s
🟦 sent np02evb02c.cern.ch	3.416 GB/s	6 kB/s
🟧 sent np02evb02d.cern.ch	3.416 GB/s	6 kB/s

Conclusion

- EOS is performing well, huge recent improvements
- EOS is a SDS (software defined storage) which can take advantages of old hardware with decent performance (over 16GB/s continuous writing)

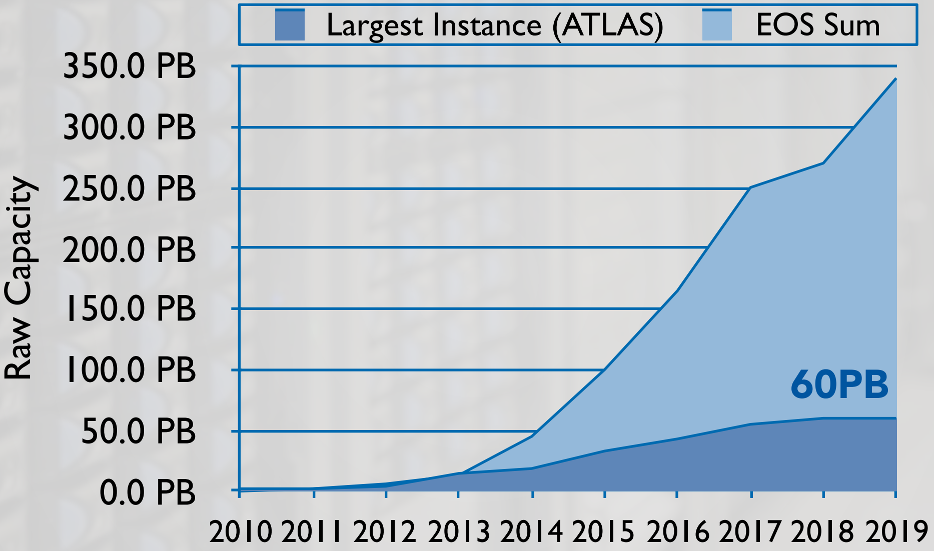
EOS Production instances @ CERN

EOS instances:

- 5 for the LHC experiments
- 9 CERNBox (EOSUSER + 5 EOSHOME + 3 EOSPROJECT)
- EOSMEDIA (photo/audio/video)
- EOSPUBLIC (Open Data and non-LHC experiments)
- EOSBACKUP (backup for CERNBox)
- 5 for various tests

~1500 storage nodes

~60k disks

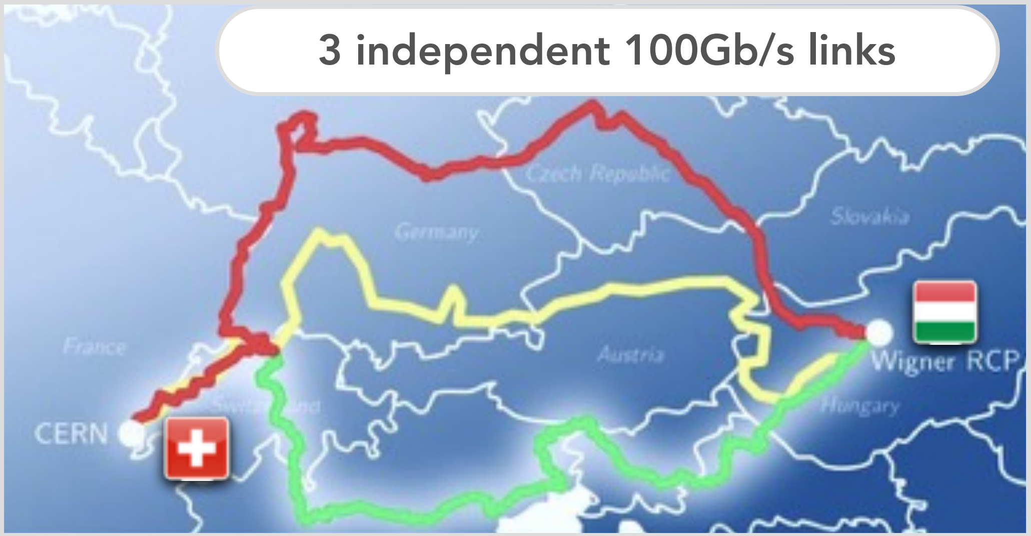


Hardware evolution

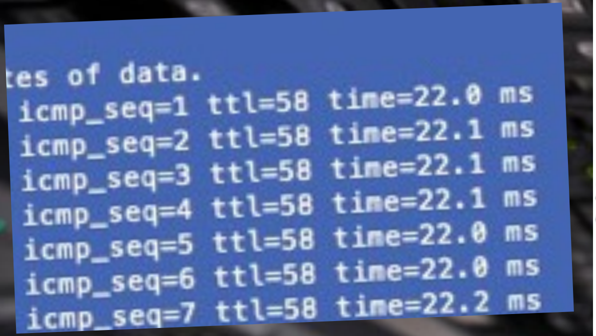
- Profiting from ***economy of scale***
  - minimise price per GB
- System Unit:
  - 8 physical cores (16 virtual) 64-128GB RAM
  - disk-tray of 24x 4-6-10-12TB HDDs

- Running different generations
  - 2 trays per system unit
  - 4 trays per system unit
  - 8 trays per system unit

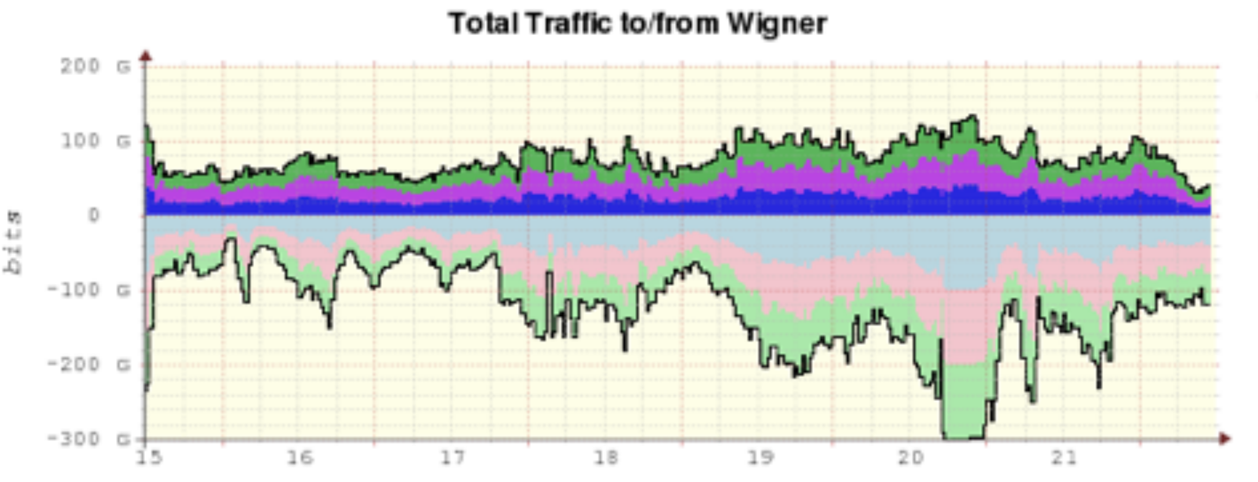
Wigner Computer Centre



3 independent 100Gb/s links



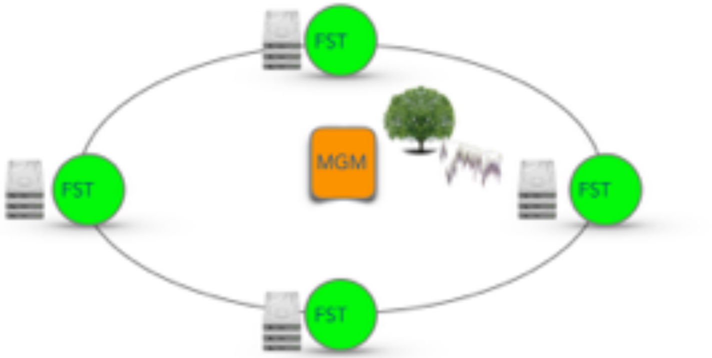
```
tes of data.
 icmp_seq=1 ttl=58 time=22.0 ms
 icmp_seq=2 ttl=58 time=22.1 ms
 icmp_seq=3 ttl=58 time=22.1 ms
 icmp_seq=4 ttl=58 time=22.1 ms
 icmp_seq=5 ttl=58 time=22.0 ms
 icmp_seq=6 ttl=58 time=22.0 ms
 icmp_seq=7 ttl=58 time=22.2 ms
```



