
CERN

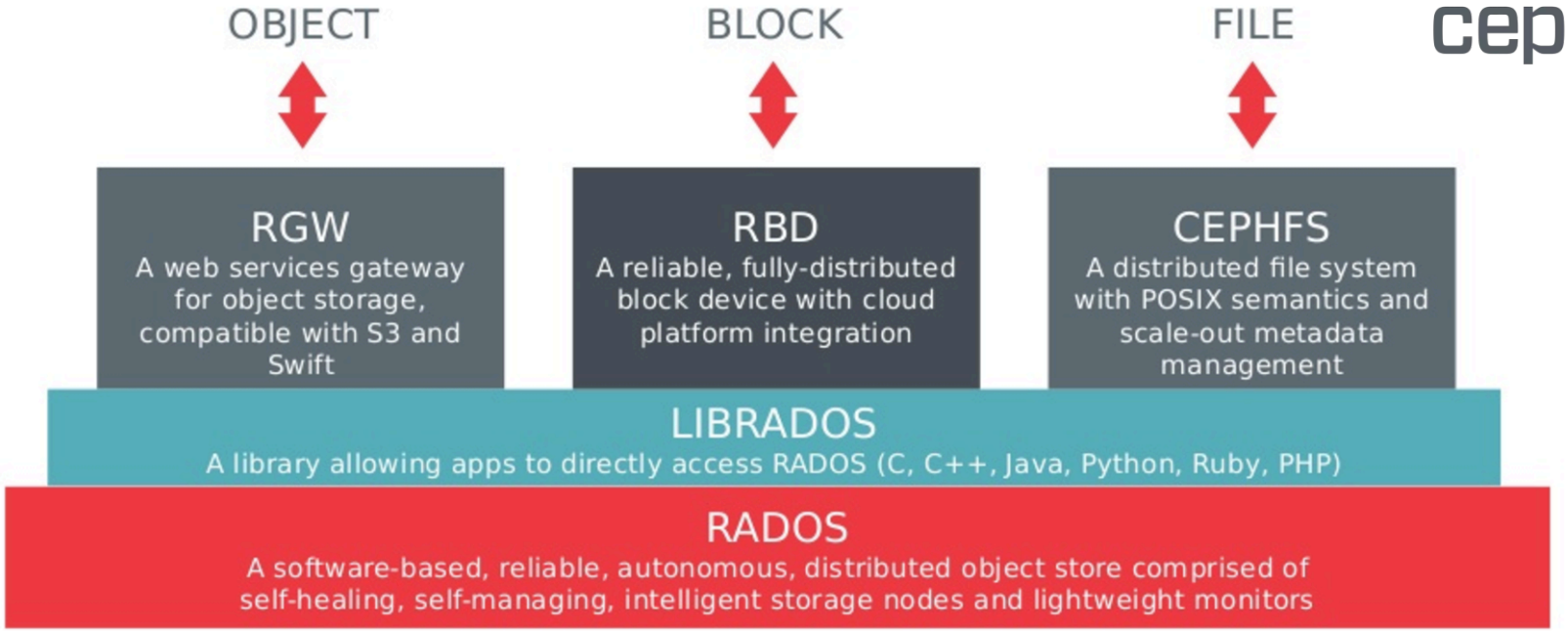
cephS3@CERN

Dan van der Ster
CERN IT Storage Group
12 September 2017 – WLCG PreGDB

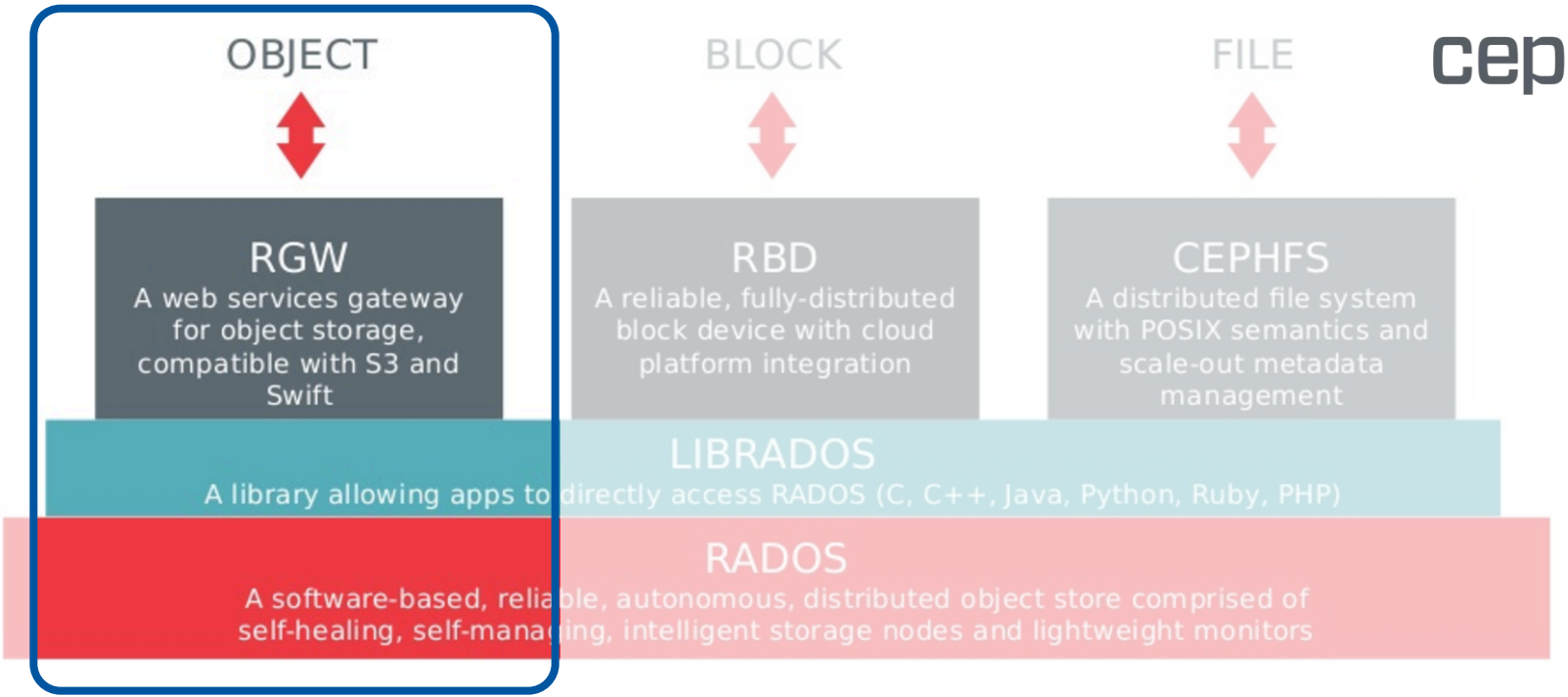
What's So Bad About POSIX I/O?

https://www.nextplatform.com/2017/09/11/whats-bad-posix-io/

Ceph at a Glance



Ceph at a Glance



Ceph S3 at CERN: cs3.cern.ch

- 865TB cluster, 369TB RAW used.
  - 32 OSD servers, 3 virtual Ceph mons, 10 virtual S3 gateways.

- Running Ceph jewel 10.2.8, servers run CentOS 7.3.
- S3 data pool uses **erasure-coding**: 4 data + 2 parity stripes.

- Buckets accessible as <bucket>.cs3.cern.ch or cs3.cern.ch/<bucket>
- http or https; Comodo SSL certificate for (*.)cs3.cern.ch

RGW Details

- Each radosgw is a (CERN OpenStack) m2.medium VM:
  - 4GB RAM, 10Gbps network (shared on host).
  - RGW's spread across all availability zones.
  - Ceph config: ops/usage logs disabled; 512 civetweb threads, 8 rados handles, 512 rgw threads

- Haproxy frontends to handle ssl and balance the load:
  - Haproxy running on same VMs as rgw.
  - Two haproxy "backends": 5 rgw's dedicated to ATLAS buckets, 5 rgw's for everyone else
    - Bucket mapping is hardcoded – we update as needed.

