



Ceph

Loïc Dachary

Artisan Logiciel Libre



- A l'aube des temps
- Un métier à échelle humaine
- Participer à une révolution
- L'avenir du stockage
- Travailler avec des visionnaires de talent

Developeur Ceph 9/212

```
$ git shortlog -nes | head -10
```

```
17425 Sage Weil <sage@inktank.com>
```

```
2898 Yehuda Sadeh <yehuda@inktank.com>
```

```
2185 Greg Farnum <greg@inktank.com>
```

```
1690 Samuel Just <sam.just@inktank.com>
```

```
1391 Colin P. McCabe <colinm@hq.newdream.net>
```

```
1053 Josh Durgin <josh.durgin@inktank.com>
```

```
942 Danny Al-Gaaf <danny.al-gaaf@bisect.de>
```

```
884 John Wilkins <john.wilkins@inktank.com>
```

```
589 Loic Dachary <loic@dachary.org>
```

```
490 Dan Mick <dan.mick@inktank.com>
```

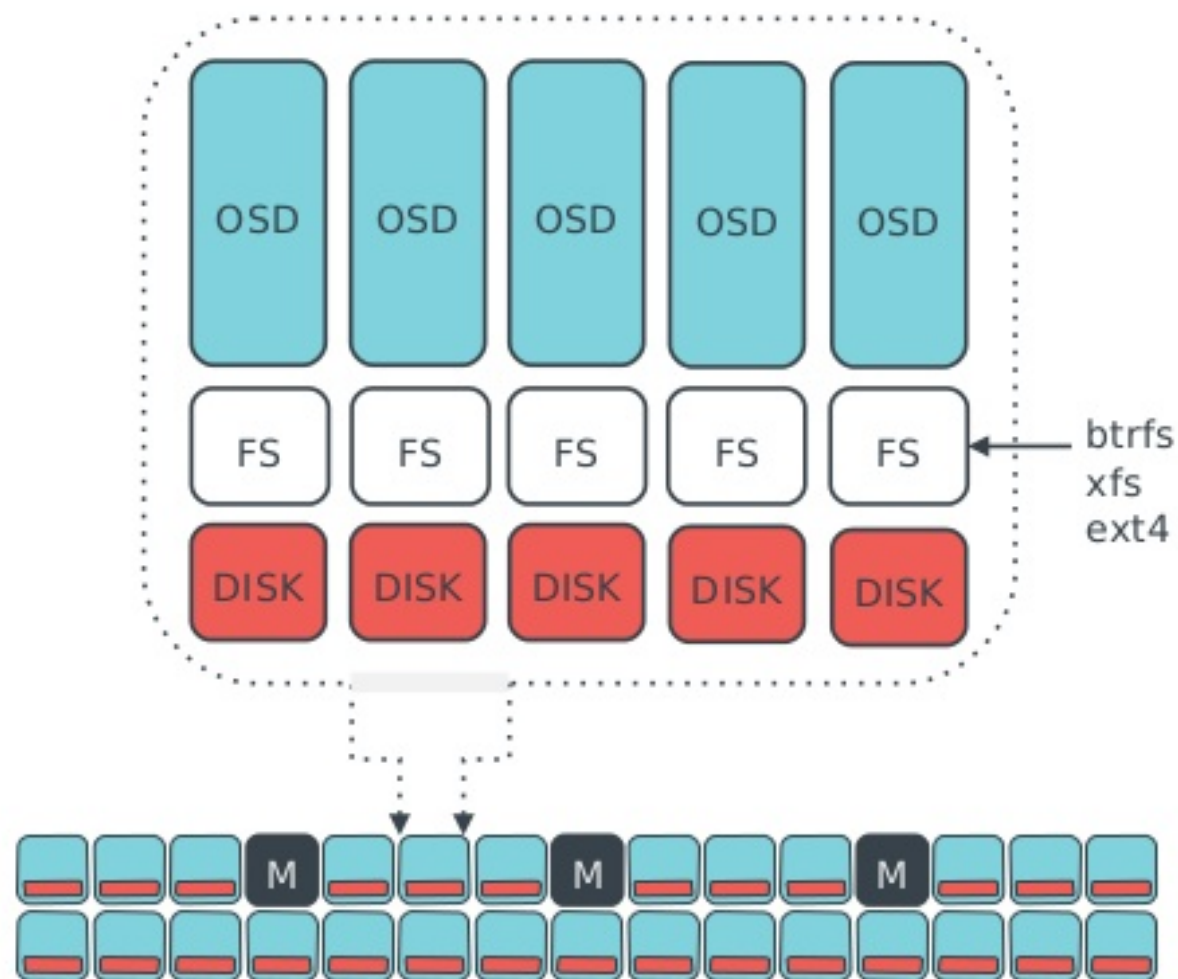
C'est quoi Ceph ?

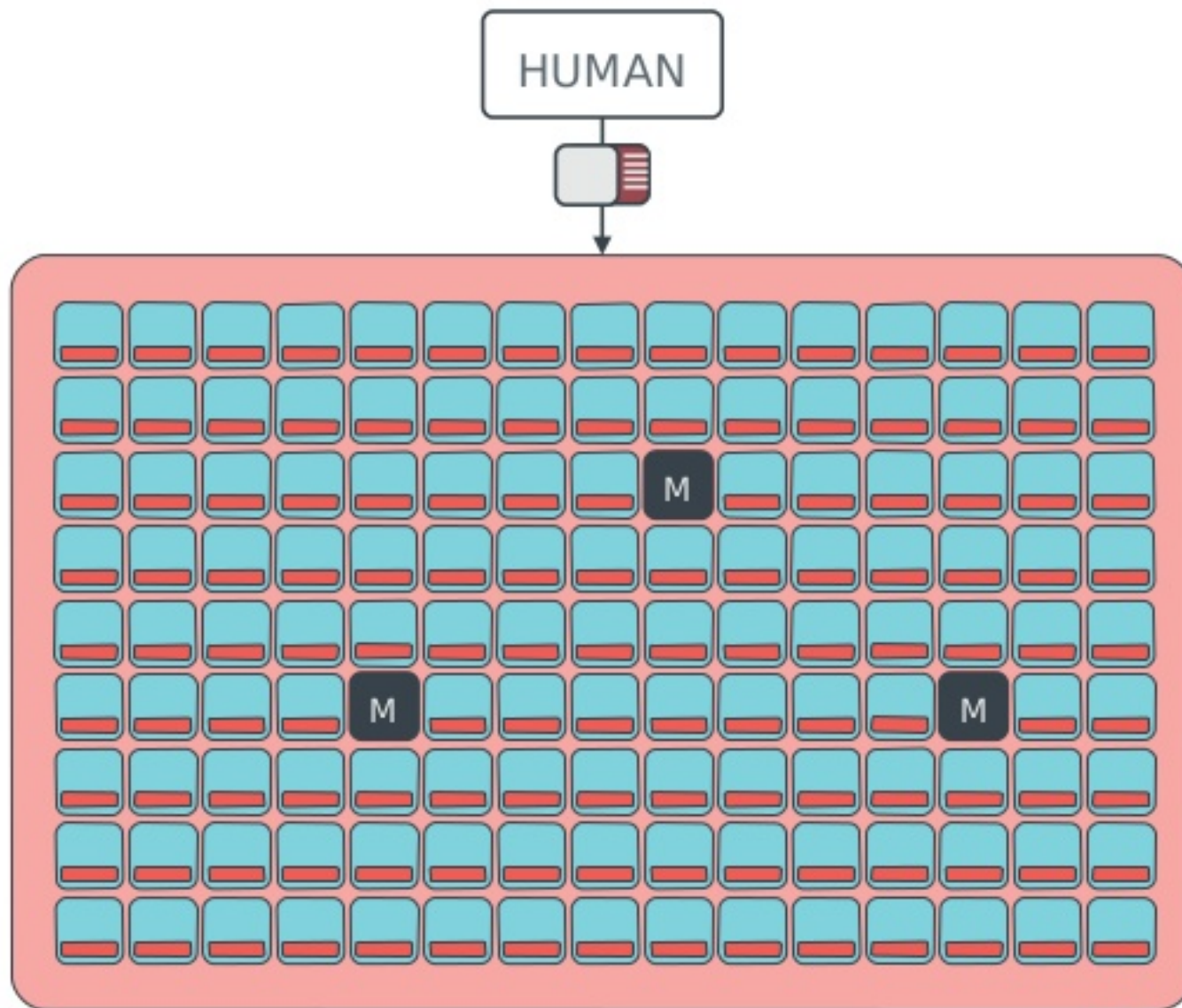
- Un stockage d'objets
- Distribue et équilibre les données
- Les opérations de base : get / put / delete
- L'unité de stockage est l'objet
- Stockage redondant