Total Traffic to/from Wigner

Distributed Storage Setup

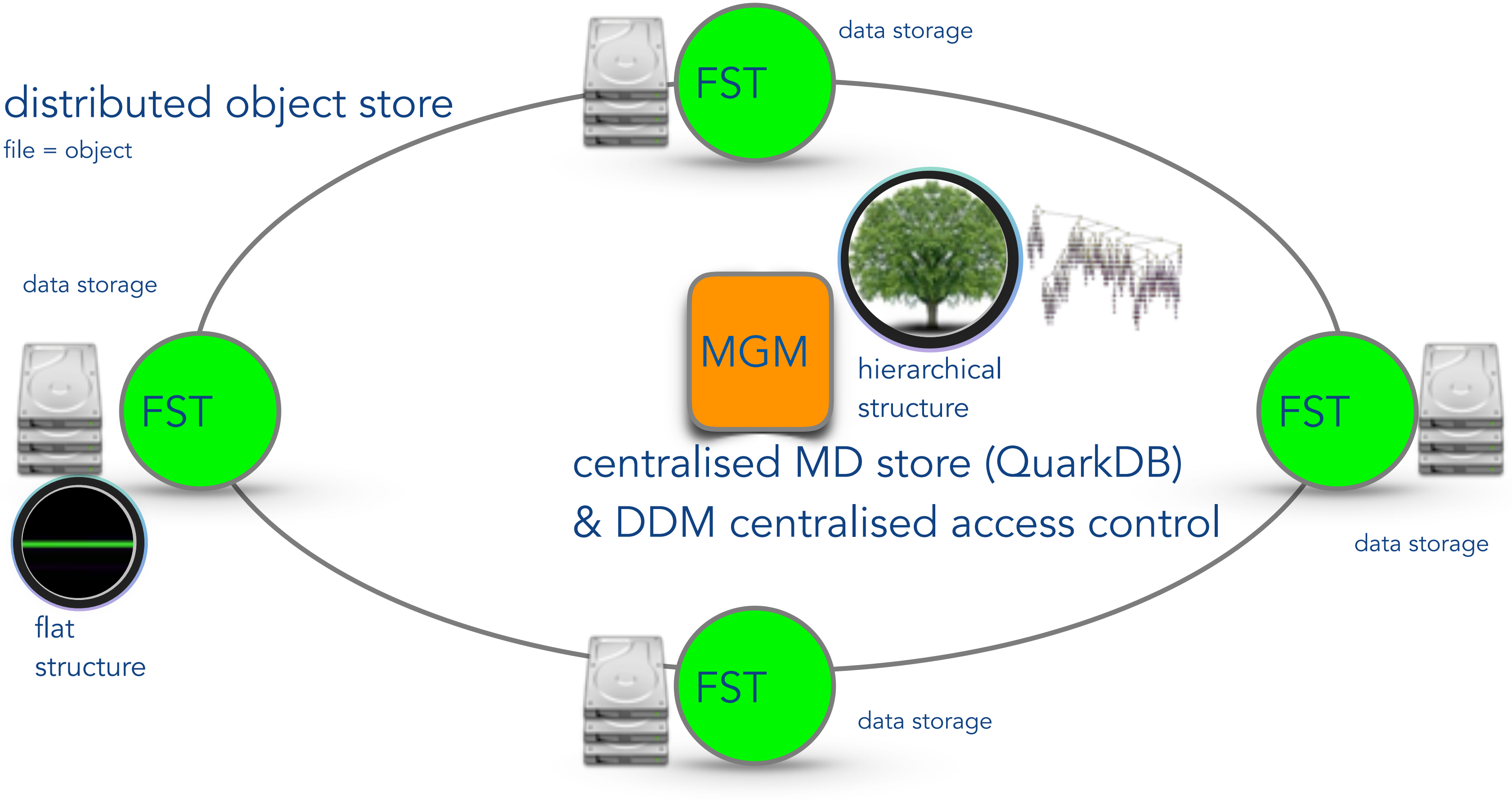
EOS lake



multi-site storage native EOS JBOD
centralised namespace in KV store for meta data
distributed object store for data

- simple design, deployment & configuration
  - central **HA** namespace **MGM** (persistency in QuarkDB)
  - distributed file storage **FST** (single daemon on top of disk(s) + key to connect)
    - deployment as service or container
    - each service is geographically tagged e.g. EOS::CERN::513
  - **placement policies** defined globally or per directory/subtree taking into account geographical tags of FSTs and location of clients

EOS - Distributed Architecture



distributed object store

file = object

data storage

FST

data storage

FST

flat

structure

MGM

hierarchical

structure

centralised MD store (QuarkDB)

