Des trucs sur CEPH

Par Rustace
& Rezyo

Need another one? Here it is!

1. Pourquoi CEPH ?

A part “feur”, “coubeh” et autres réponses stupides

Notre cahier des charges était plutôt simple, il nous fallait différentes propriétés:

- La redondance des données
- La disponibilité des service
- La simplicité d'exploitation
- Un système de backup cool

—Le wiki MiNET

Aspect **01**

Distribué

On peut répartir les données sur plusieurs serveurs

Aspect **02**

Robuste

Pas un point unique de défaillance

Aspect **03**

Extensible

On peut facilement augmenter la quantité de stockage

Aspect **04**

Compatible

Pas eu besoin de changer toute notre infra de l'époque

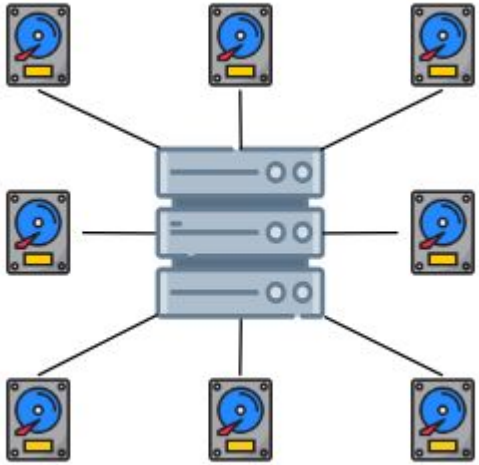
C pa fassil

Whoa!

C’est cool tout ça, il y a pas de défaut ?

2. Banalités

Les trucs de base



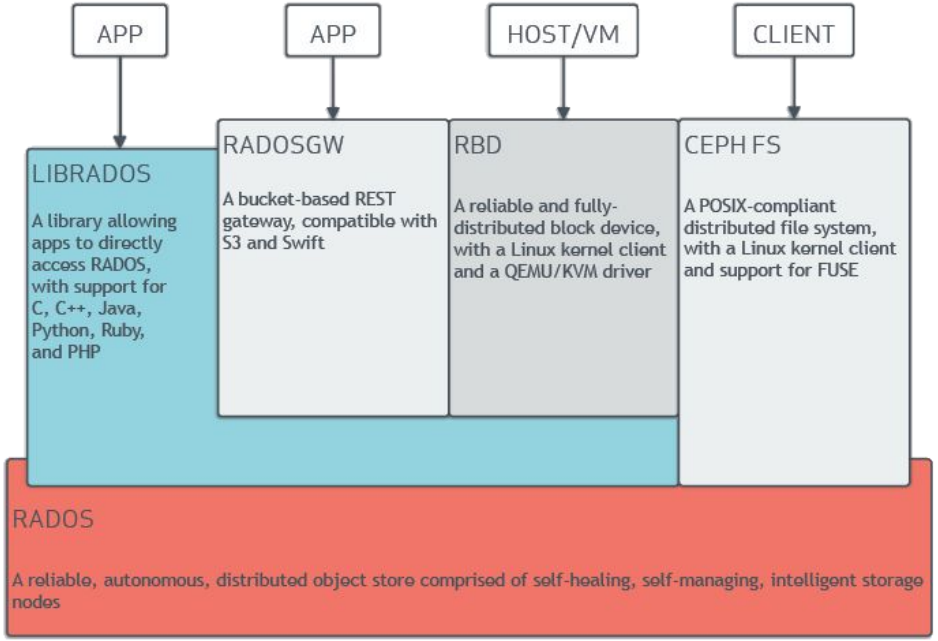
Les storage pools

1 Pool = 1 Noeud

Le backend RADOS

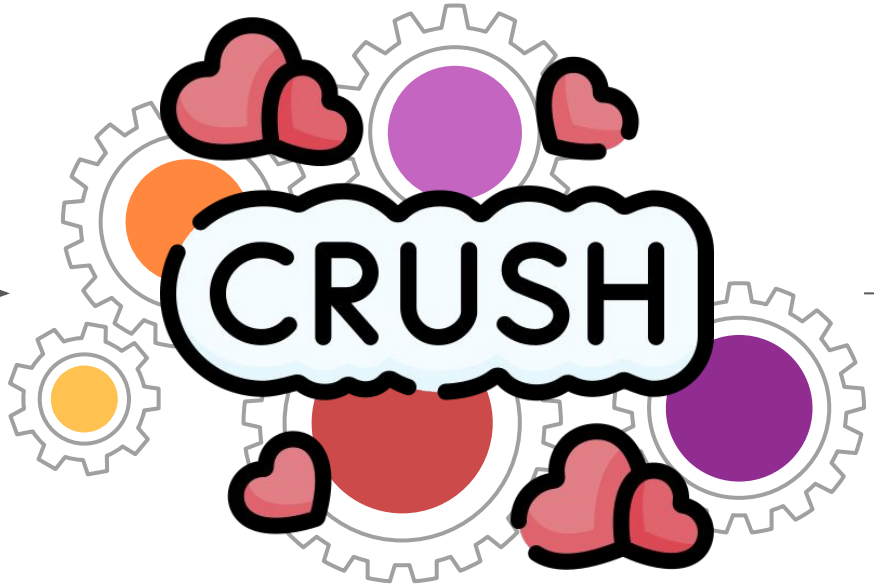
C'est le gestionnaire de stockage

```
#include <rados/librados.h>
```