Users/Usage/Quotas

- ATLAS (~160TB):
  - *atlaspanda*: event service + panda logs
  - Also *atlasdatafed* (300GB) and *atlaseventindextest (2.2TB)*
- CMS (~100TB):
  - *cmsboinc*: staging area for boinc inputs/outputs
- IT resource accounting (500GB)
- Also test accounts for CVMFS, Gitlab-CI, DynaFed, OpenStack Trove, Terraform, Oracle backups

- We set a bytes quota per user account. (e.g. atlaspanda has a 200TB quota)
  - Also possible: num entries quotas, num buckets quotas, quota per bucket
- Querying space used is difficult: with S3, you must iterate over all objects (e.g. du).

Getting usage with swift API

S3 does not have a summary usage method, but swift does:

```
@aiadm27 ~> swift -A http://cs3.cern.ch/auth/v1.0 -U dvanders:swift -K xxx stat
                   Account: v1
                Containers: 3
                   Objects: 52907
                     Bytes: 246986100596
                X-Timestamp: 1505127994.37179
X-Account-Bytes-Used-Actual: 247098736640
                X-Trans-Id: tx000000000000006a704b-0059b66e39-666462-default
              Content-Type: text/plain; charset=utf-8
             Accept-Ranges: bytes
```

See https://developer.openstack.org/api-ref/object-storage/?expanded=show-account-metadata-detail

Things to investigate

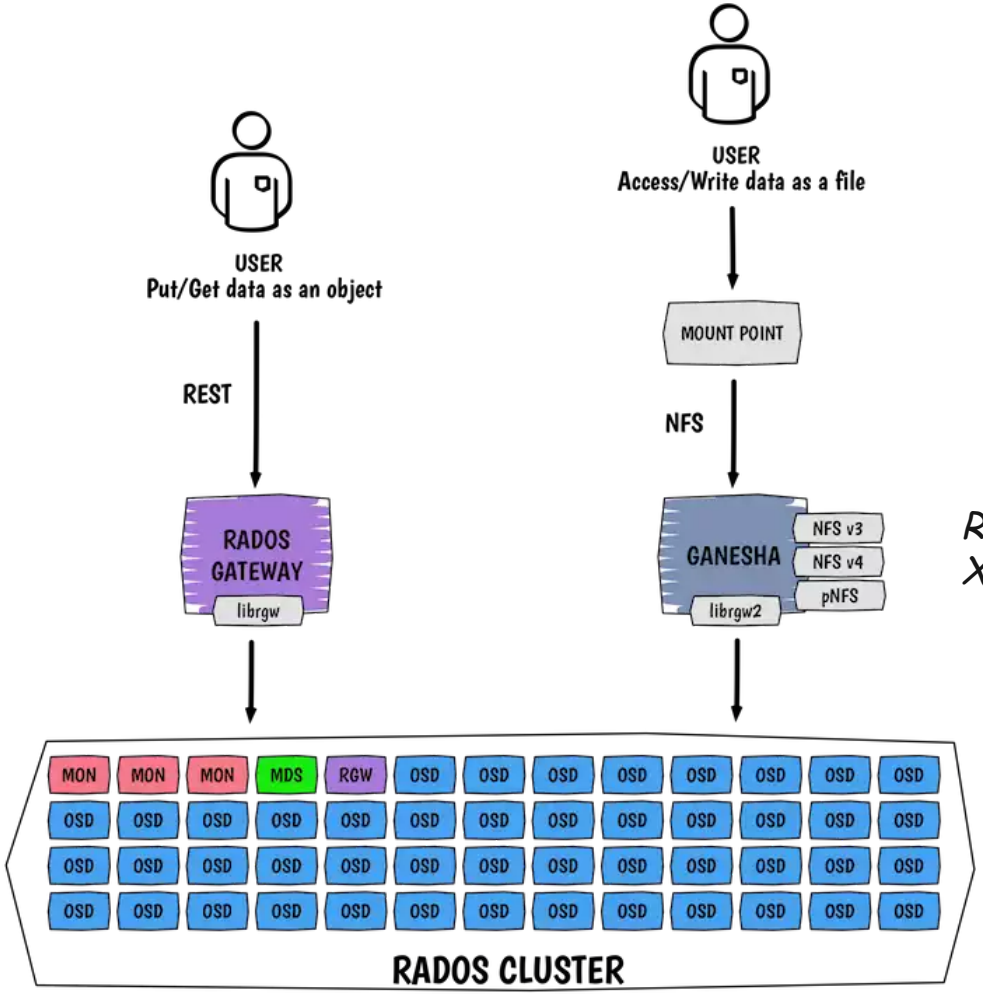
- Update to Ceph Luminous:
  - Object/Bucket policies for object expiration.
  - Huge buckets (via dynamic bucket resharding).
  - “Perfect” data balancing: less overhead, more space for user data
  - Object metadata search: http://ceph.com/rgw/new-luminous-rgw-metadata-search/