& DDM centralised access control

data storage

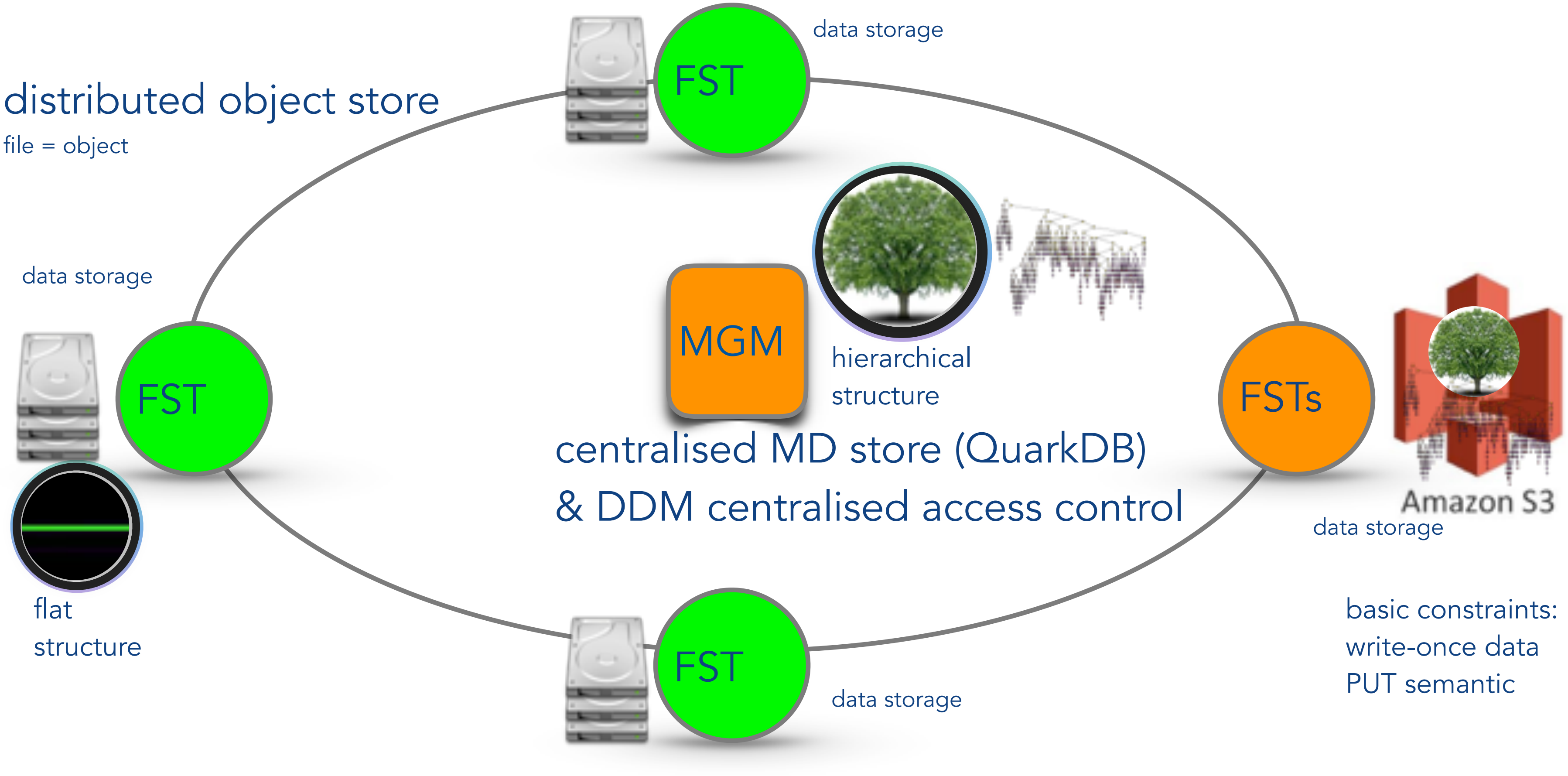
FST

data storage

FST

data storage

EOS - External Storage Systems



distributed object store

file = object

data storage

FST

data storage

FST

flat

structure

MGM

hierarchical

structure

centralised MD store (QuarkDB)

& DDM centralised access control

FSTs

Amazon S3

data storage

FST

data storage

basic constraints:

write-once data

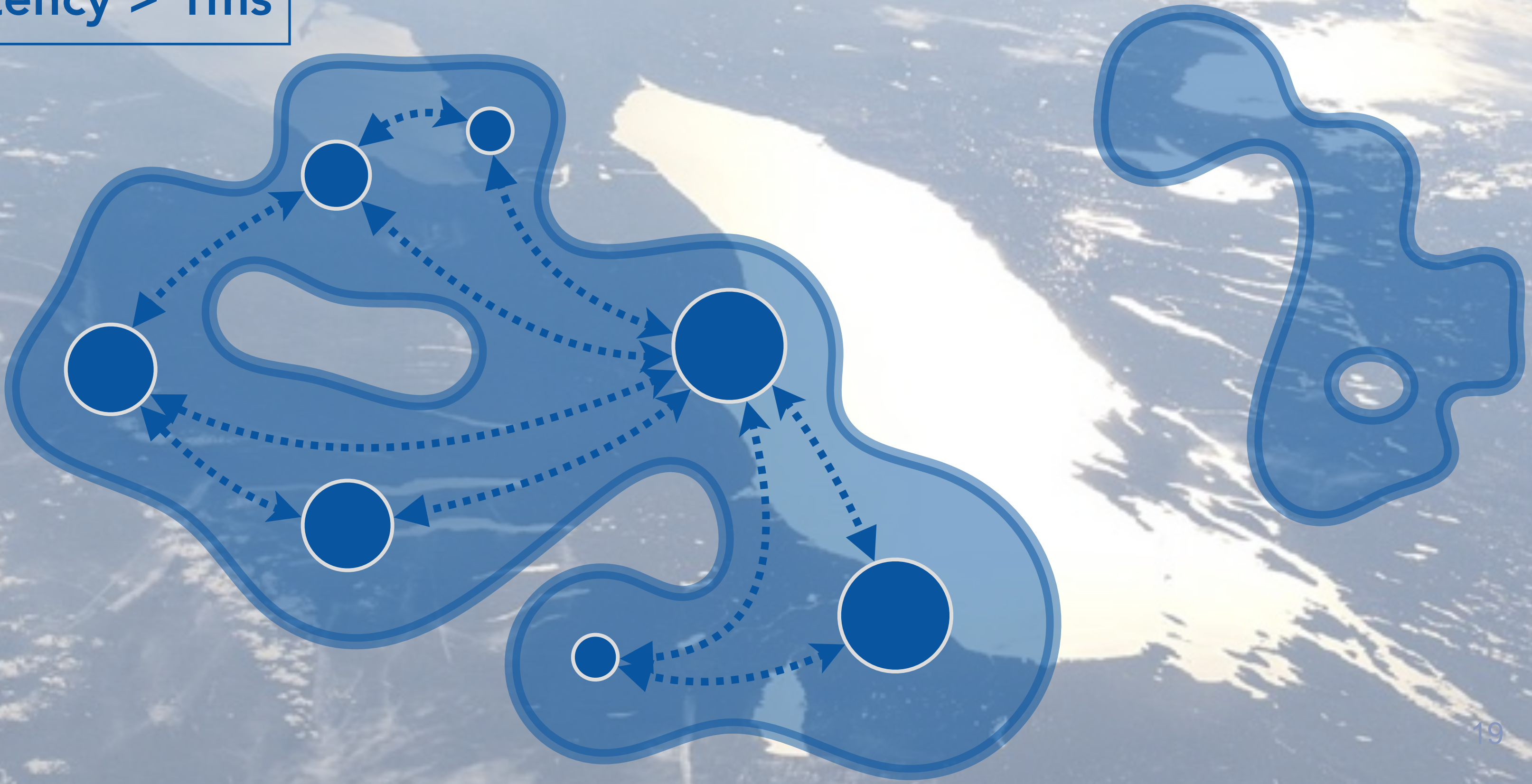
PUT semantic

R&D EOS “Data Lake”

exploring the geographical distributed EOS capabilities

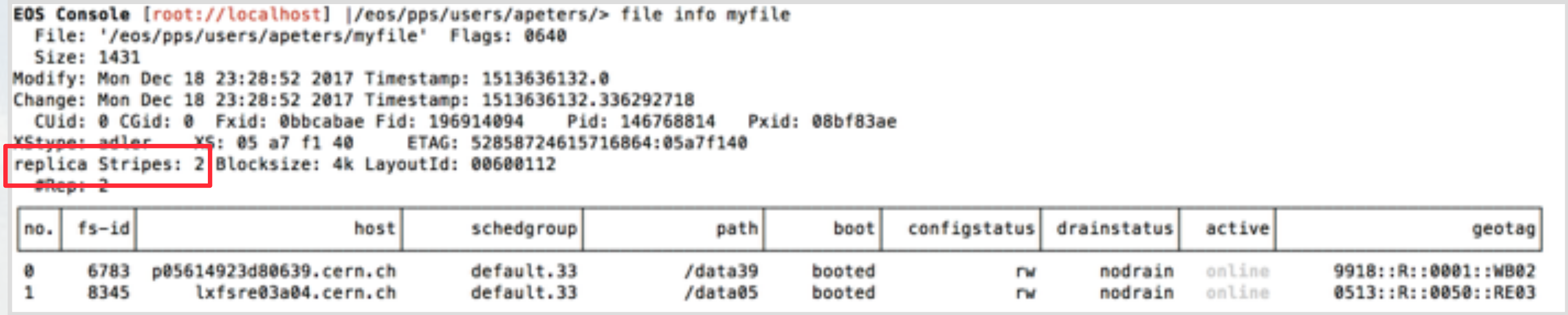
n sites
k replicas
k<<n
latency > 1ms

High-Luminosity LHC
CERN-SKA partenship



EOS - File Layouts

EOS files described by static layout (type + parameters e.g replica:2)