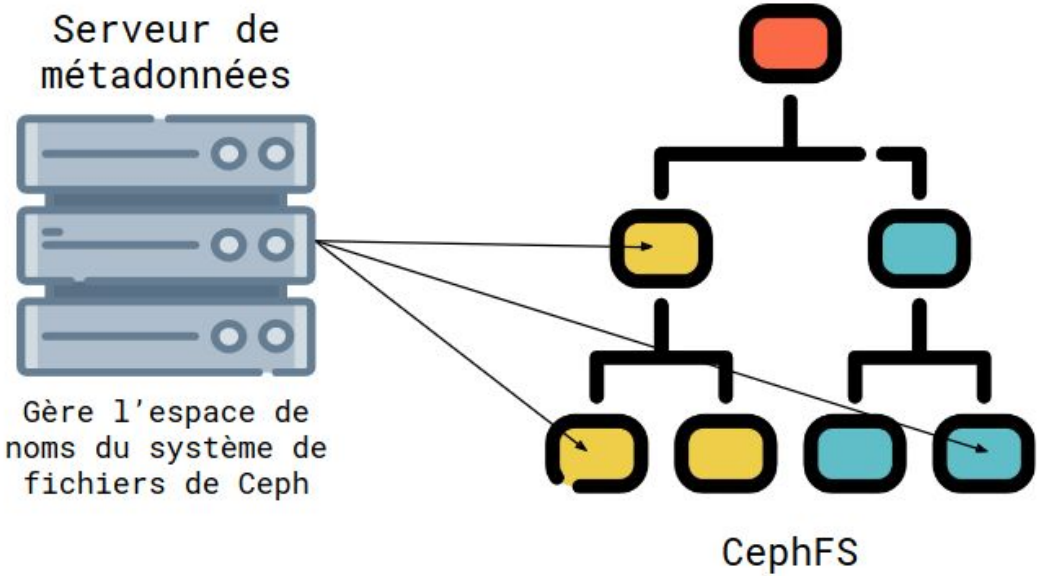


L’algorithme CRUSH

Crush map,
Chemin de fichier



ID de fichier,
Emplacement


Serveur de métadonnées
Gère l'espace de noms du système de fichiers de Ceph
CephFS

MDS

3. Architecture

Comment ça marche tout ce
petit monde ?