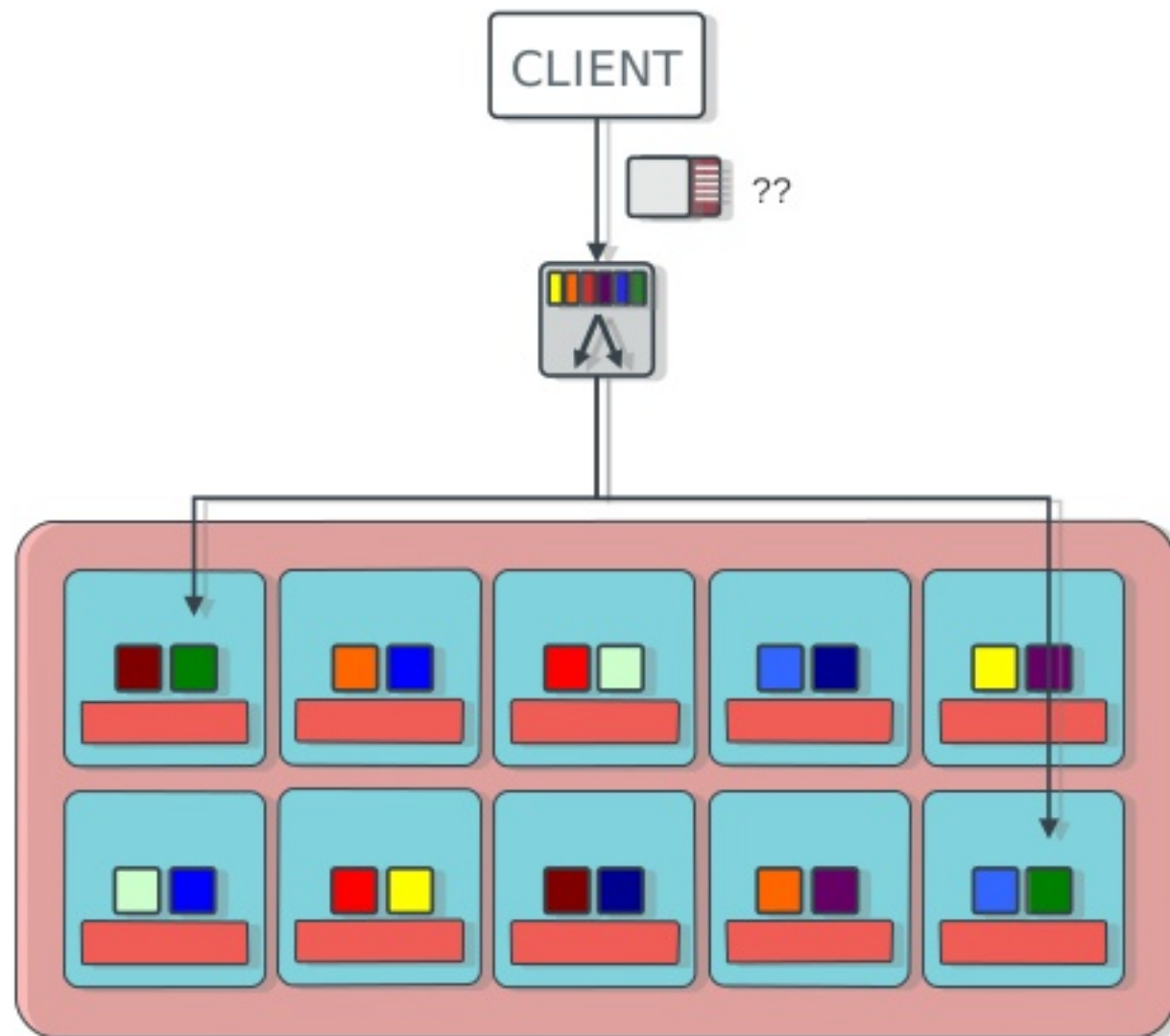
C'est qui Ceph ?

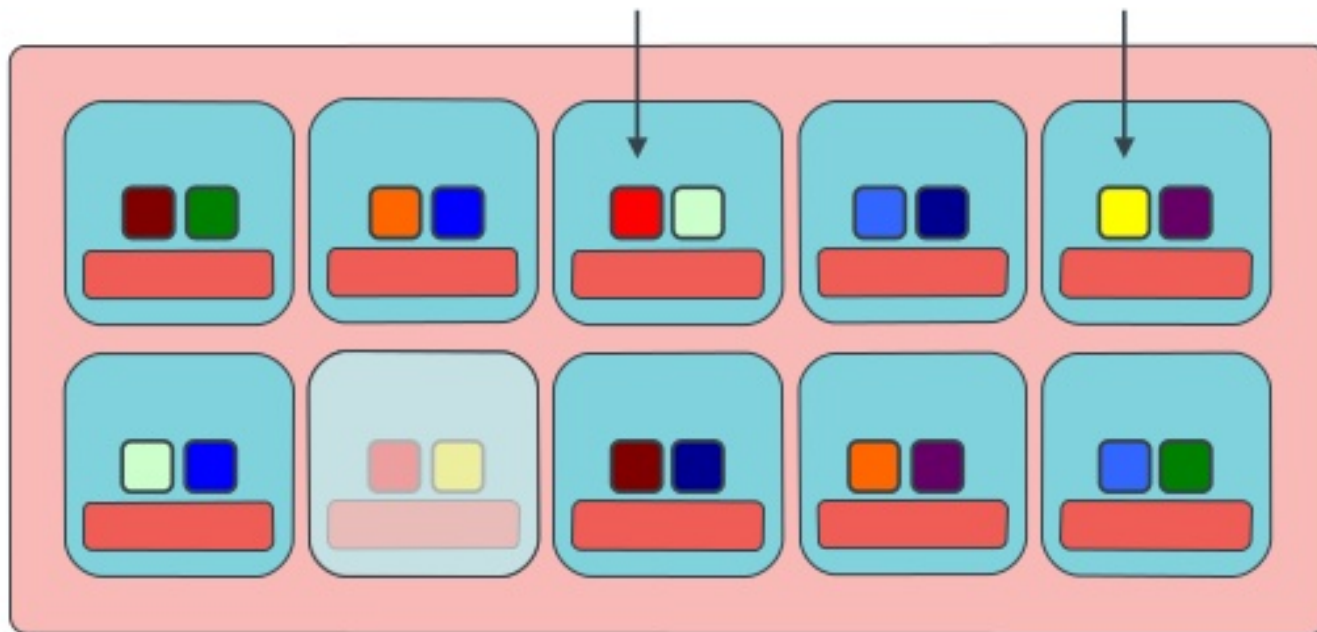
Commits by organizations

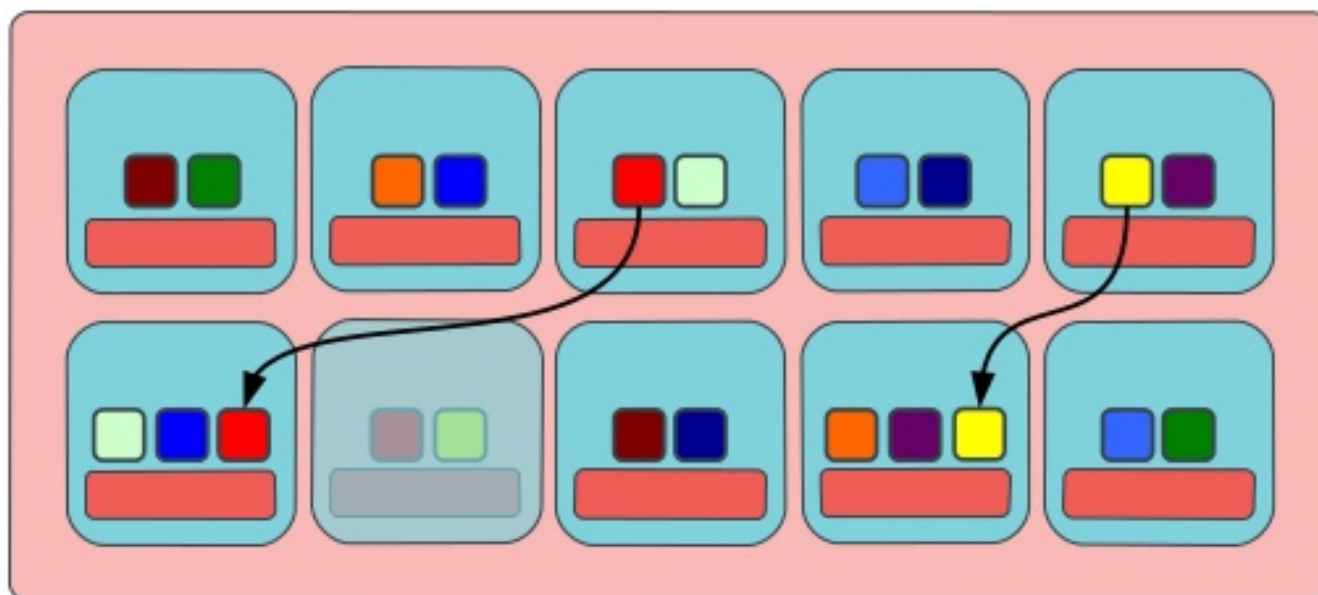
- 1 21495 Inktank <contact@inktank.com>
- 2 1421 DreamHost <contact@dreamhost.com>
- 3 946 Deutsche Telekom <contact@telekom.de>
- 4 622 Cloudwatt <libre.licensing@cloudwatt.com>
- 5 456 Intel <contact@intel.com>