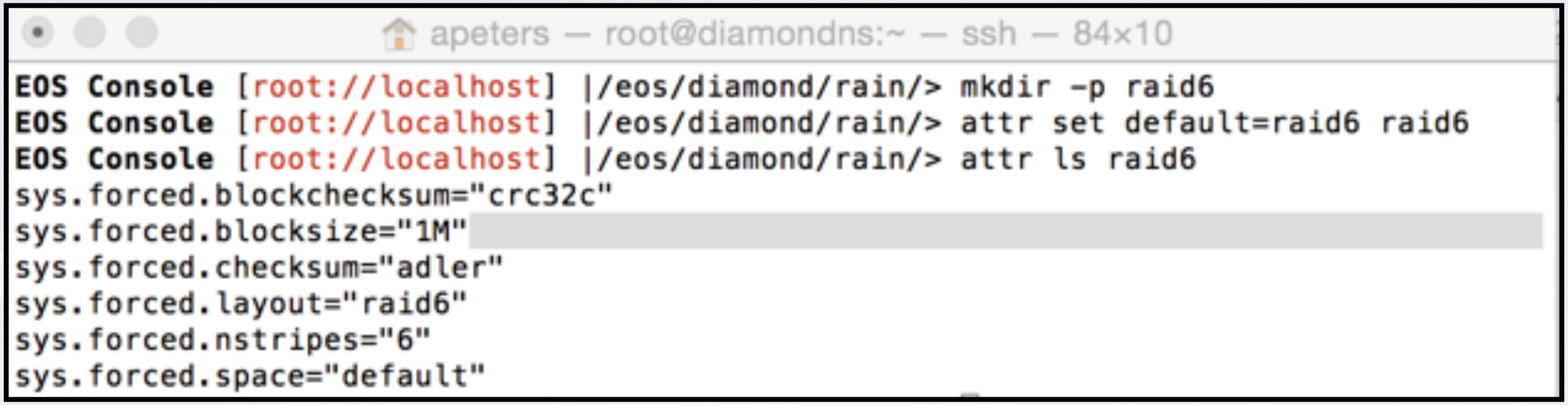
```
EOS Console [root://localhost] |/eos/pps/users/apeters/> file info myfile
  File: '/eos/pps/users/apeters/myfile'  Flags: 0640
  Size: 1431
Modify: Mon Dec 18 23:28:52 2017 Timestamp: 1513636132.0
Change: Mon Dec 18 23:28:52 2017 Timestamp: 1513636132.336292718
  CUid: 0 CGid: 0  Fxid: 0bbcabae Fid: 196914094    Pid: 146768814   Pxid: 08bf83ae
XStype: adler    XS: 05 a7 f1 40      ETAG: 52858724615716864:05a7f140
replica Stripes: 2 Blocksize: 4k LayoutId: 00600112
  #Rep: 2
┌───┬──────┬────────────────────────┬────────────────┬────────────────┬───────┬────────────┬───────────┬──────┬────────────────────┐
│no.│ fs-id│                    host│      schedgroup│            path│   boot│configstatus│drainstatus│active│              geotag│
└───┴──────┴────────────────────────┴────────────────┴────────────────┴───────┴────────────┴───────────┴──────┴────────────────────┘
 0      6783  p05614923d80639.cern.ch       default.33          /data39  booted           rw     nodrain  online  9918::R::0001::WB02
 1      8345     lxfsre03a04.cern.ch       default.33          /data05  booted           rw     nodrain  online  0513::R::0050::RE03
```

Available layout types:

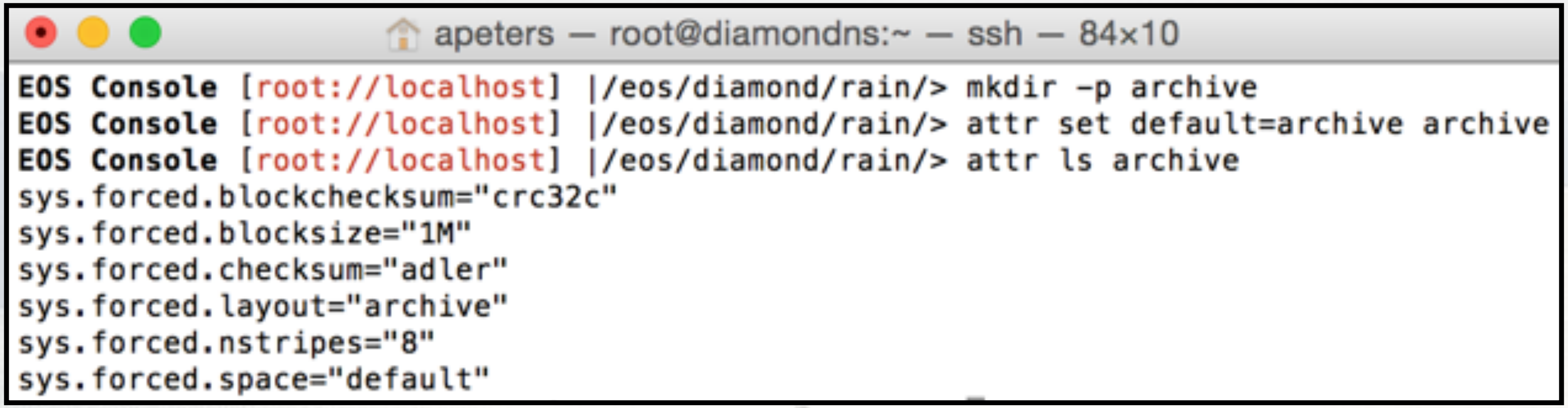
- replication layouts
  - **plain** (single copy)
  - **replica** (n copies)
- erasure coding layouts (RAIN)
  - **Raid5** (n stripes, 1 parity)
  - **Raid6** (n stripes, 2 parity)
  - **Archive** (n stripes, 3 parity)
  - **Qraid** (n stripes, 4 parity)

- http://eos-docs.web.cern.ch/eos-docs/using/rain.html
- http://eos-docs.web.cern.ch/eos-docs/using/policies.html