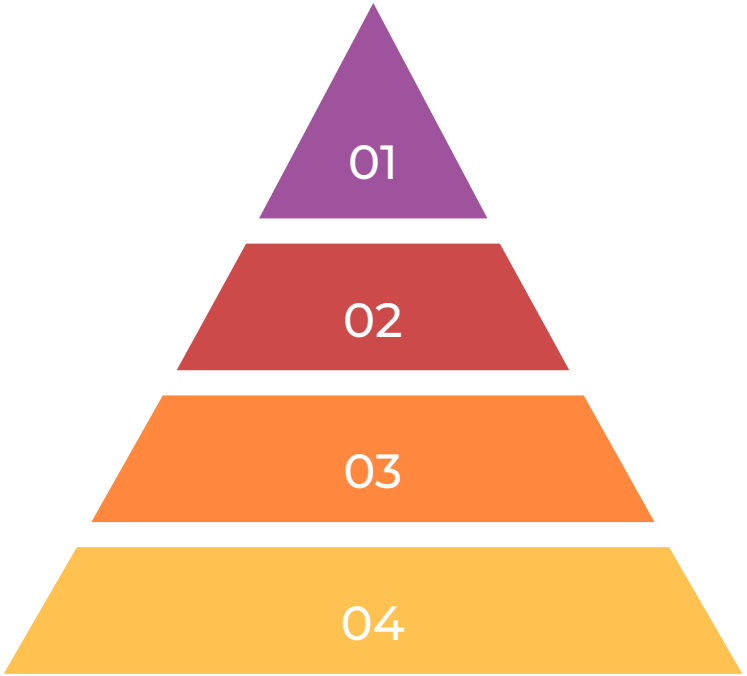
Les différents acteurs de CEPH

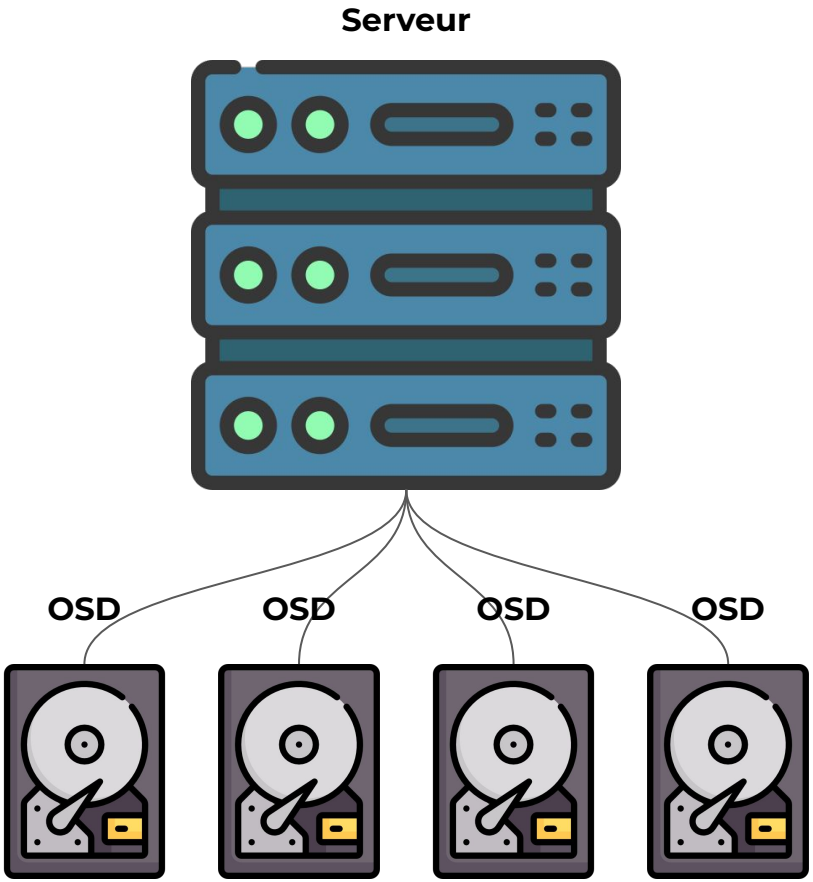
01 Manager
Connaît l'état du système

02 Moniteur
Maintient une carte du système

03 BlueStore
Les MDS

04 OSD
Le véritable stockage

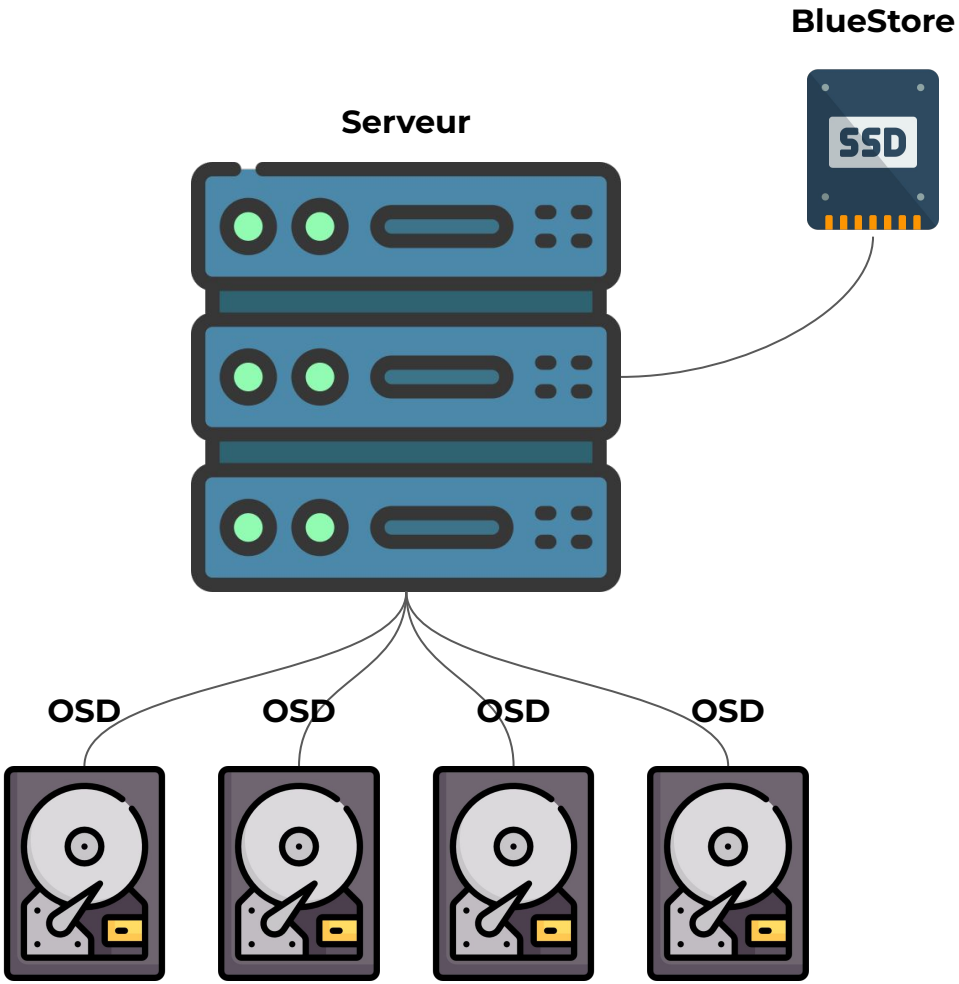




Les OSDs pour le stockage

Pour Object Storage Daemon