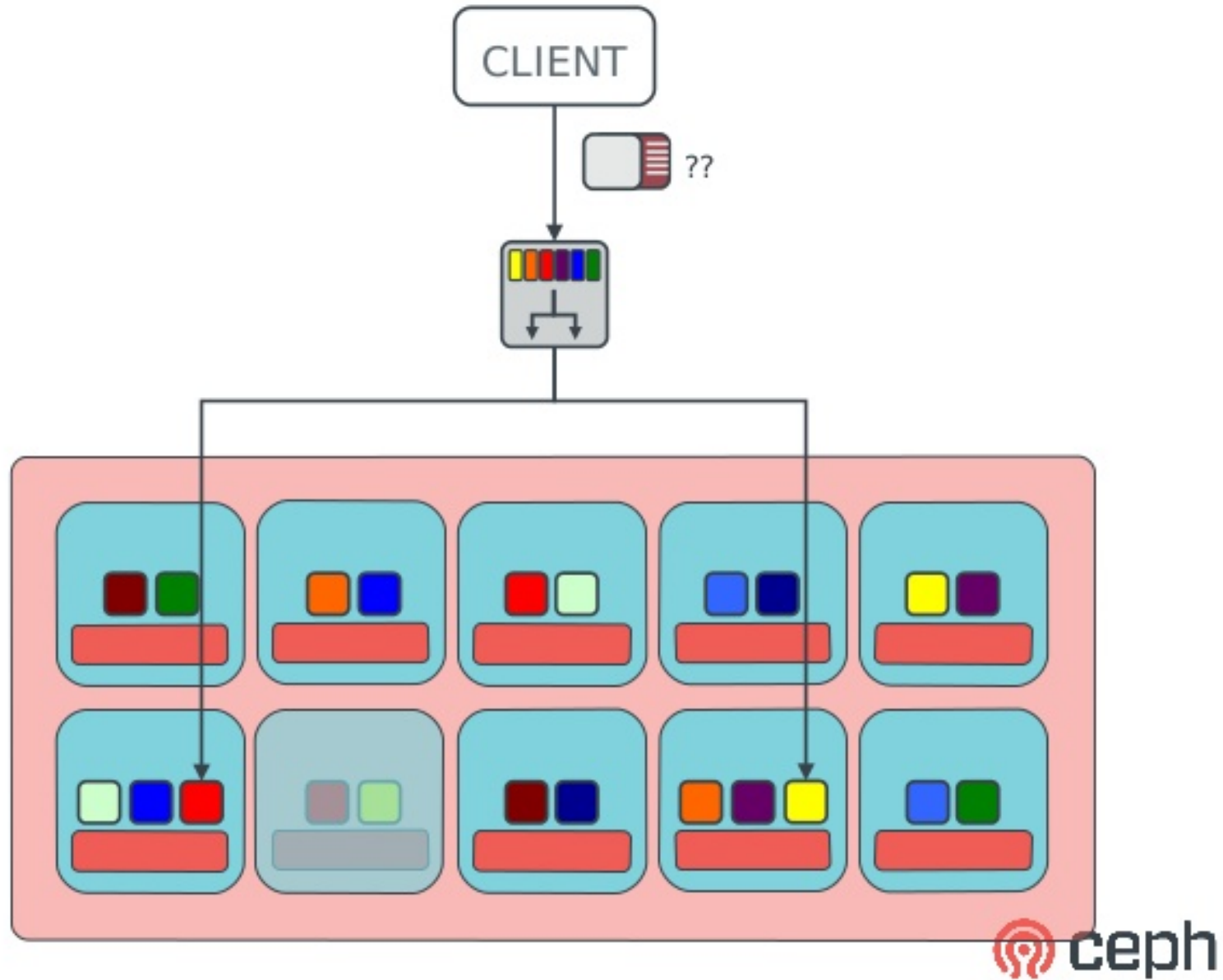


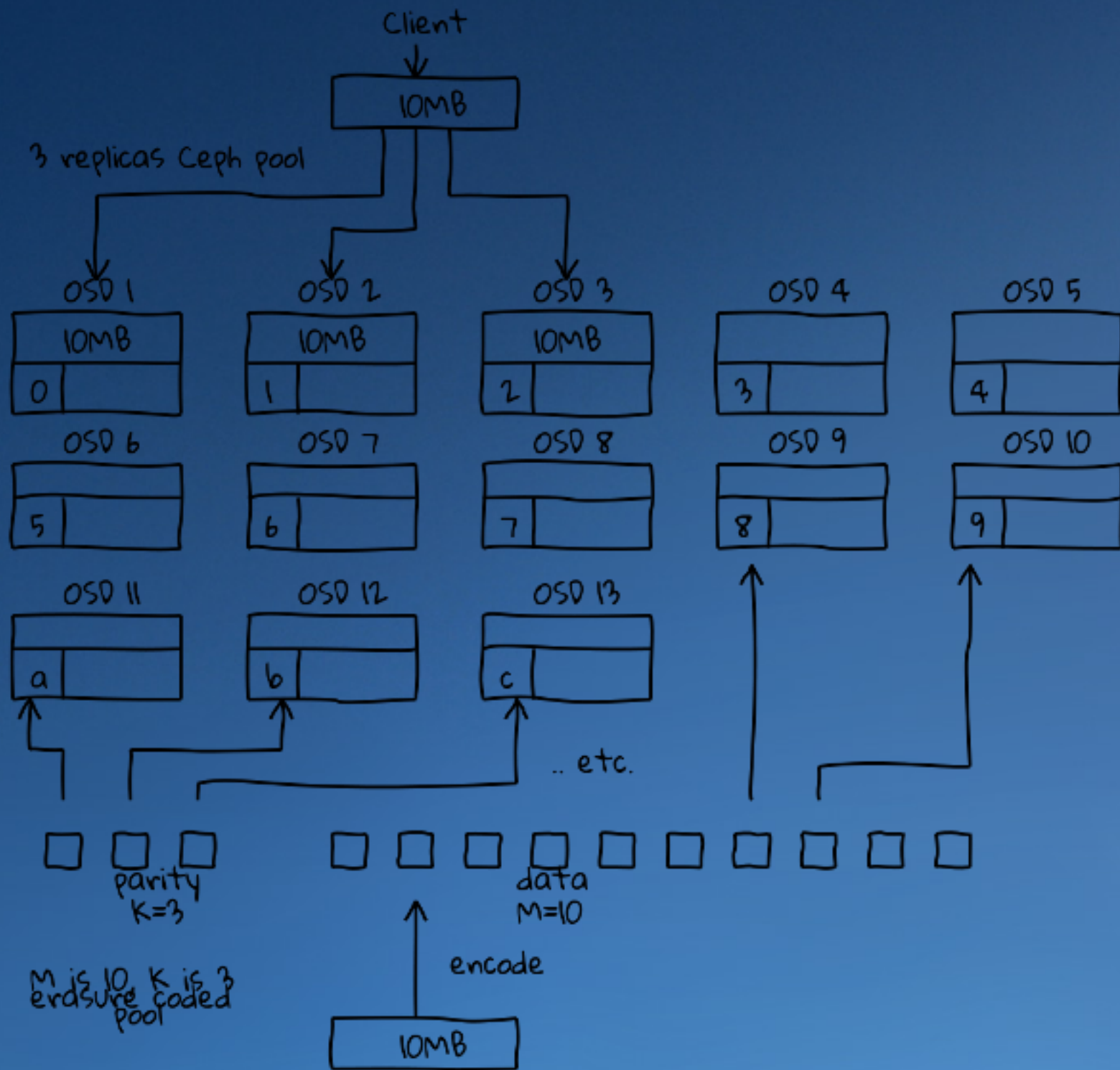






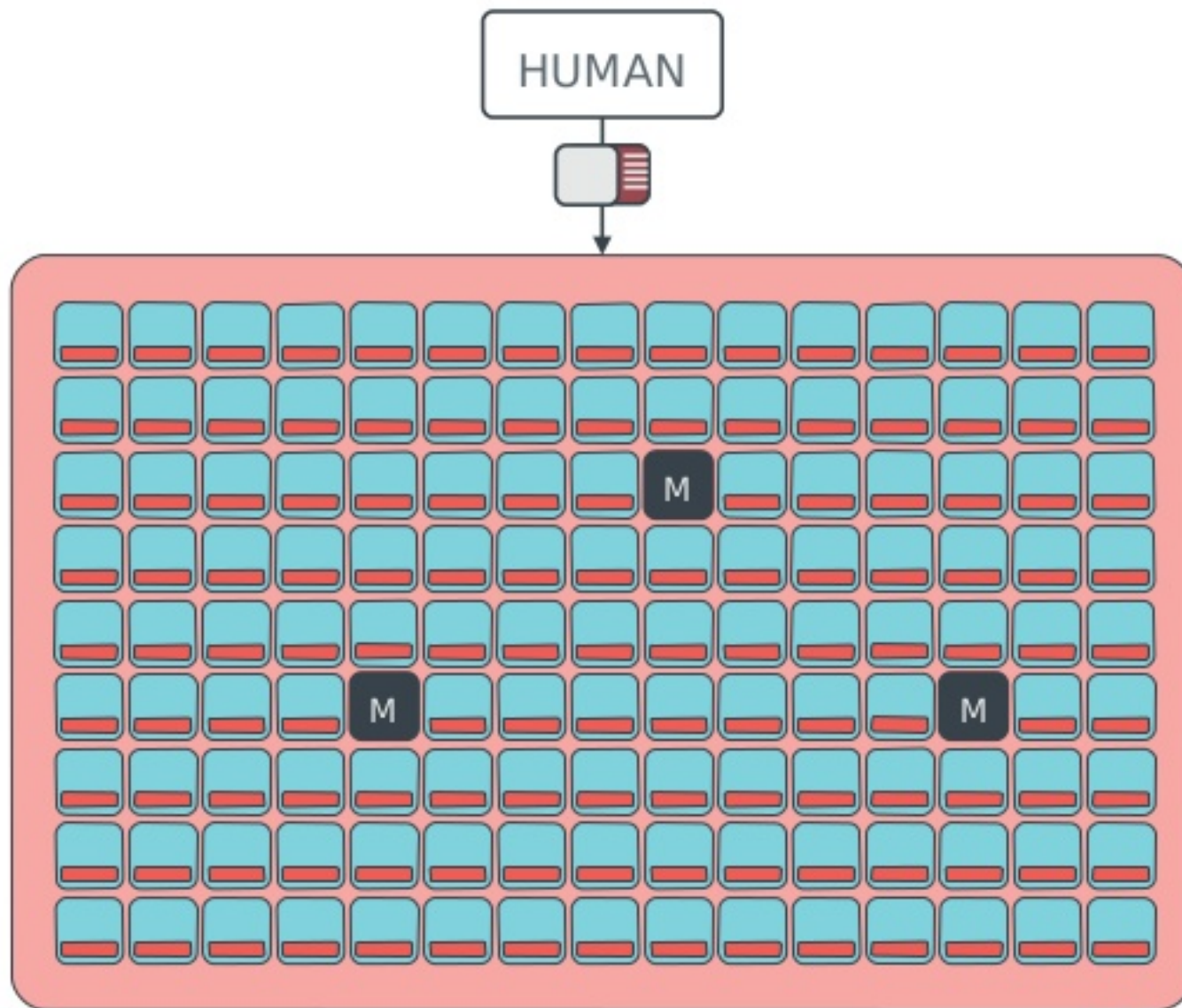




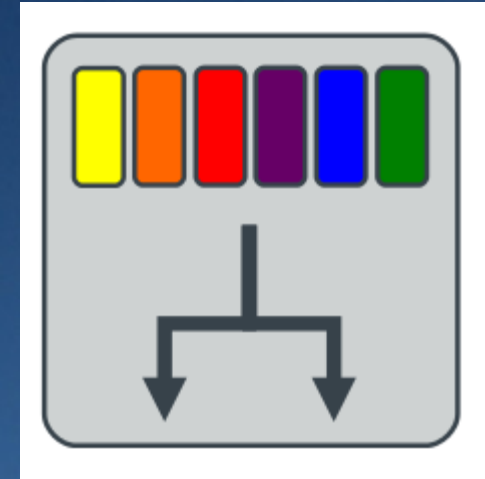


Monitors

- Maintain cluster membership and state
- Provide consensus for distributed
- decision-making
- Small, odd number
- These do not serve stored objects to
- clients