EOS - File Layouts


apeters — root@diamondns:~ — ssh — 84×10

```
EOS Console [root://localhost] |/eos/diamond/rain/> mkdir -p raid6
EOS Console [root://localhost] |/eos/diamond/rain/> attr set default=raid6 raid6
EOS Console [root://localhost] |/eos/diamond/rain/> attr ls raid6
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="1M"
sys.forced.checksum="adler"
sys.forced.layout="raid6"
sys.forced.nstripes="6"
sys.forced.space="default"
```


apeters — root@diamondns:~ — ssh — 84×10

```
EOS Console [root://localhost] |/eos/diamond/rain/> mkdir -p archive
EOS Console [root://localhost] |/eos/diamond/rain/> attr set default=archive archive
EOS Console [root://localhost] |/eos/diamond/rain/> attr ls archive
sys.forced.blockchecksum="crc32c"
sys.forced.blocksize="1M"
sys.forced.checksum="adler"
sys.forced.layout="archive"
sys.forced.nstripes="8"
sys.forced.space="default"
```

http://eos-docs.web.cern.ch/eos-docs/using/policies.html

EOS - Layout Conversions

http://eos-docs.web.cern.ch/eos-docs/configuration/lru.html

automatic policies how or where files are stored

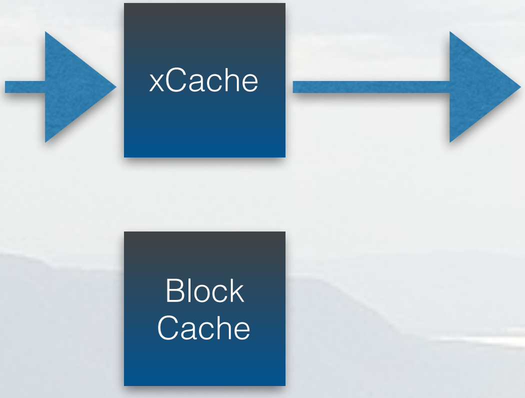
	on creation	**after 1 month**	**after 3 month**	**after 6 month**
	replica:3 + dyn. caching	RAIN: (4,2) no dyn. caching	replica: 1 + 1 tape copy	1 tape copy
on disk	300% + dyn.	150%	100%	0%
on tape	0%	0%	100%	100%

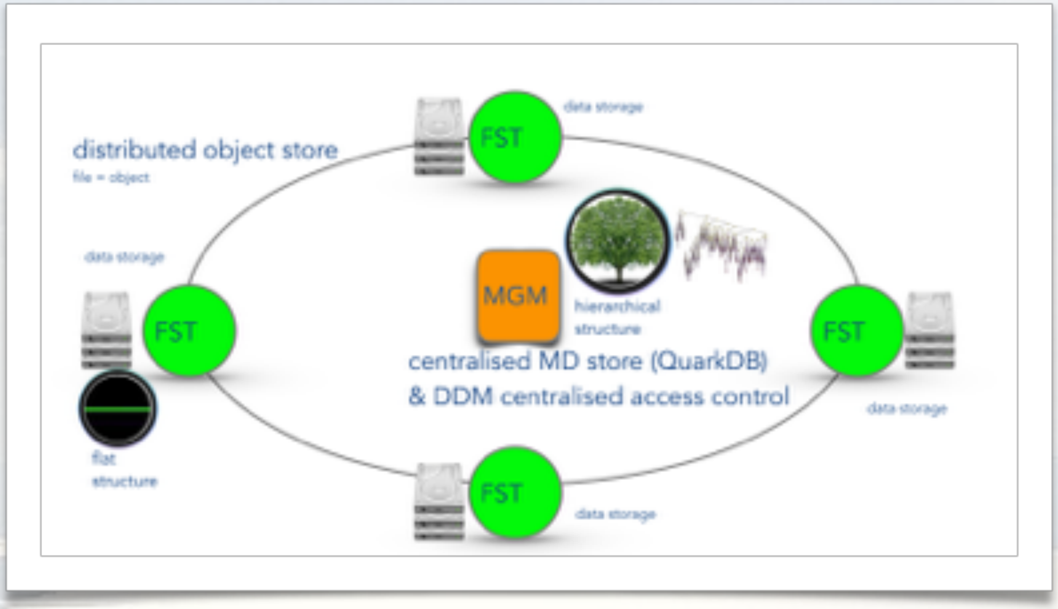
Simple extended attribute language to express time-based conversions is available.

Within the XDC project we are commissioning a conversion engine allowing to instantiate conversion jobs based on arbitrary namespace queries and a standardised QOS interface.

EOS - Adding unmanaged ro-caches

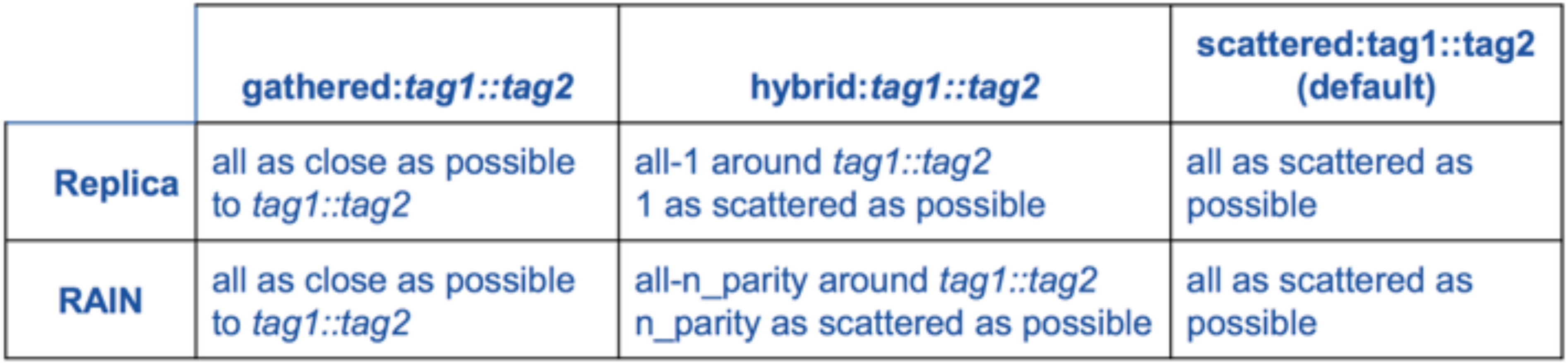
EOS default protocol is XRootD
- any combination with XRootD components & plug-ins is supported





Remote running xCache
(XRootd + cache plug-in)

EOS - Placement Policies



	gathered:*tag1::tag2*	**hybrid:*tag1::tag2***	**scattered:tag1::tag2 (default)**
Replica	all as close as possible to *tag1::tag2*	all-1 around *tag1::tag2* 1 as scattered as possible	all as scattered as possible
RAIN	all as close as possible to *tag1::tag2*	all-n_parity around *tag1::tag2* n_parity as scattered as possible	all as scattered as possible

Specify placement policies **in multiple contexts**

- Set placement policy in a directory
  ```
  eos attr set sys.forced.placementpolicy=gathered:site2 /eos/demo
  ```

- Specify placement policy in an explicit file conversion
  ```
  eos file convert /eos/demo/passwd replica:2 default scattered
  ```

- Set placement policy in an automatic conversion (LRU converter)
  ```
  eos attr set 'sys.conversion.*=00600112|scattered' /eos/demo
  ```

http://eos-docs.web.cern.ch/eos-docs/configuration/geotags.html
http://eos-docs.web.cern.ch/eos-docs/configuration/geoscheduling.html

EOS - Information

EOS Open Storage

CERN storage technology used at the Large Hadron Collider (LHC)