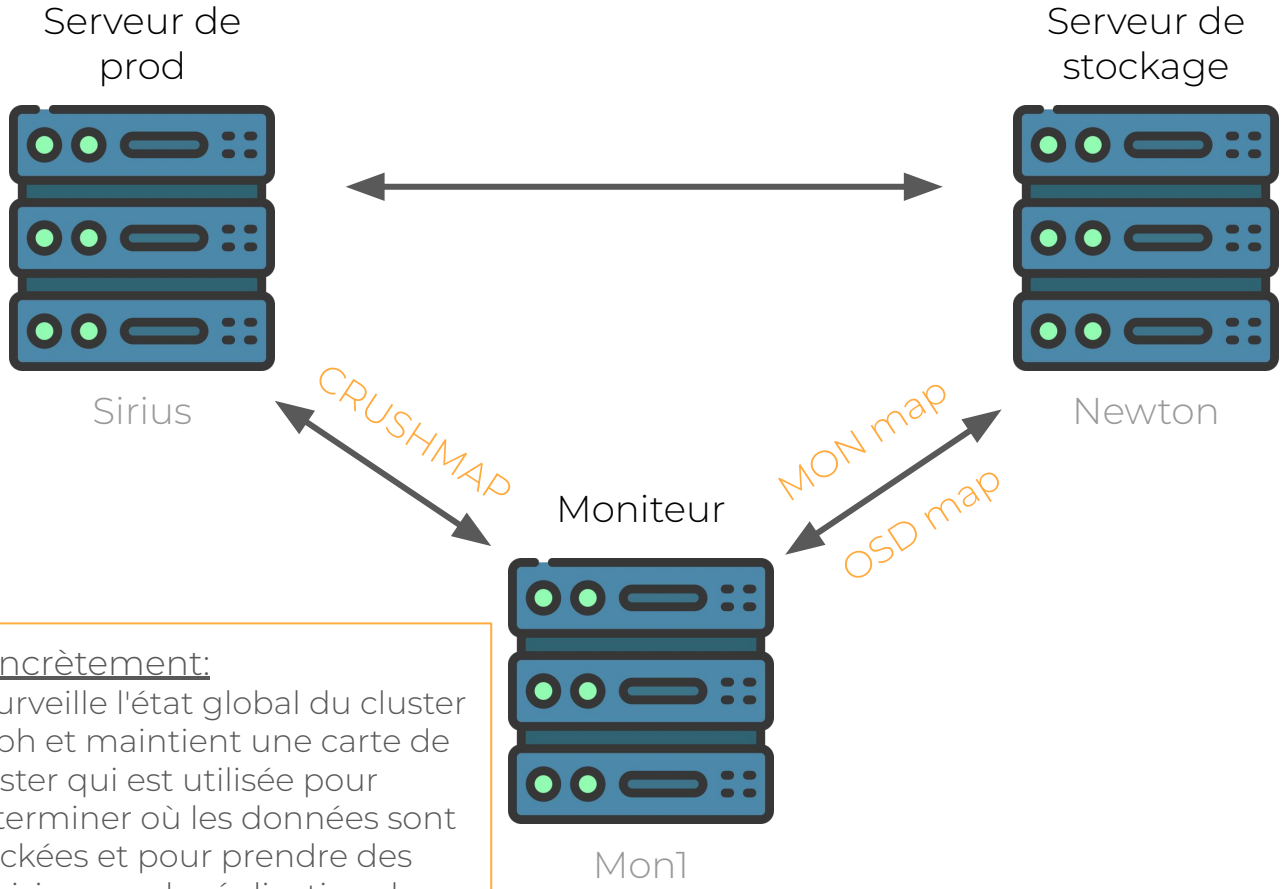
- S'occupe du stockage des données
- Assure la redondance
- 1 daemon par disque



Les BlueStores pour les métadonnées

Les fichiers sont stockés en brut sur les OSDs, les BlueStores contiennent les informations sur ces fichiers

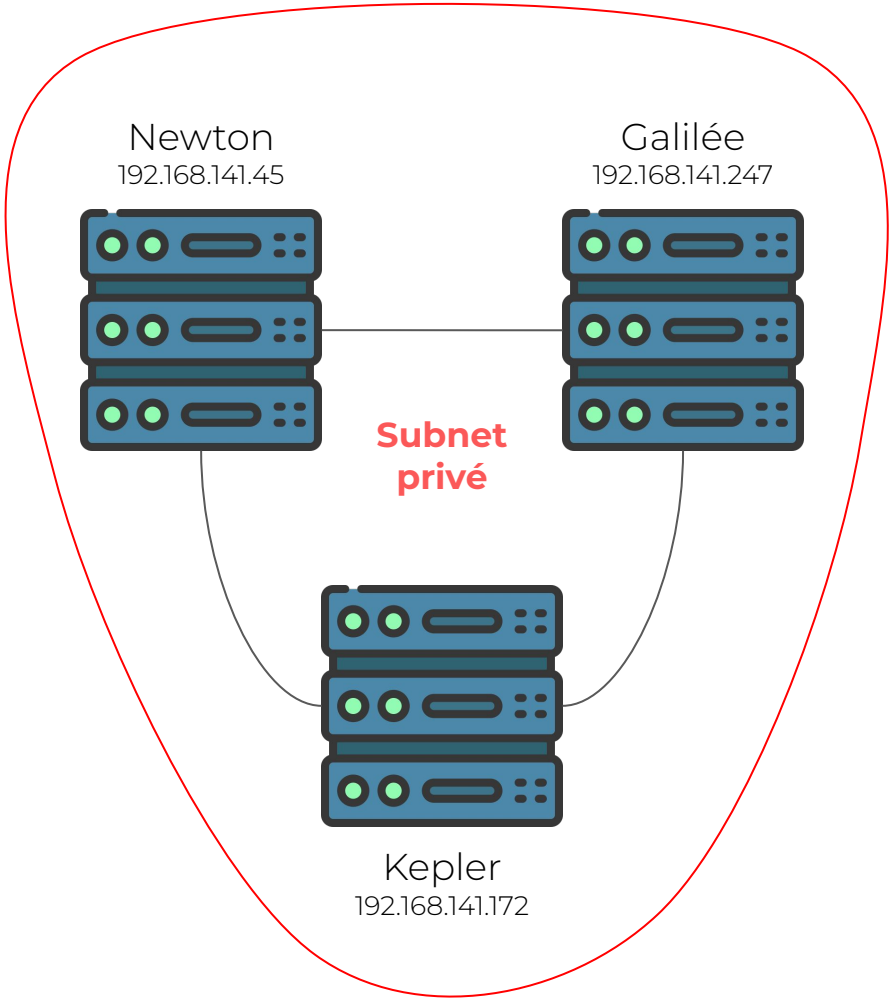
S'apparente (grossièrement) à une table d'association de blocs mémoire