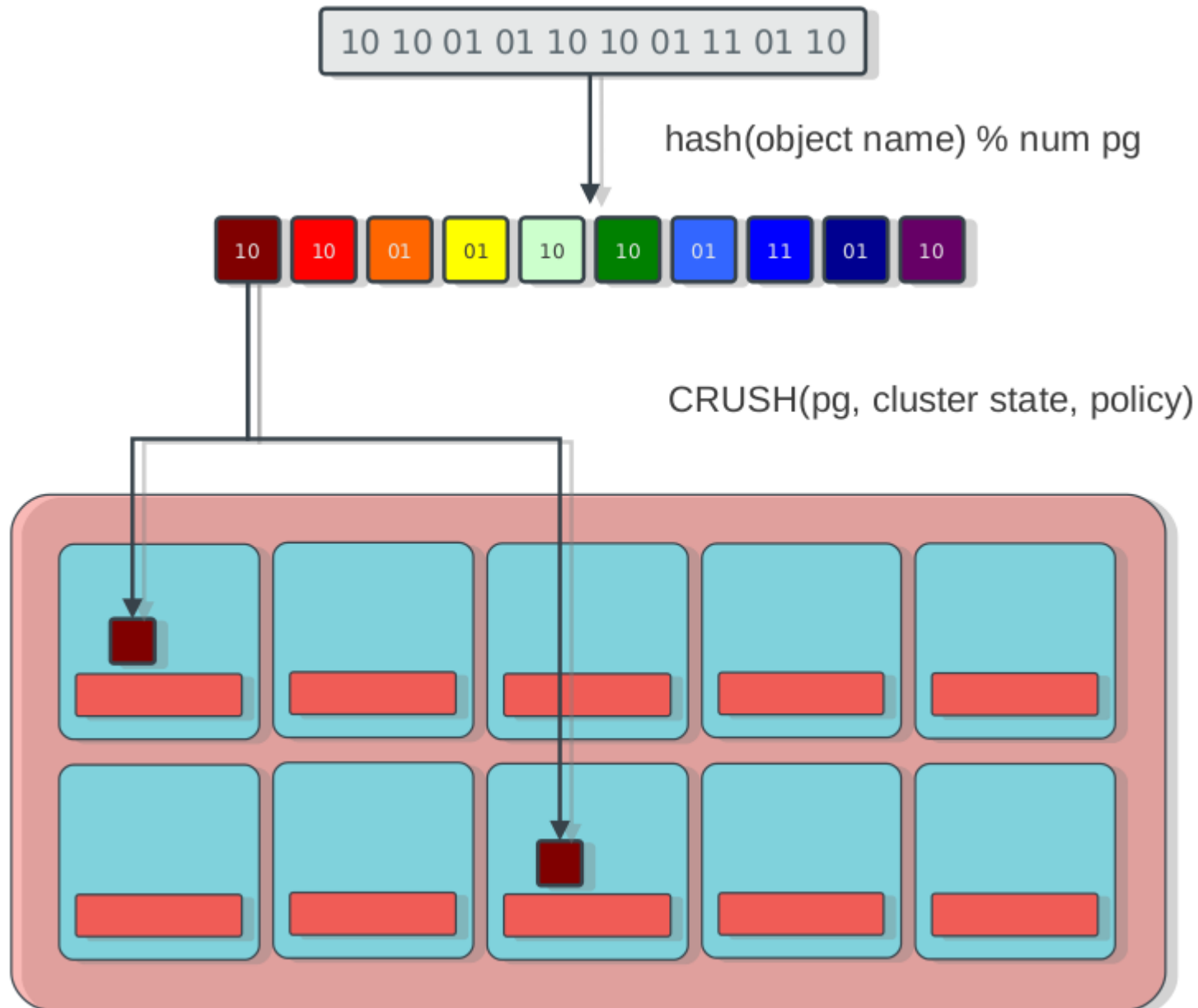


Crush

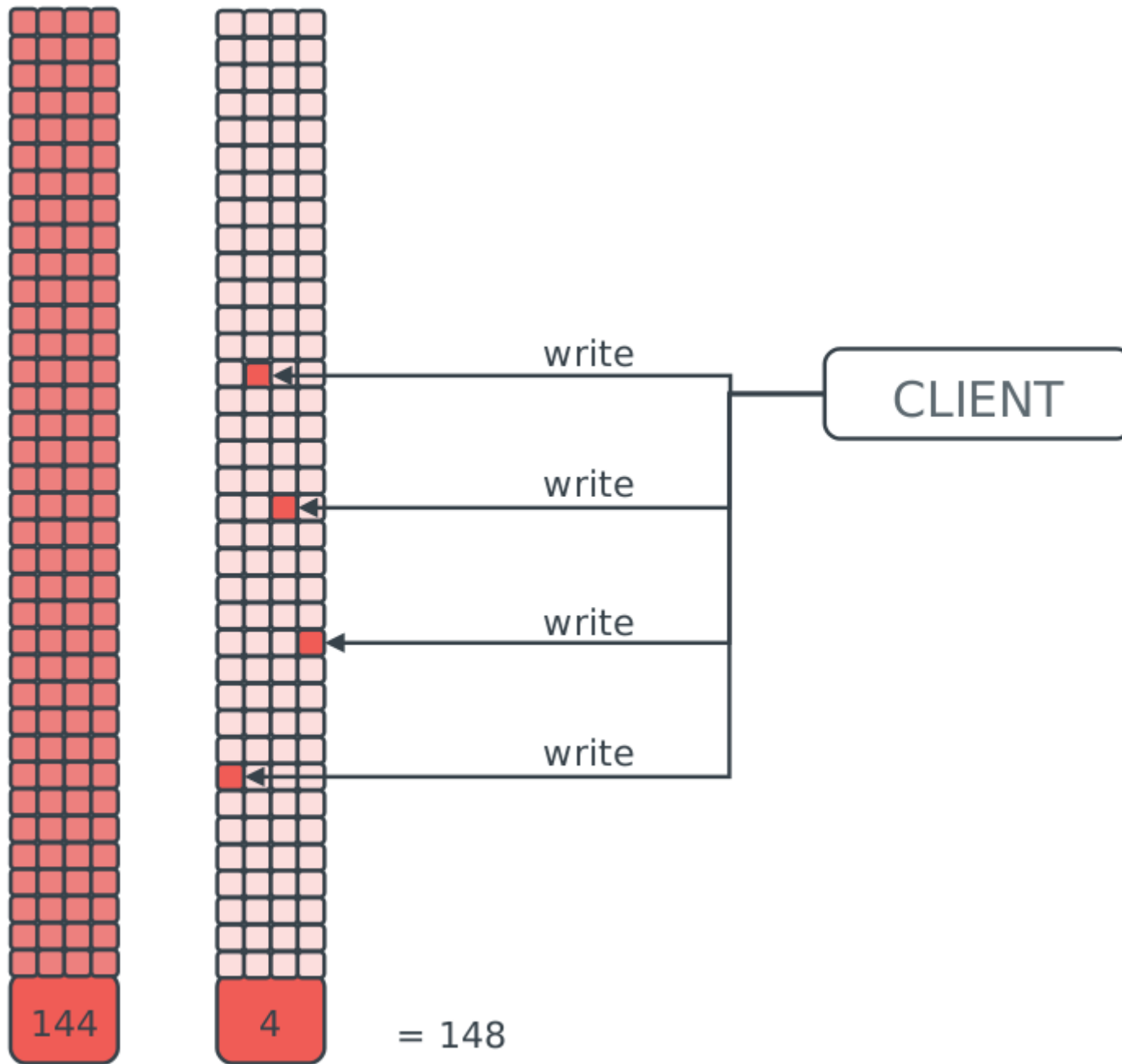


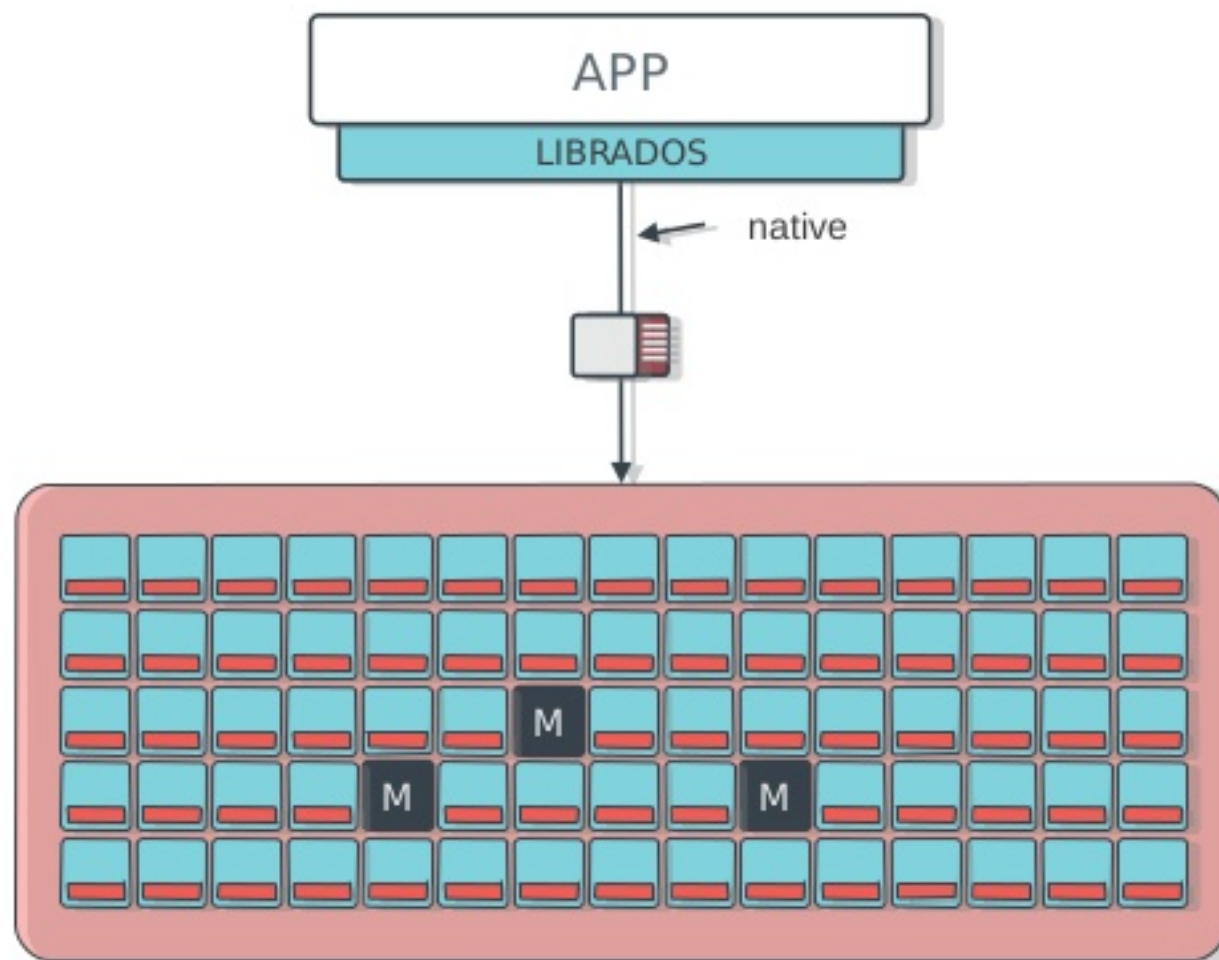
- Pseudo-random placement
- algorithm
 - Fast calculation, no lookup
 - Repeatable, deterministic
 - Ensures even distribution
 - Stable mapping
- Limited data migration
- Rule-based configuration
 - specifiable replication
 - infrastructure topology aware
 - allows weighting

