A Ceph plugin for XrootD

Sébastien Ponce
sebastien.ponce@cern.ch

CERN

January 28th 2015

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Outline

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Introducing Ceph

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Ceph is an object Store

- An Object is
  - a piece of data
  - metadata (extended attributes)
  - a unique identifier
- Different from a file system
  - Get/Put semantic
  - No file descriptor
  - Flat namespace
    aka no namespace
  - "small" objects (few MBs)

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

At the core : CRUSH

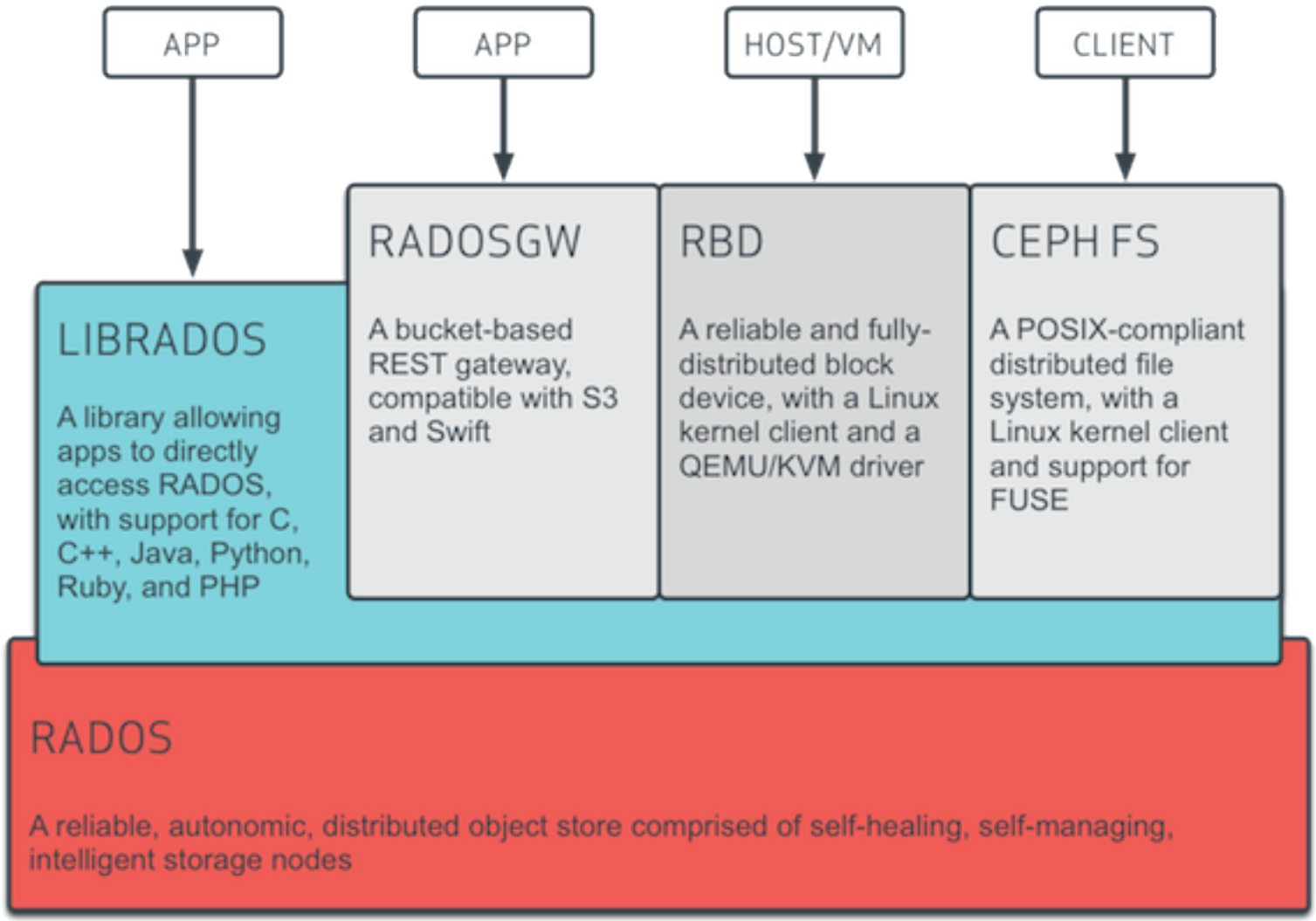
Controlled Replication Under Scalable Hashing