Les moniteurs

Où cé kel son lait donné?

Concrètement:
Il surveille l'état global du cluster Ceph et maintient une carte de cluster qui est utilisée pour déterminer où les données sont stockées et pour prendre des décisions sur la réplication des données



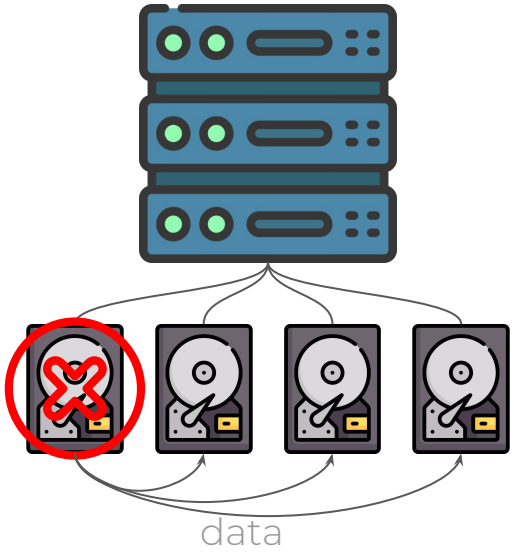
Les Managers

- Génèrent l'OSD map
- Gère les OSDs
- Rend transparent le cluster depuis tous les noeuds

La réplication des données

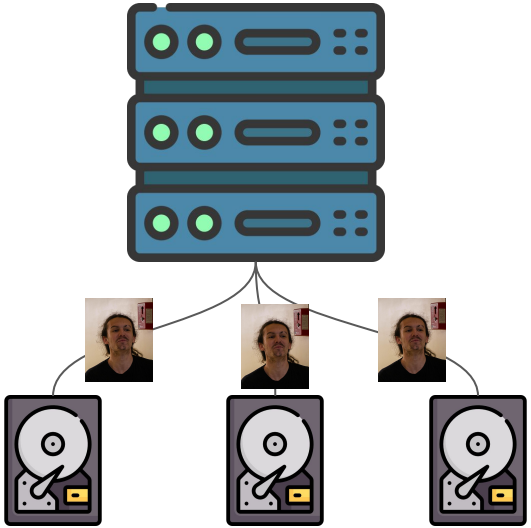
100% autonome:

Re-réplique les données dès qu'un des OSD est down

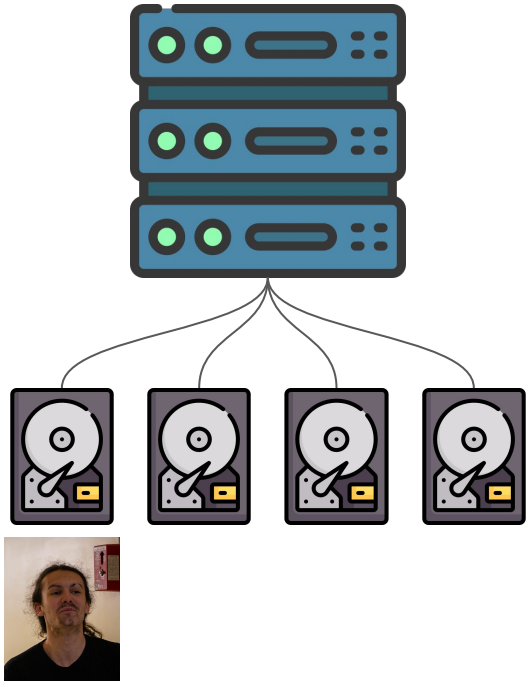


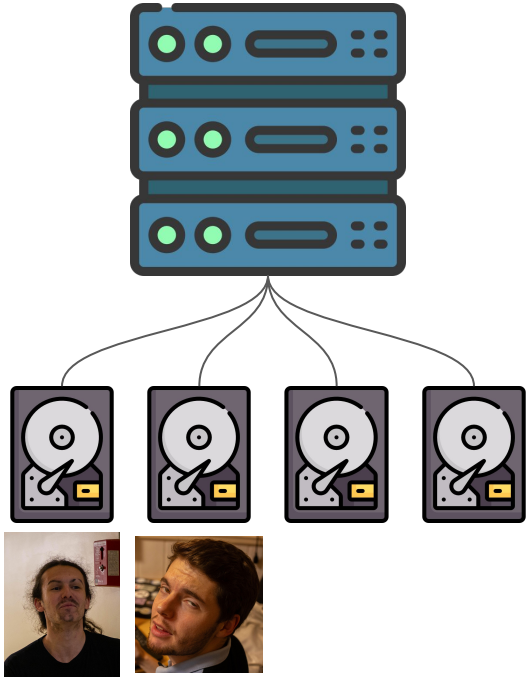
1 réplica/noeud:

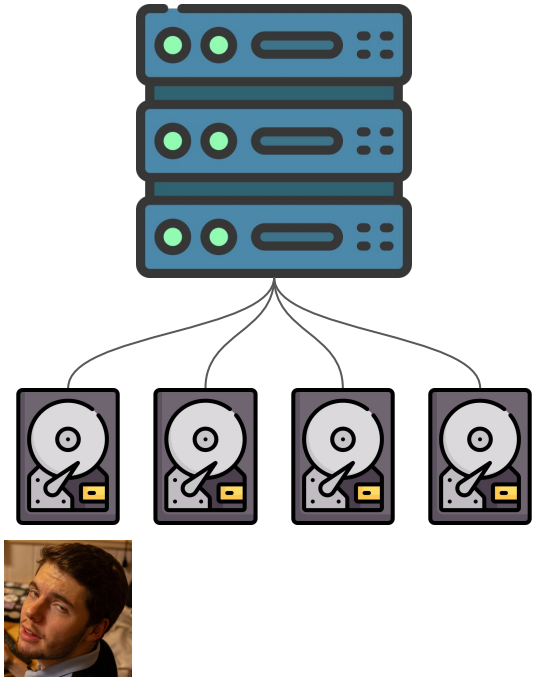
Chaque donnée répliquée ne peut jamais être 2 fois sur le même noeud