Algorithme crush



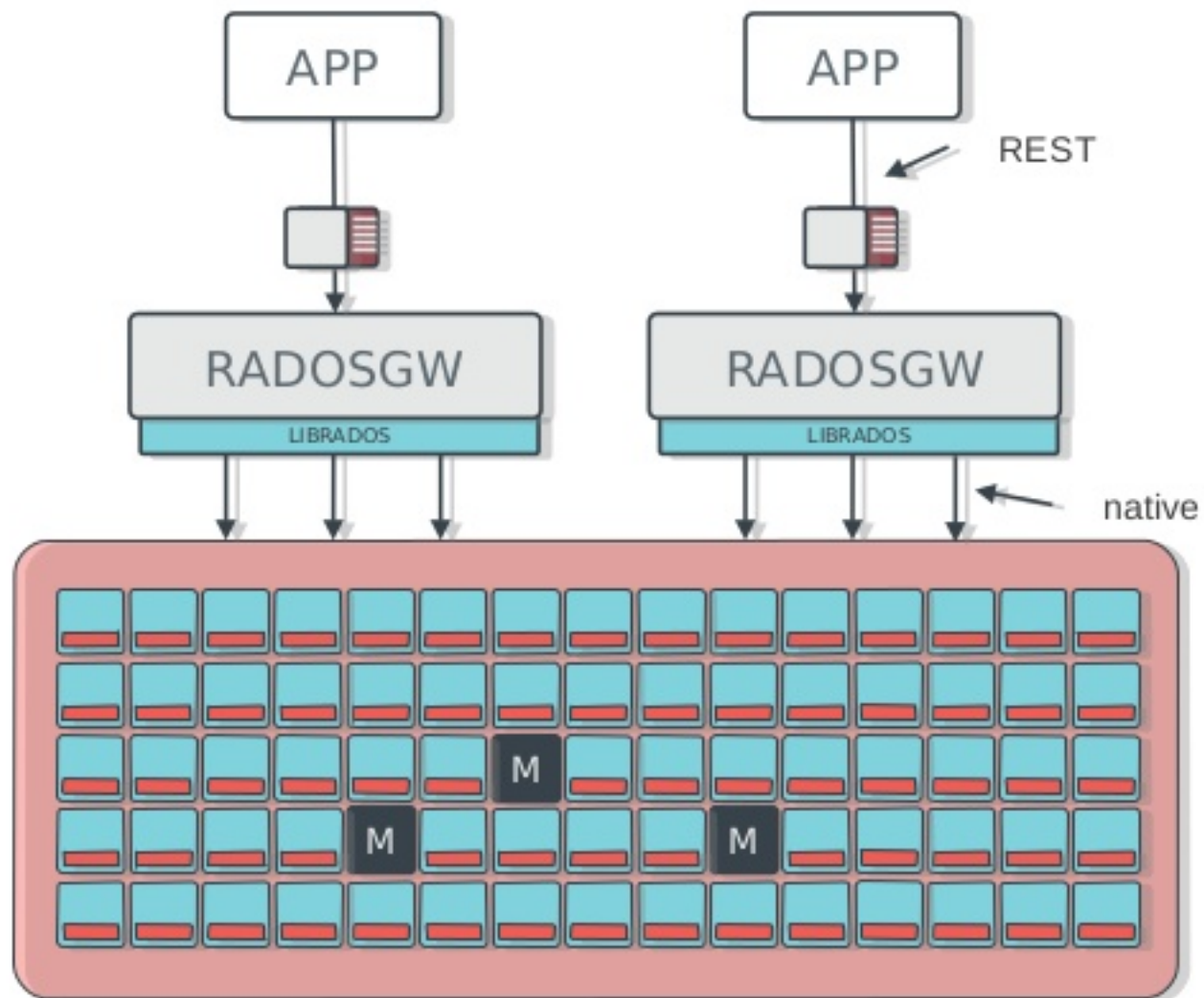
Copy on write





ReSTful Object Storage

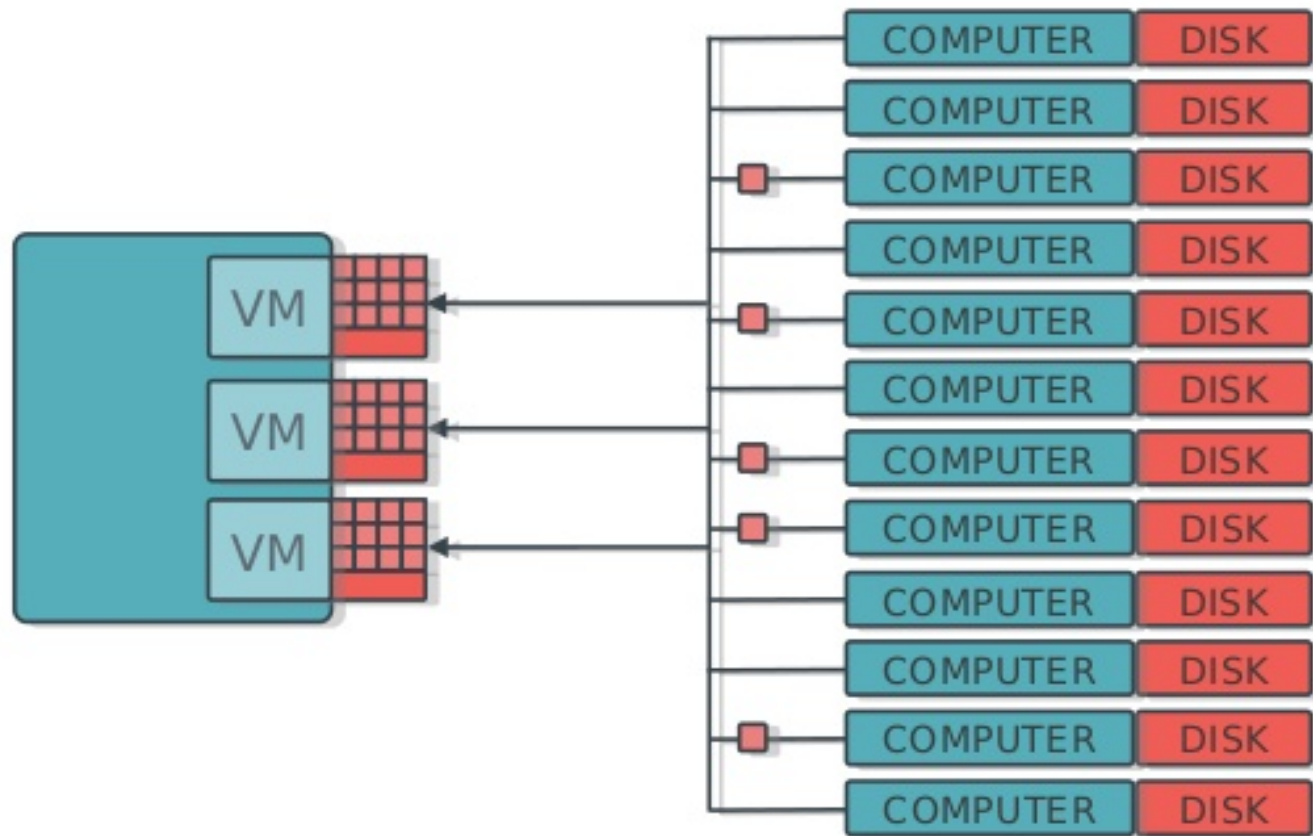
- radosgw : interface pour interagir avec les objets
- HTTP / HTTPS / JSON
- APIs S3 et Swift
- Remplacement du logiciel Swift