- Second region needed? (Ceph S3 can mirror buckets into a separate cluster)
- OpenStack Keystone integration: create S3 keys directly via OpenStack.

Things to investigate: librgw

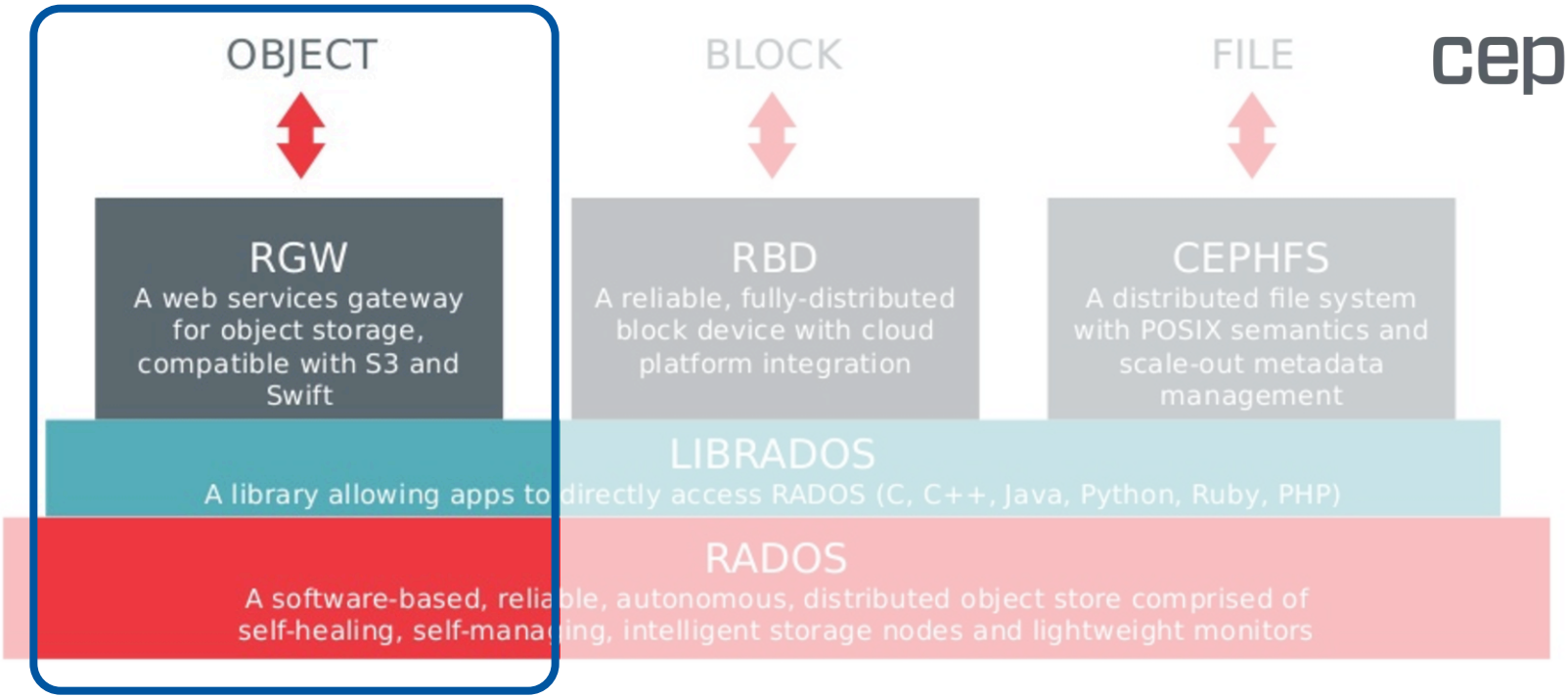
- **NFS-RGW** bridge: nfs-ganesha can re-export Ceph-S3 buckets as NFS. See **FSAL_RGW**.
  - Useful to give legacy clients access into S3.
  - https://www.sebastien-han.fr/blog/2016/12/23/Ceph-Rados-Gateway-and-NFS/