Réplication ×2

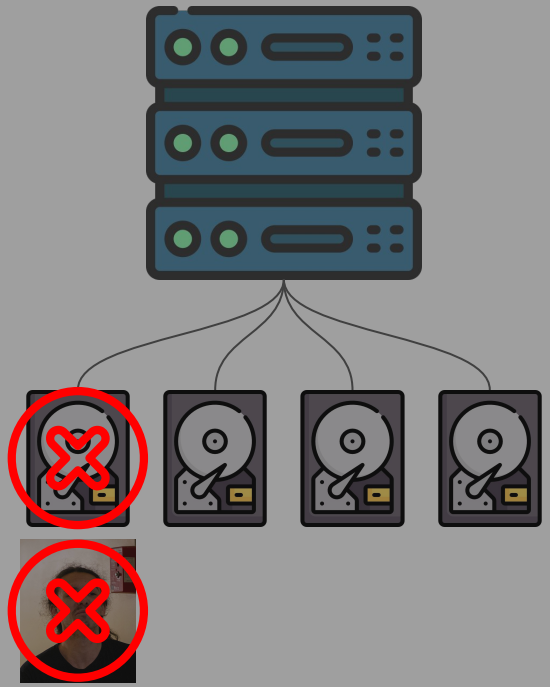


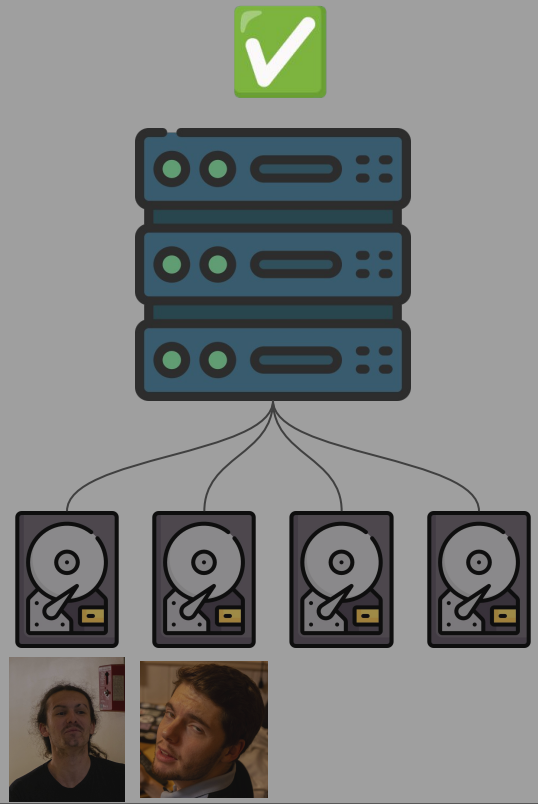


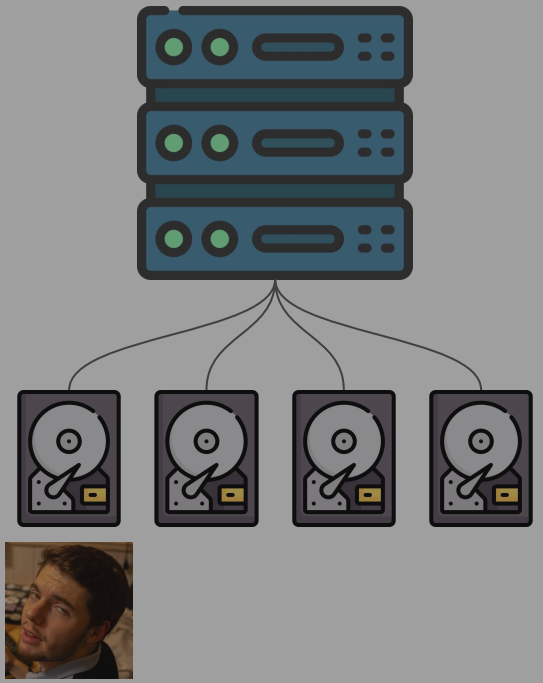


Réplication ×2

Perdre 1 OSD



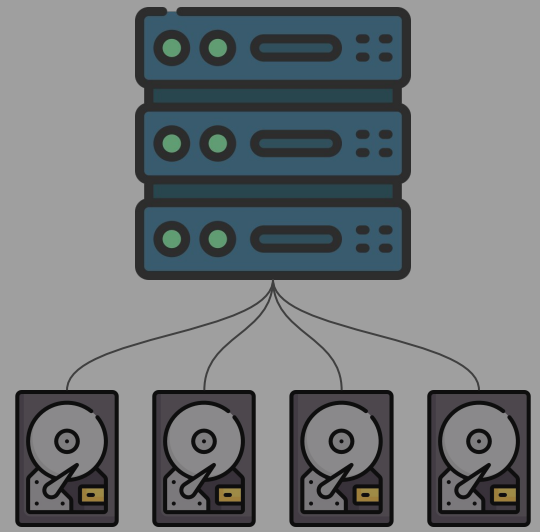


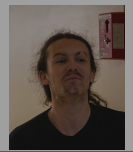


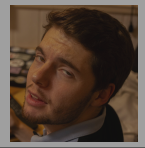
Réplication ×2

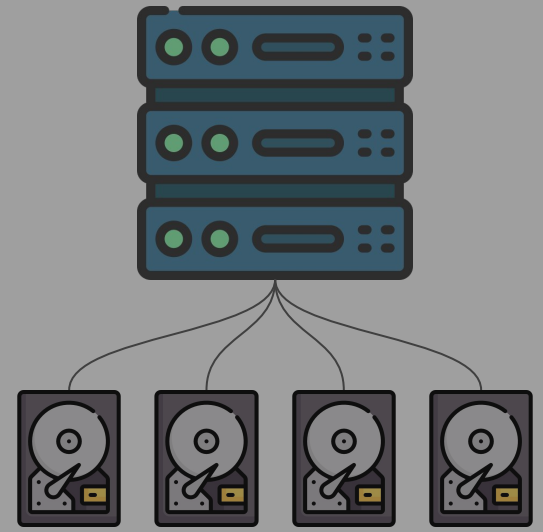
Perdre 1 Serveur