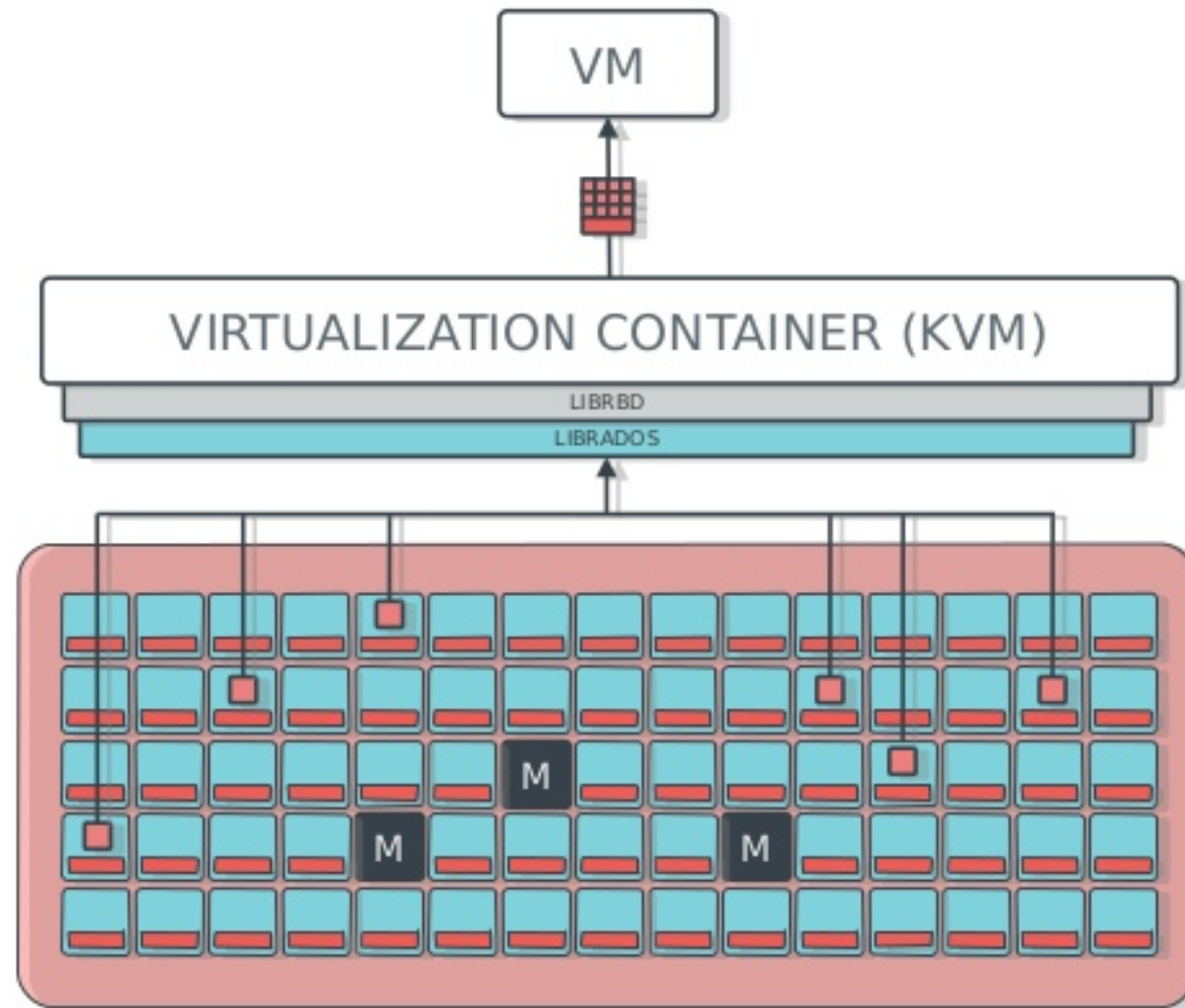


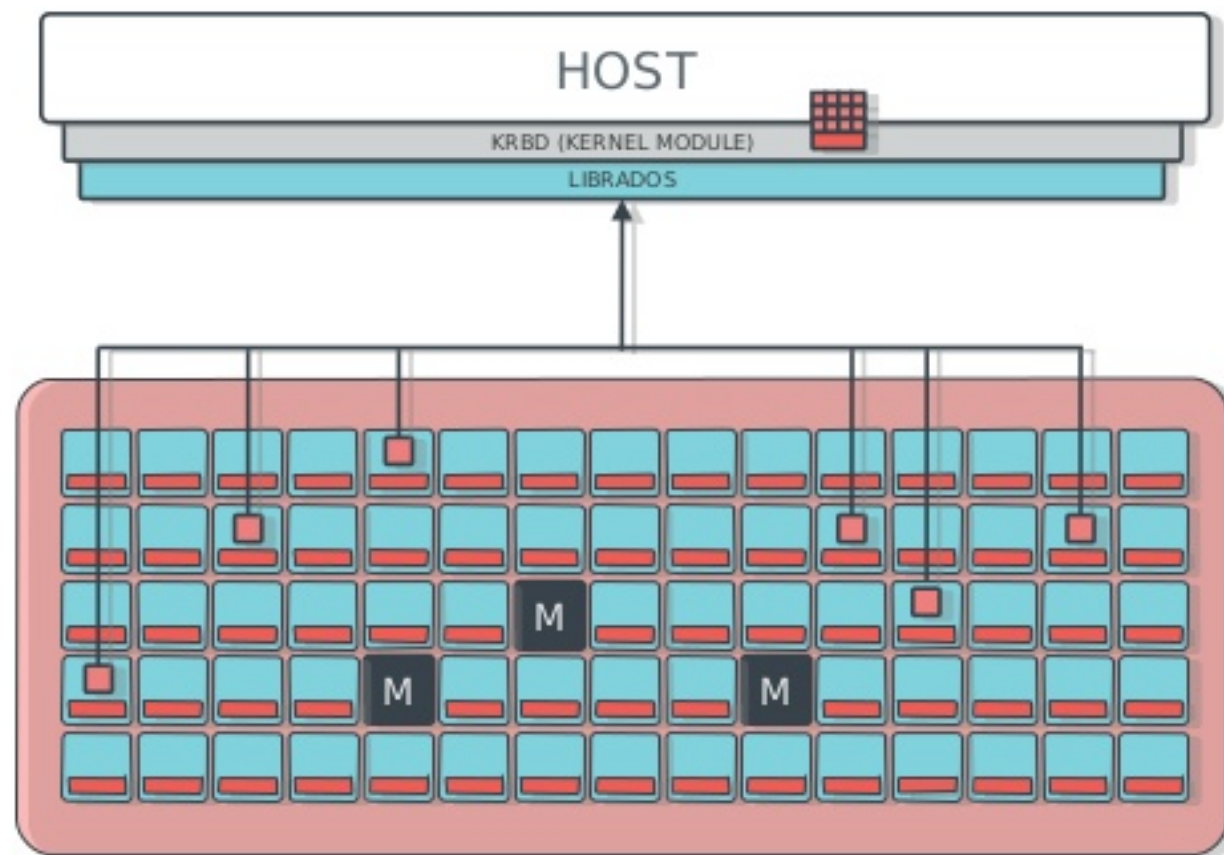
RBD Rados Block Device



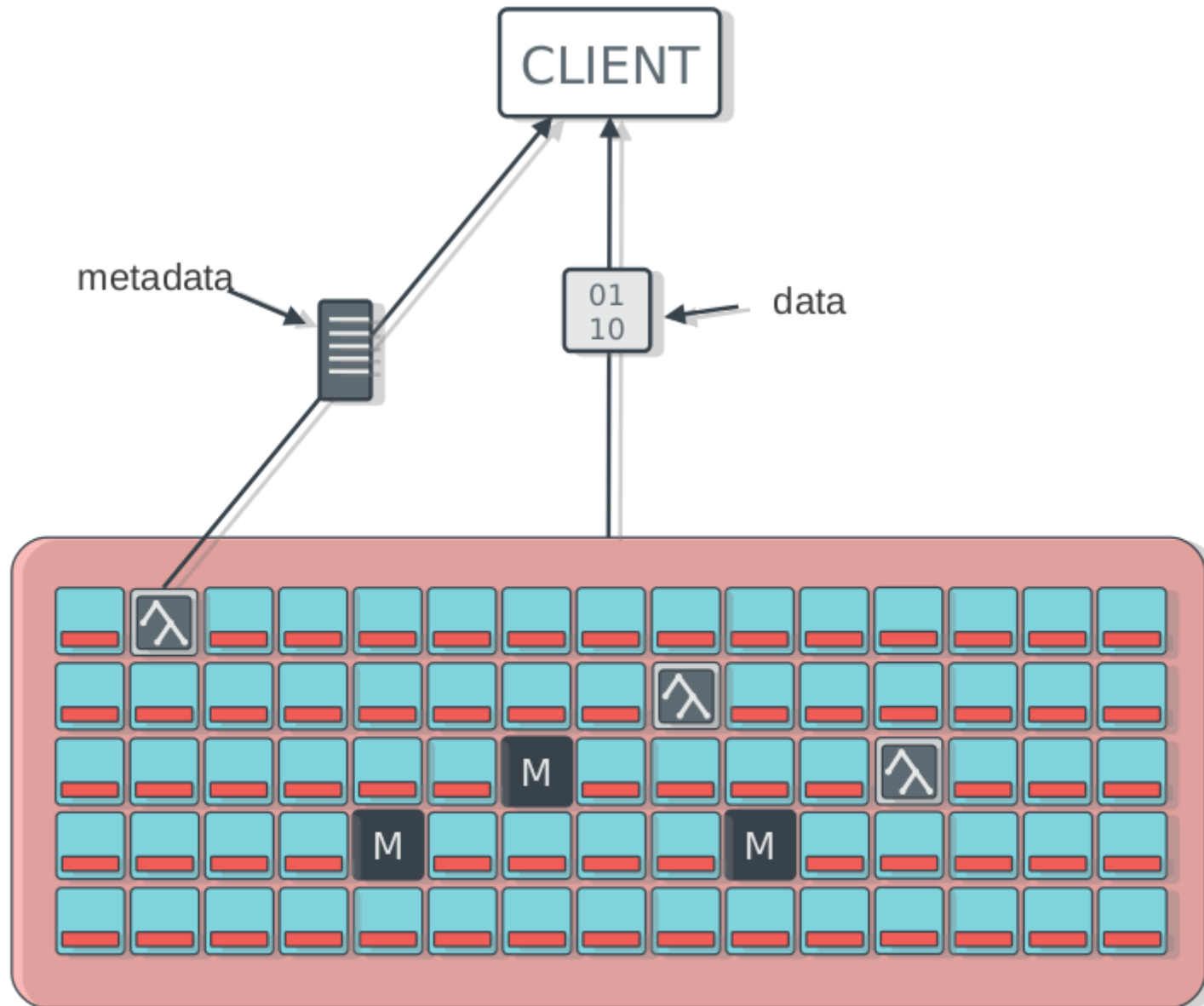
- /dev/rbd0 avec krbd
- /dev/vdb avec librbd dans qemu/kvm
- librbd est ce qu'OpenStack utilise
- comme si on disposait d'un disque