`http://ceph.com/papers/weil-crush-sc06.pdf`

- a clever data placement algorithm ...
  - supporting cluster maps (e.g. rows/cabinet/shells/devices)
  - approximating a uniform probability distribution
  - with completely deterministic mapping
  - but pseudo-random distribution
  - minimizing data movements on cluster evolution
- ... for a very scalable object store
  - no central catalog of object placement
  - placement computation done by clients
  - configurable replication (file by file)
  - erasure coding available

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

CEPH interfaces



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

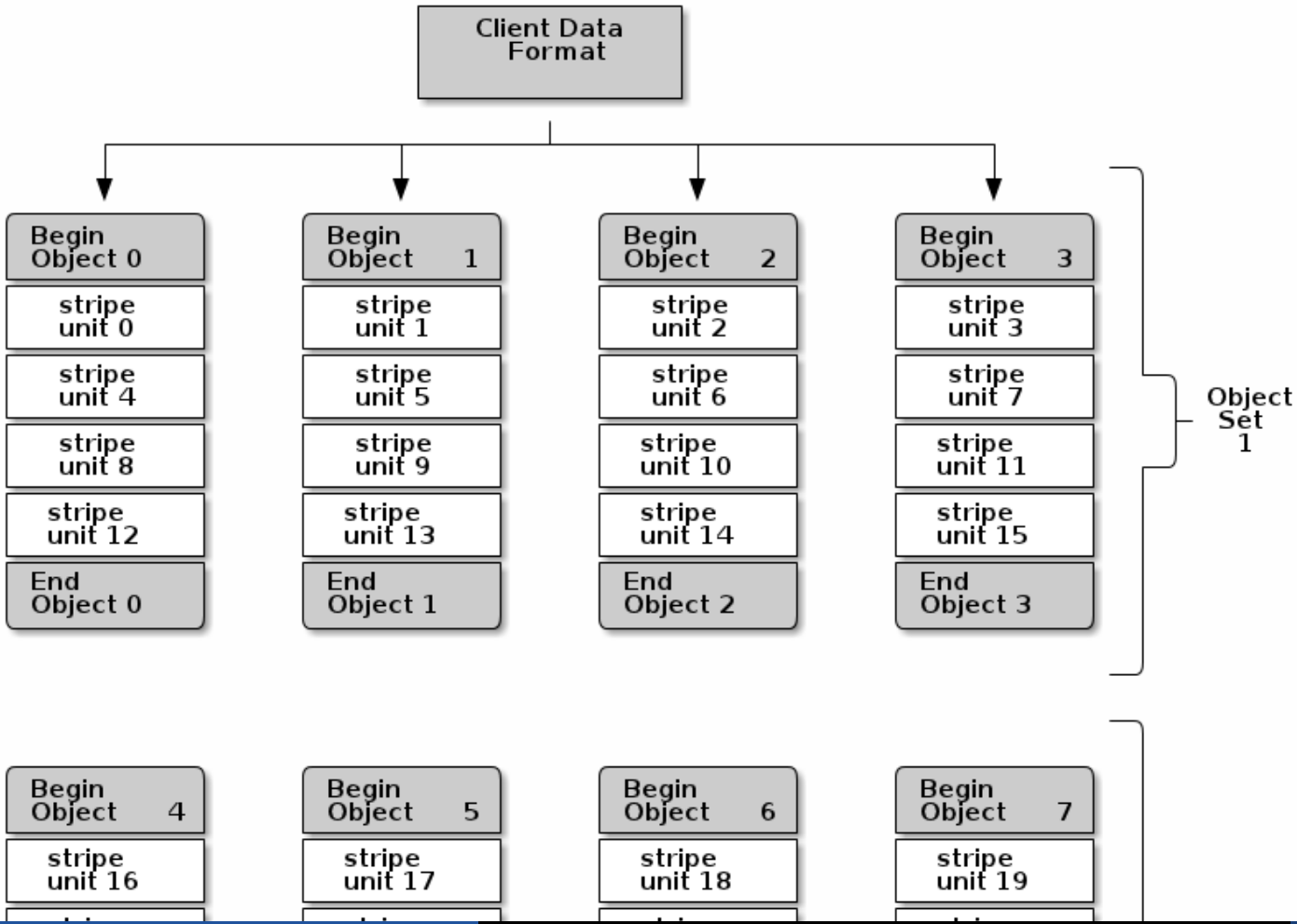
librados - libradosstriper

- librados
  - the natural object store interface
  - stong atomicity/snapshoting features
  - has object store limitations :
    - no namespace
    - objects should be small (logical I/O unit, order of MBs)
    - no parallel I/O within an object
- libradosstriper
  - adds striping on top of librados
  - reusing the striping of Ceph FS and RBD
  - available starting with giant release