Web

https://eos.web.cern.ch/

Documentation

http://eos-docs.web.cern.ch/eos-docs/

Source Code

https://gitlab.cern.ch/dss/eos/

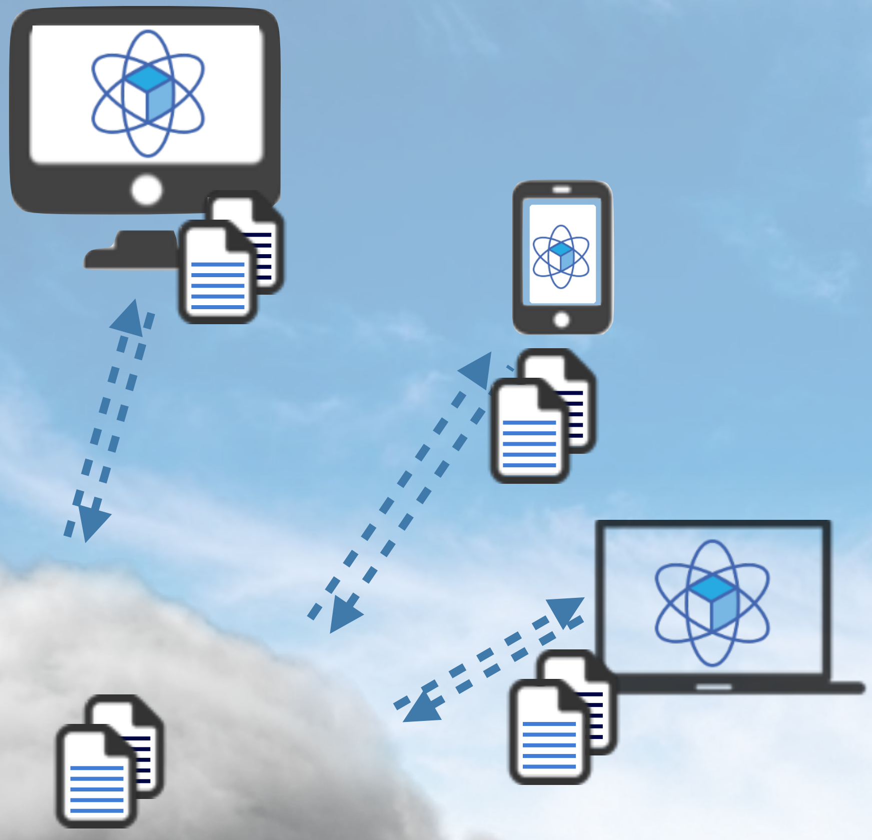
Community Exchange

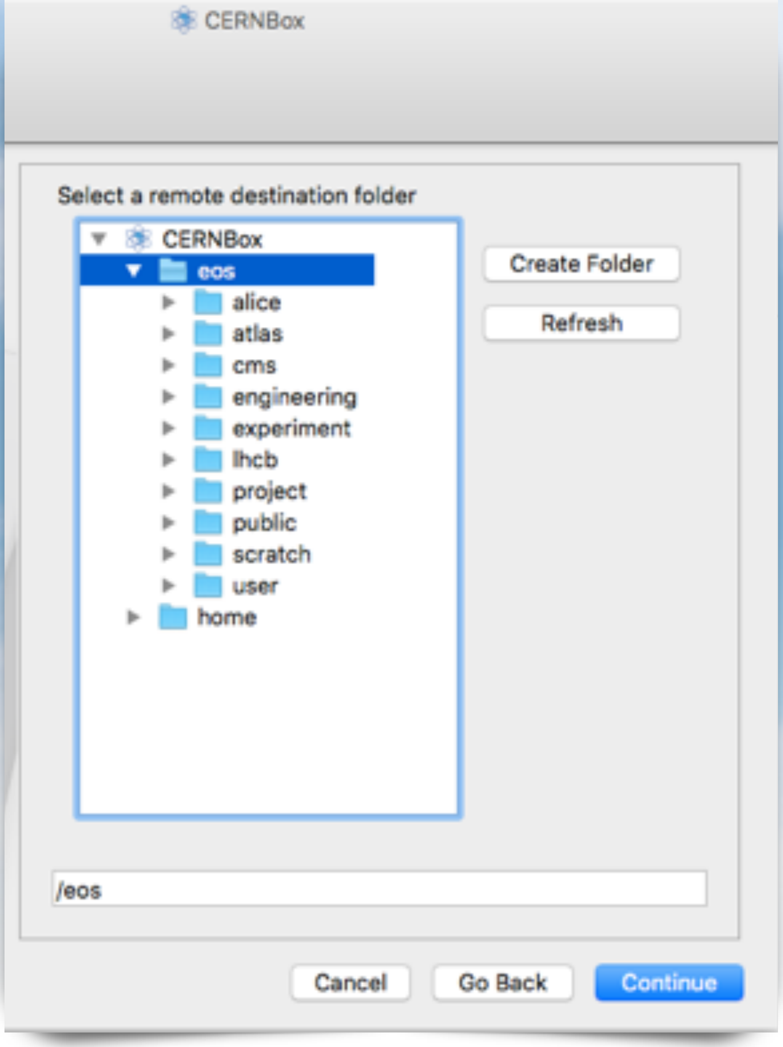
https://eos-community.web.cern.ch/

CERNBox: the CERN cloud storage driven by EOS

CERNBox powered by EOS

Bring data closer to our users: CERNBox





Select a remote destination folder

- CERNBox
  - eos
    - alice
    - atlas
    - cms
    - engineering
    - experiment
    - lhcb
    - project
    - public
    - scratch
    - user
  - home