- **librgw**: the radosgw has been refactored into a library. Allows (trusted) apps to have a sort of local S3 gateway to Ceph.
  - Xrootd or GridFTP bridges to RGW should be doable.
  - The lib: https://github.com/ceph/ceph/blob/master/src/rgw/librgw.cc
  - Example usage: https://github.com/nfs-ganesha/nfs-ganesha/tree/next/src/FSAL/FSAL_RGW

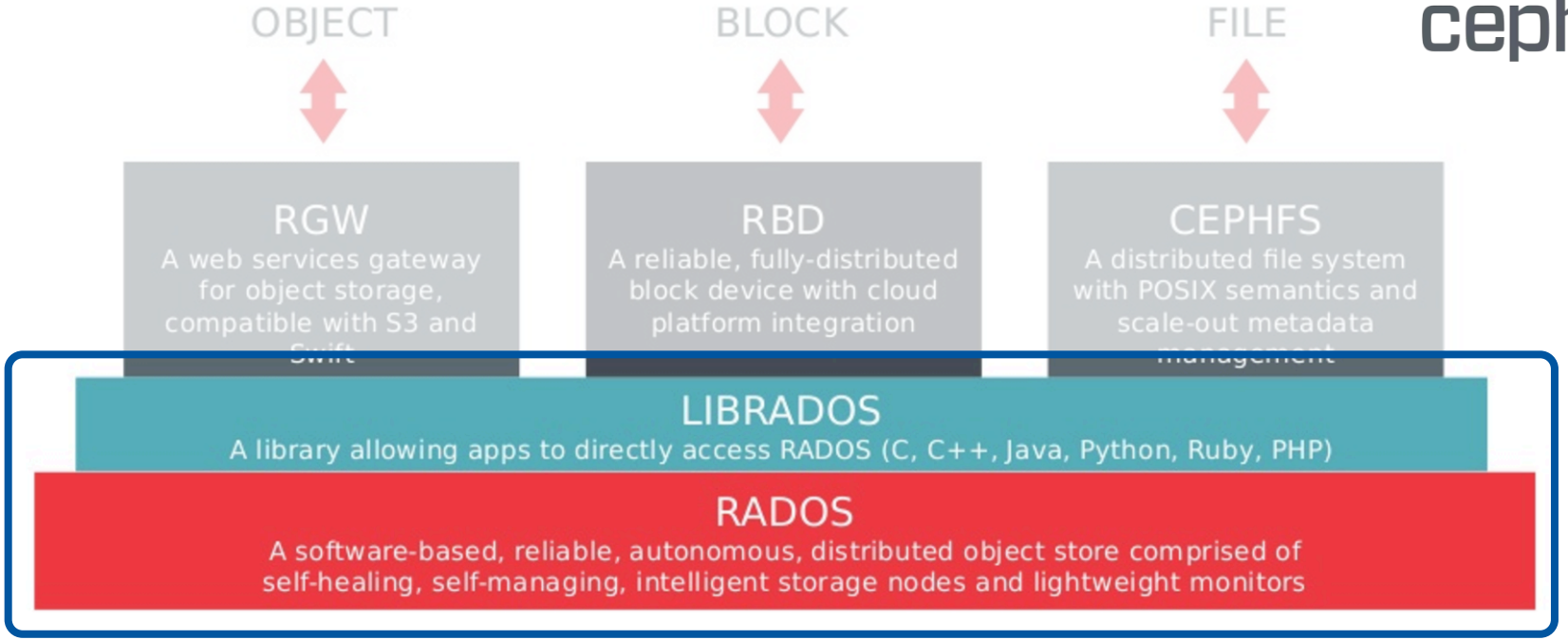


Repeat for Xrootd/GridFTP

Ceph at a Glance

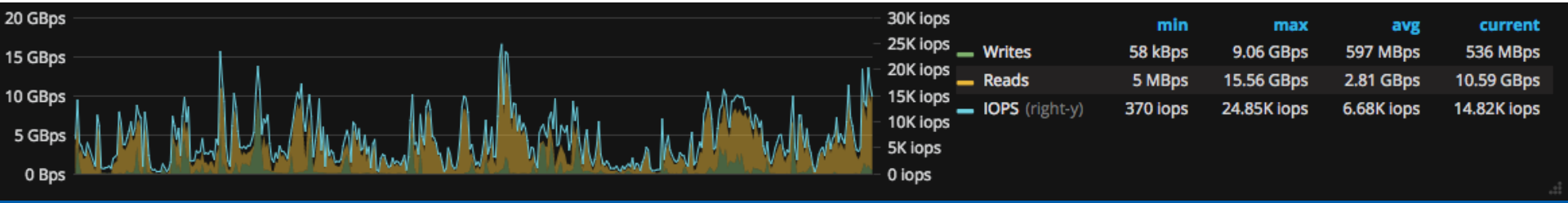


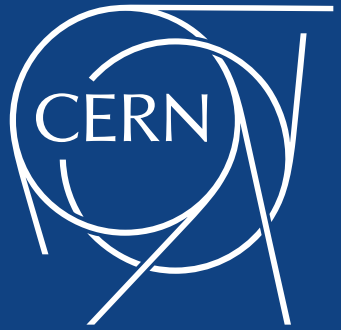
Ceph at a Glance



RADOS Object Storage

- CASTOR ALICE backed by **Ceph RADOS**
- Thin CASTOR gateways using **libradosstriper**: write striped objects to a Ceph RADOS pool
  - Useful for faster parallel streaming, and to keep object sizes small.
  - Xrootd gateways use lots of memory buffering RADOS IO – improvements needed.
- 1.6PB used x2 replicas: has been reliable, but we want to move to EC.
  - 42 disk servers, 4PB RAW. Peaks of >15GBps reading, >9GBps writing.




CERN

radosgw configuration

```
[client.radosgw]
  keyring = /etc/ceph/ceph.client.radosgw.keyring
  rgw content length compat = true
  rgw dns name = cs3.cern.ch
  rgw enable ops log = false
  rgw enable usage log = false
  rgw frontends = civetweb port=8080 num_threads=512
  error_log_file=/var/log/radosgw/civetweb.error.log
  rgw num rados handles = 8
  rgw swift url = http://cs3.cern.ch
  rgw thread pool size = 512
```