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

libradosstriper

Defaults : 1 stripe, stripe unit = object size = 4 MB



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

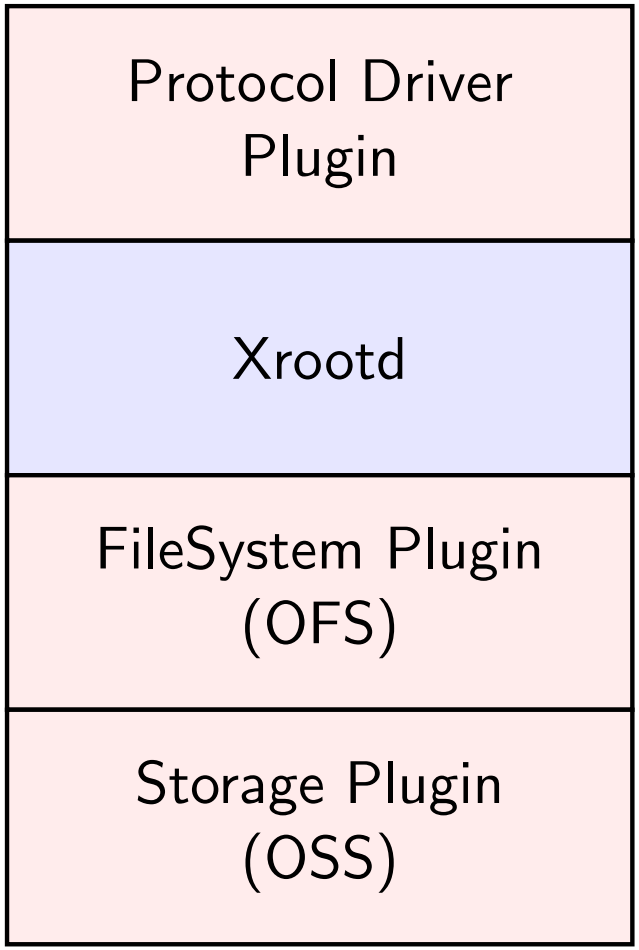
The Ceph plugin of XrootD

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

An OSS plugin

OSS = Open Storage System

- XrdOSS / XrdOssDf interfaces
  - implementing entry points to the underlying storage
  - e.g. open, read, write, close
- compatible with any OFS or Protocol plugin
  - e.g. Castor, EOS, HTTP



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Based on libradosstriper

PROs

- Taking full benefit of the distributed object store
- No file size limit
- Embedded striping and parallel I/O
  - parallel I/O is not activated by default (nb stripes = 1)

A small drawback

- No directory listing

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

External attributes

Why to deal with external attributes ?

- Xrootd has no client interface for them
- OSS plugins do not support external attributes
- ... but an OFS plugin may need them !

How they are supported

- via the XrdSysXattr interfaces intoduced in Xrootd 4.1
- implemented via an extra plugin

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Practical usage and status

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Practical usage (basic)

Config file syntax

```
ofs.osslib     libXrdCeph.so
ofs.xattrlib   libXrdCephXattr.so
all.export     *?
```

Usage

```
# xrdcp myfile root://myserver/myfile
  [1000MB/1000MB][100\%][=======][17.24MB/s]
# rados ls | grep myfile
  myfile.0000000000000000
  myfile.0000000000000001
# xrdcp root://myserver/myfile myfile2
  [1000MB/1000MB][100\%][=======][16.38MB/s]
```

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Multi user, multi pools

Config file syntax

```
ofs.osslib libXrdCeph.so [[user@]pool]
```

Extended file syntax

```
[[user@]pool:]path
```

Examples

```
xrdcp root://myserver/mypool:myfile ...
xrdcp root://myserver/myuser@mypool:myfile ...
xrdcp root://myserver/:file_with_a:in_it ...
```

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Adapting file layout

Complete config file syntax

```
ofs.osslib <lib> [[user@]pool[,<layout>]]
 layout : nbStripes[,stripeUnit[,objSize]]
```

Complete file syntax

```
[[user@]pool[,<layout>]:]path
```

Examples

```
myuser@mypool,4:myfile
myuser@mypool,4,65536:myfile
mypool,1,33554432,33554432:myfile
```

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Current status

- code is ready and under testing
- available on github (`http://github.com/sponce/xrootd`)
- to be integrated soon into the main trunk
- will be part of release 4.2

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it