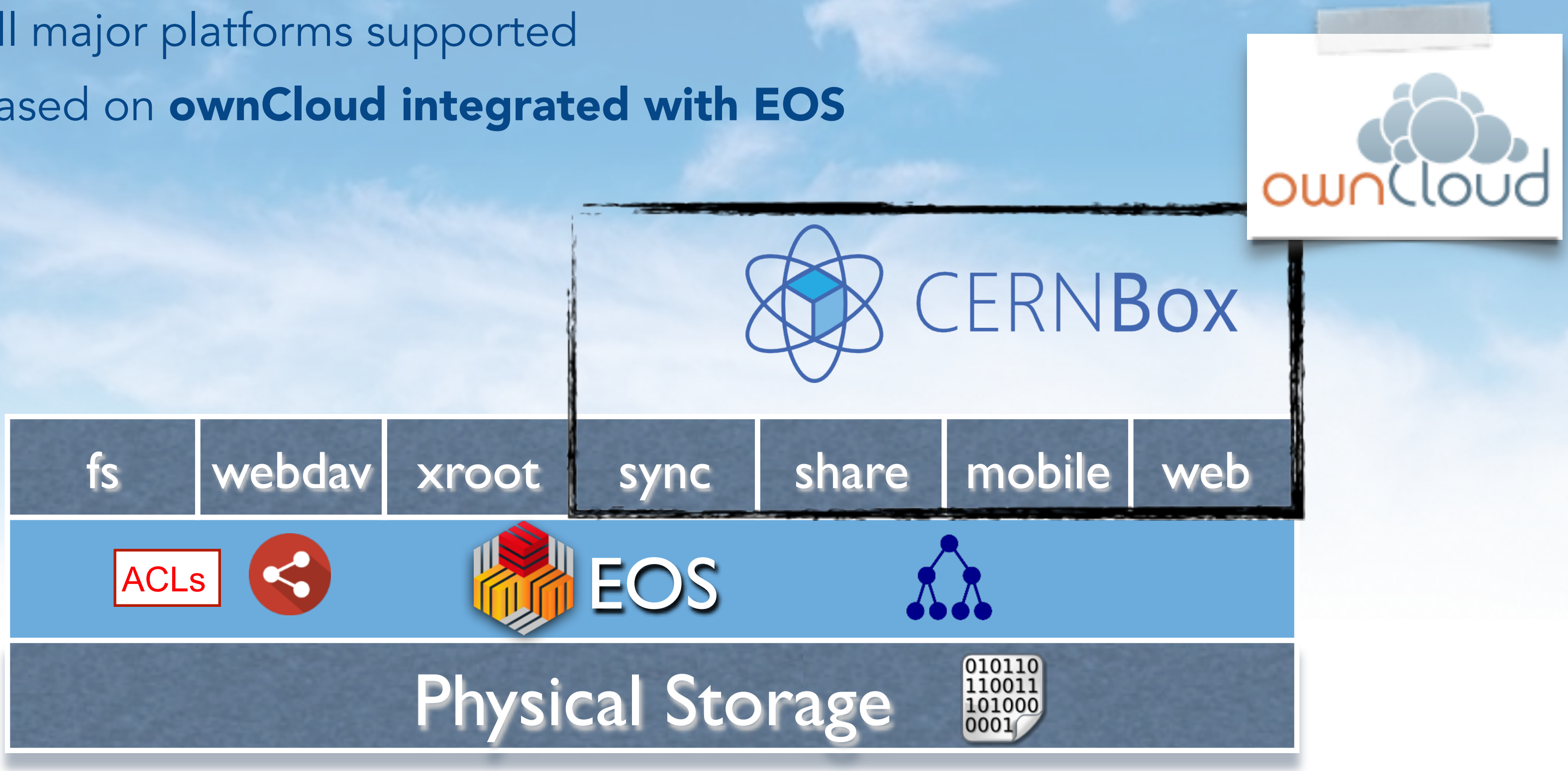
/eos

CERNBox powered by EOS

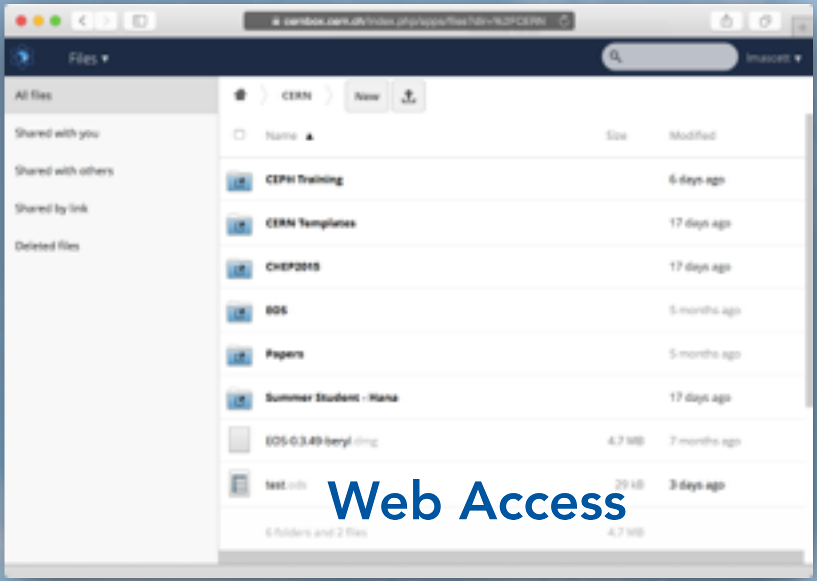
What CERNBox offers

CERNBox provides a cloud synchronisation service

- Available for all CERN users (1TB/user)
- Synchronise files (data at CERN) and offline data access
- Easy way to share with other users
- All major platforms supported
- Based on **ownCloud integrated with EOS**

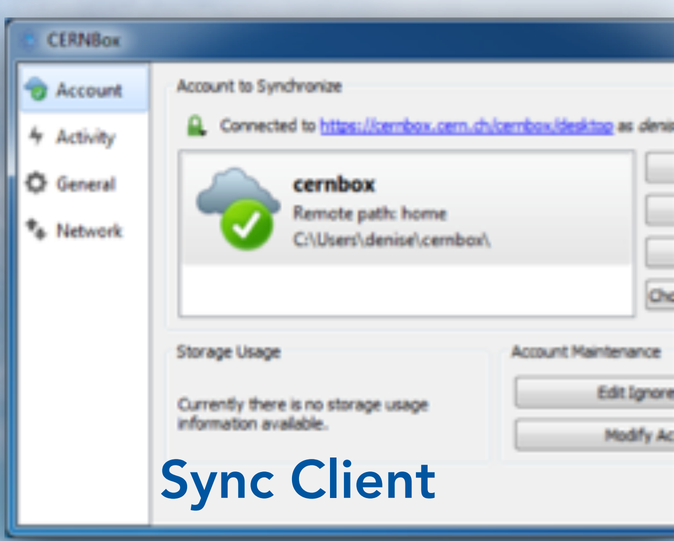


Available Access Methods



Name ▲	Size	Modified
CEPH Training		6 days ago
CERN Templates		17 days ago
CHEP2015		17 days ago
EOS		5 months ago
Papers		5 months ago
Summer Student - Hana		17 days ago
EOS-0.3.49-beryl.dmg	4.7 MB	7 months ago
test.ods	29 kB	3 days ago
6 folders and 2 files	4.7 MB	

Web Access



Account to Synchronize

Connected to https://cernbox.cern.ch/cernbox/desktop as deni

cernbox
Remote path: home
C:\Users\denise\cernbox\

Storage Usage

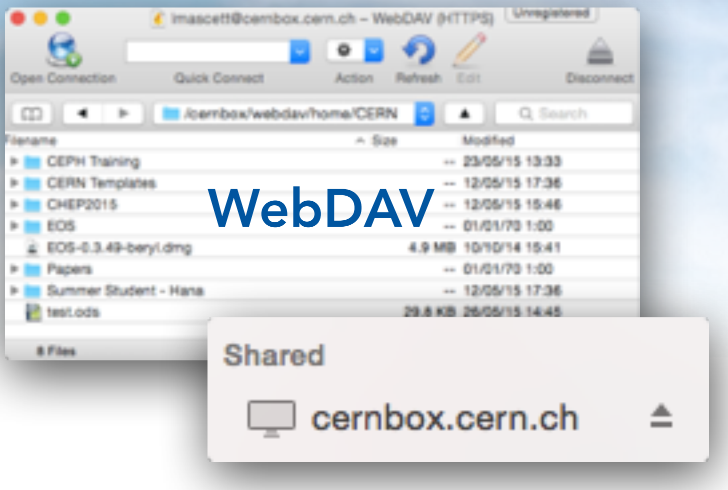
Currently there is no storage usage information available.

Account Maintenance

Edit Ignore

Modify Ac

Sync Client



Filename	Size	Modified
CEPH Training	--	23/05/15 13:33
CERN Templates	--	12/05/15 17:36
CHEP2015	--	12/05/15 15:46
EOS	--	01/01/70 1:00
EOS-0.3.49-beryl.dmg	4.9 MB	10/10/14 15:41
Papers	--	01/01/70 1:00
Summer Student - Hana	--	12/05/15 17:36
test.ods	29.8 KB	26/05/15 14:45