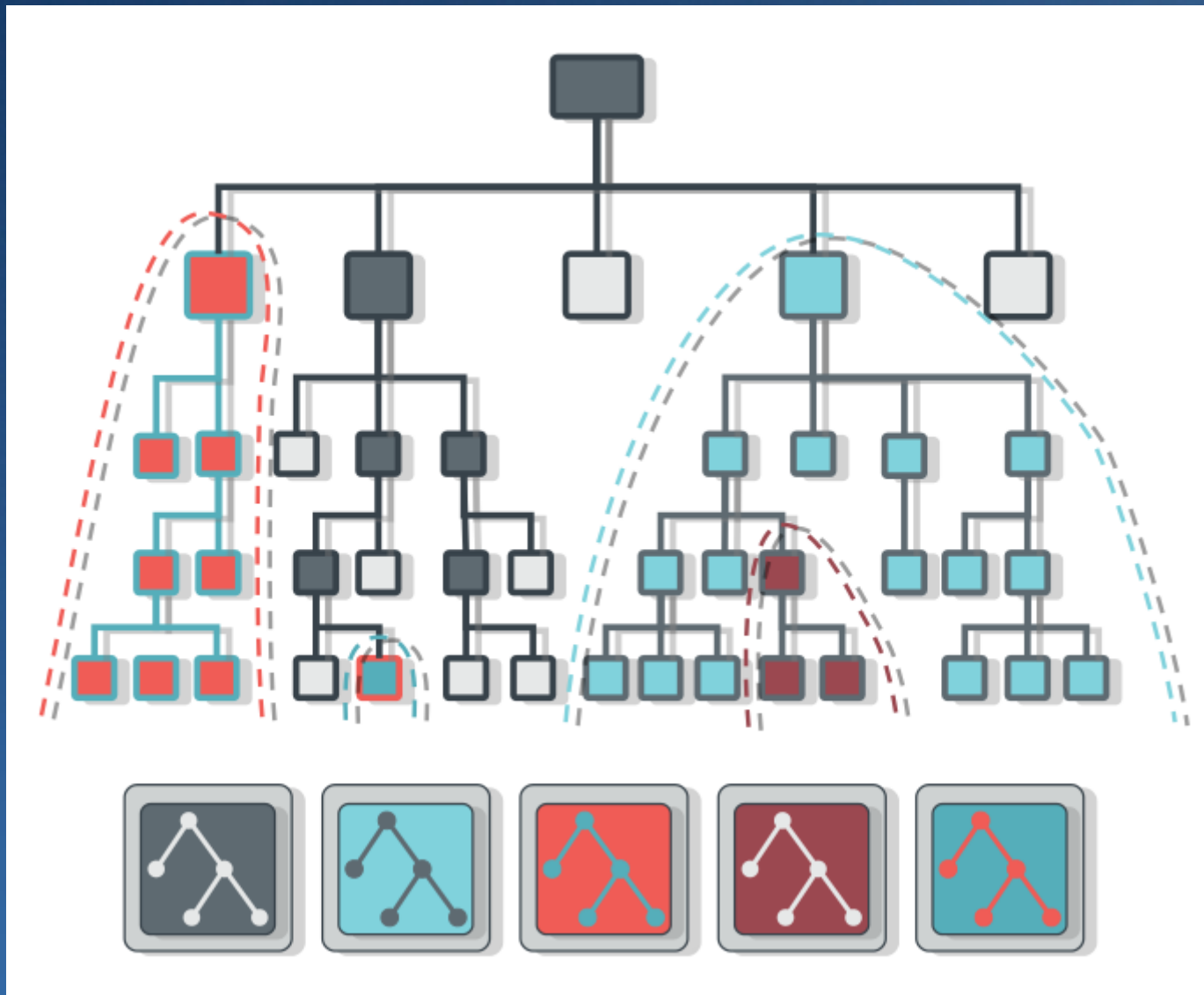




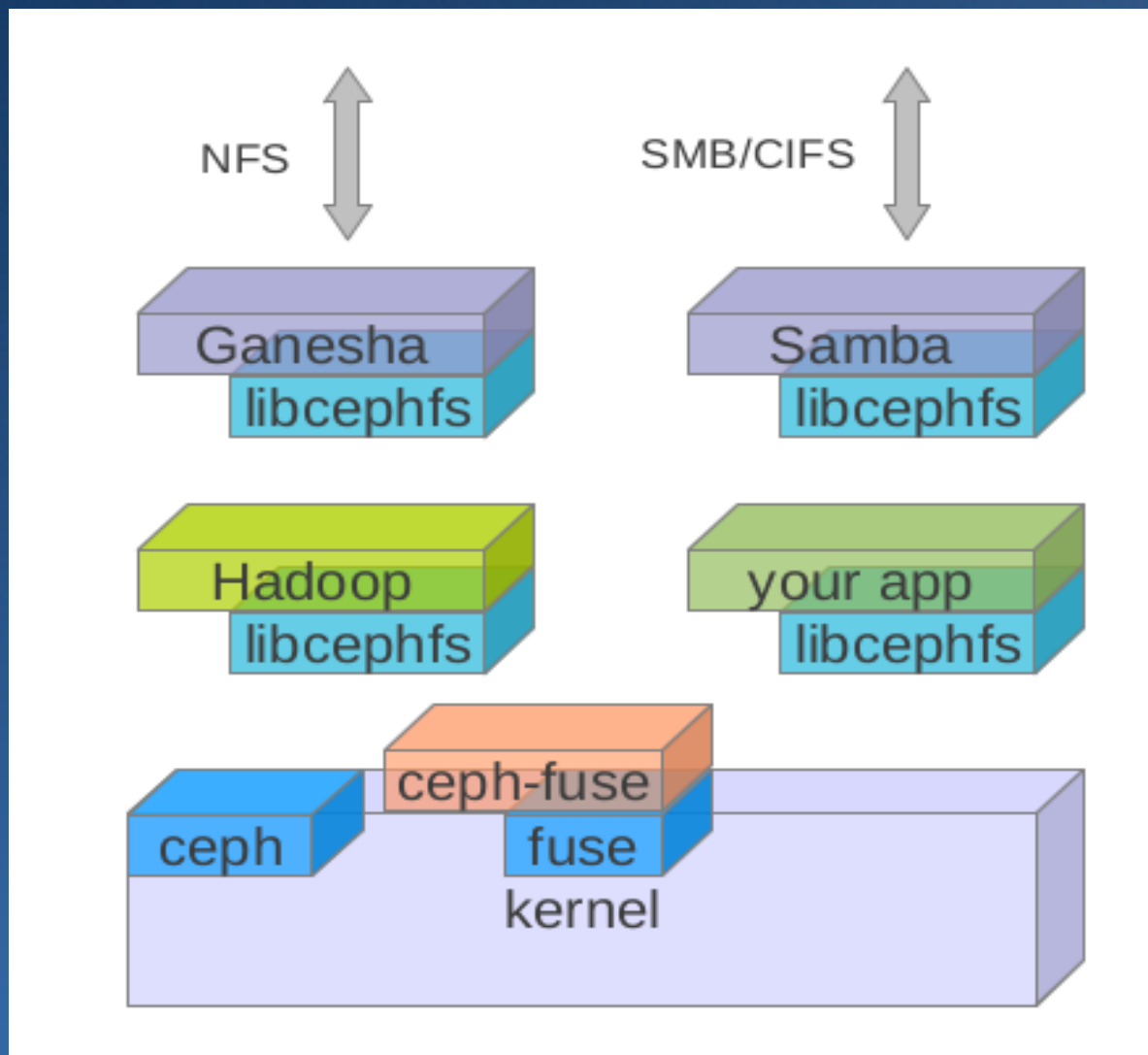
CephFS



Sous partitionnement



Protocols



Traitement près de la donnée

- Plugin installé sur les OSD
- Interface LUA
- API spécifique

Questions ?