8 Files

Shared

cernbox.cern.ch

WebDAV

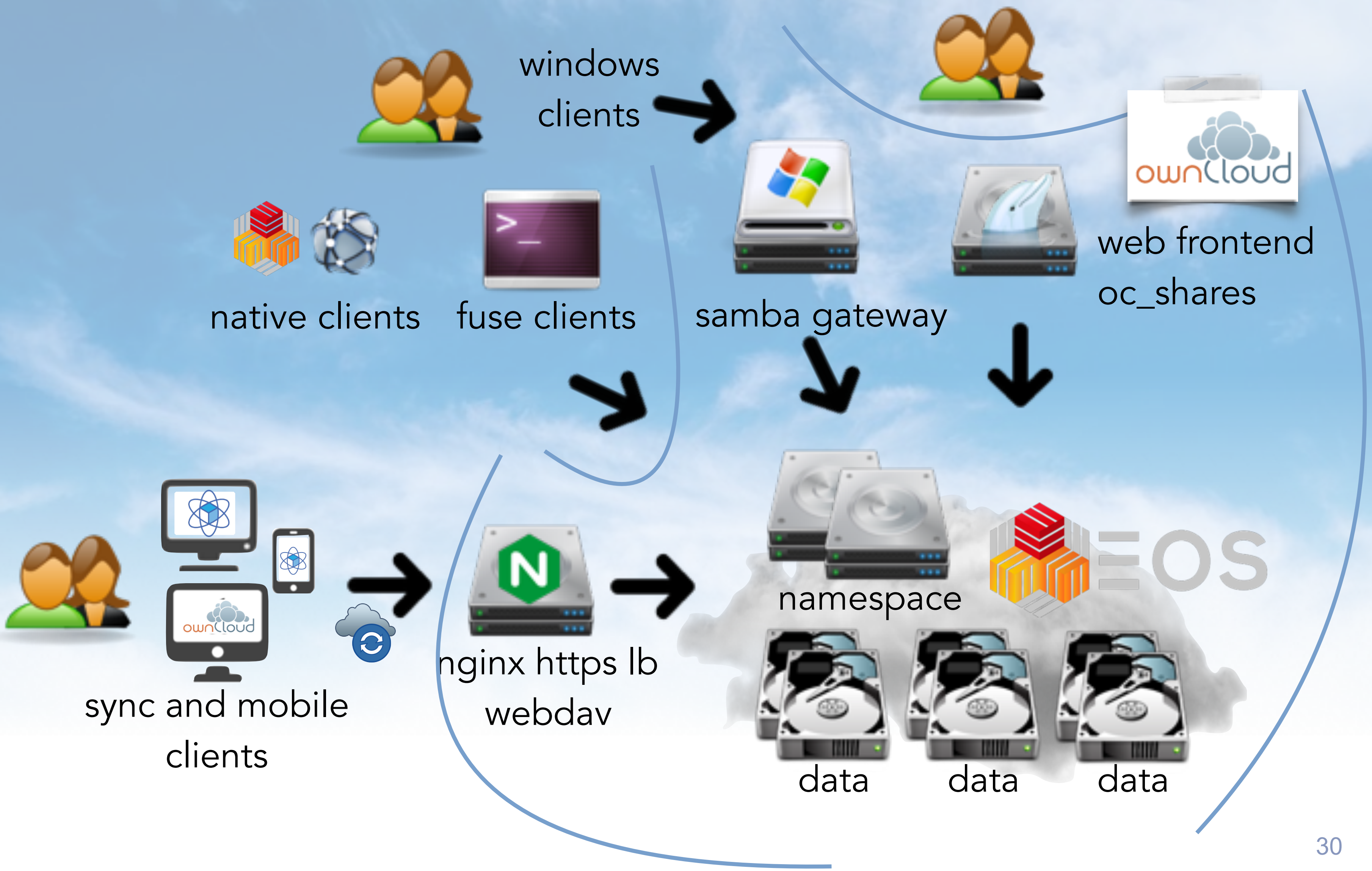
Mobile App

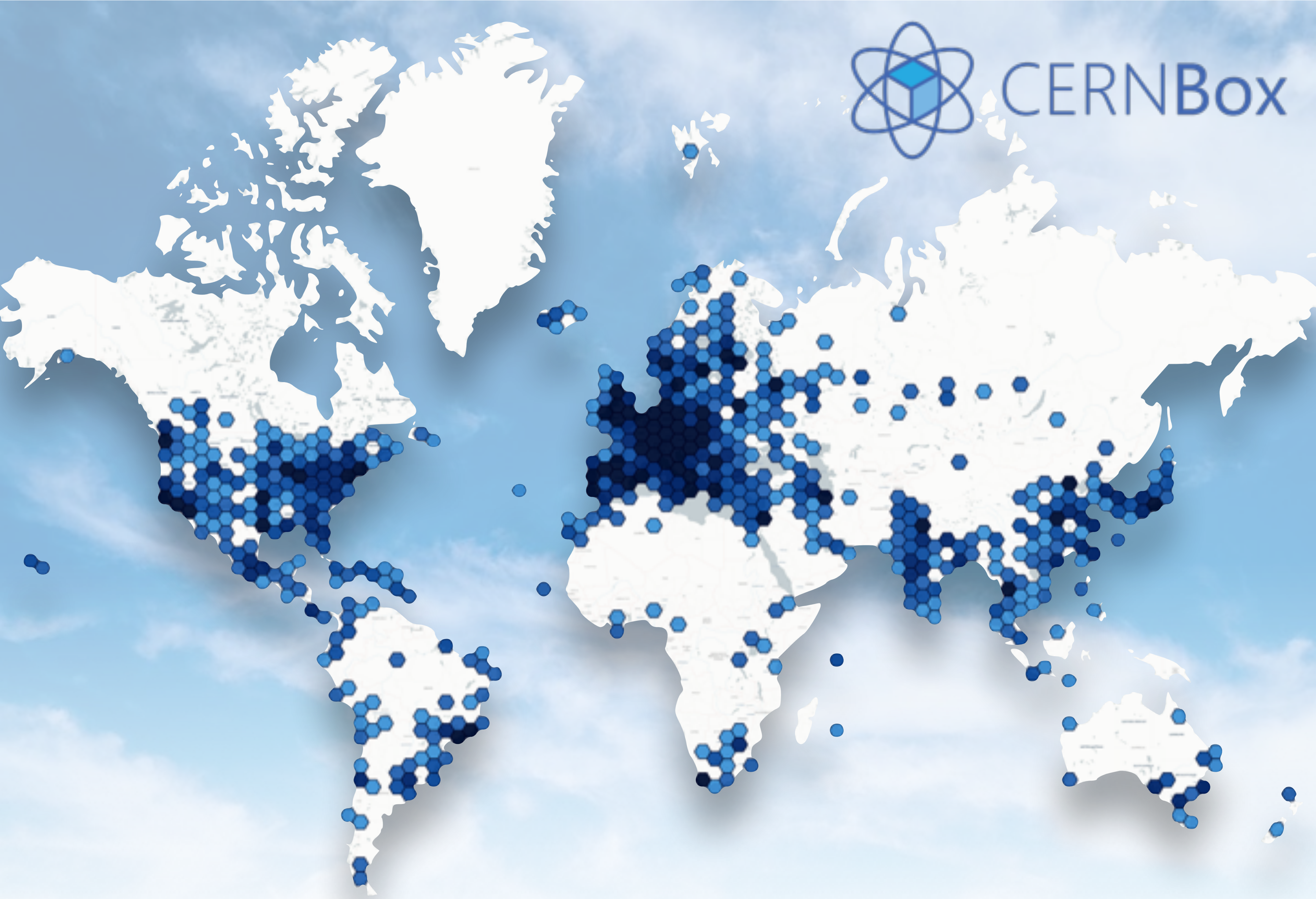
Directly from the storage backend

EOSUSER

(xroot, http, s3, ...)

EOS: the CERNBox backend





CERNBox

© OpenStreetMap contributors,© CARTO

Summary and Outlook

Summary and Outlook

EOS open storage provides a **very flexible** platform for large communities

- storage technology used to store LHC data
- more than 17k users storing data today via **CERNBox**

Demonstrated unprecedented scalability

- largest **low-cost** High-Energy Physics storage installation

CERNBox as an extension of the Desktop

- Bring data closer to our users
- New ways to interact with the data

Strategic for CERN based disk storage

Thanks for the attention!

CERN

www.cern.ch