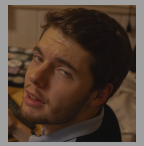
✅





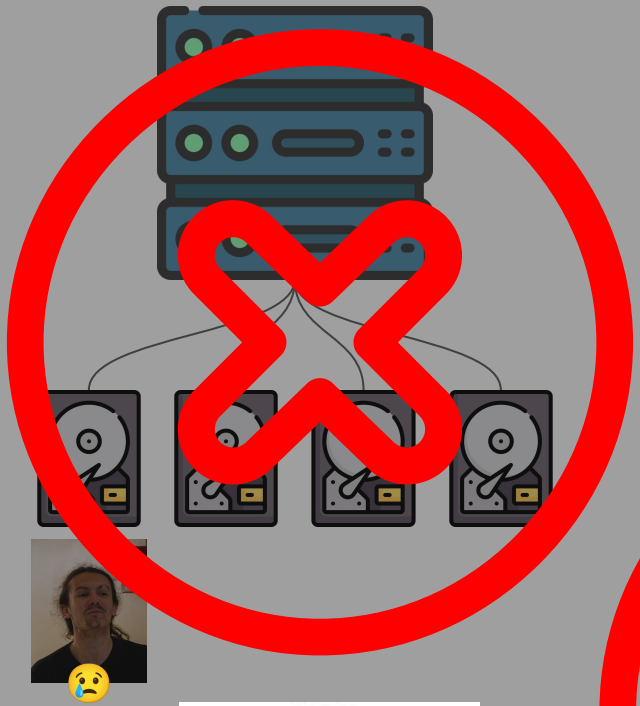


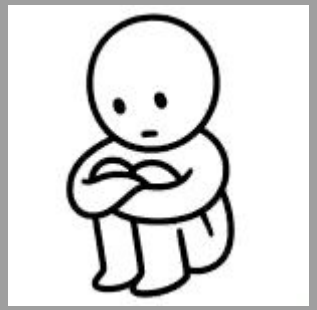


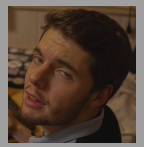


Réplication ×2

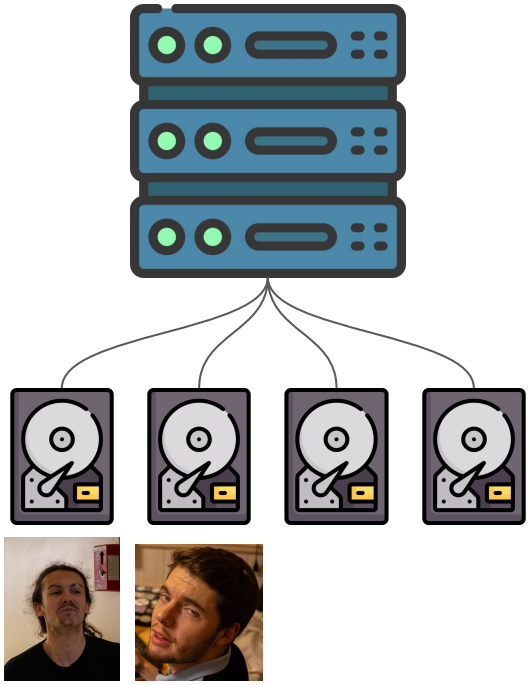
APOCALYPSES
Perdre 2 Serveur

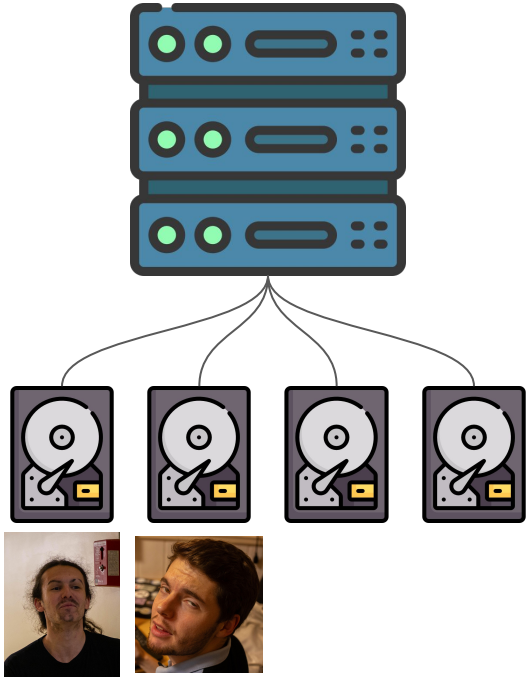


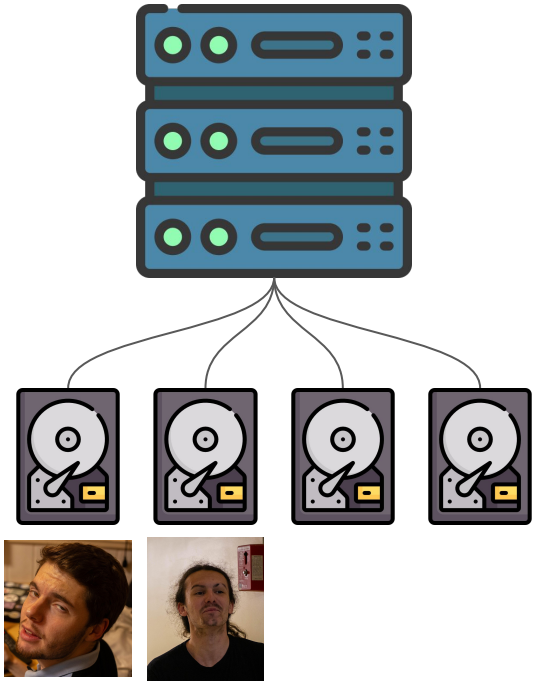




Réplication ×3

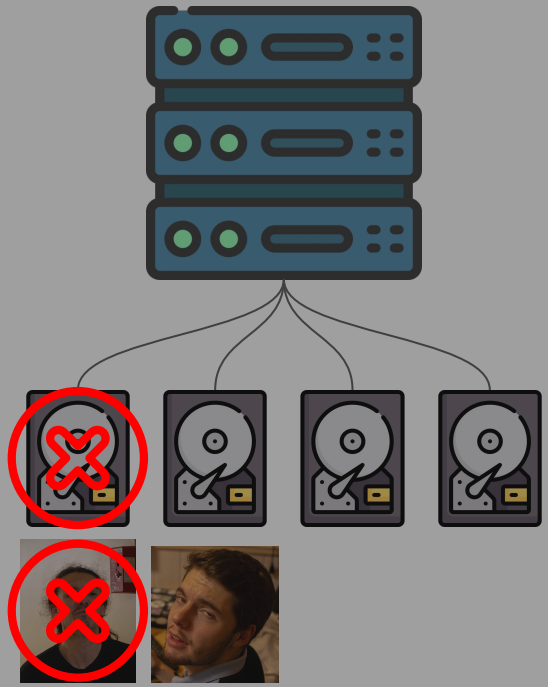




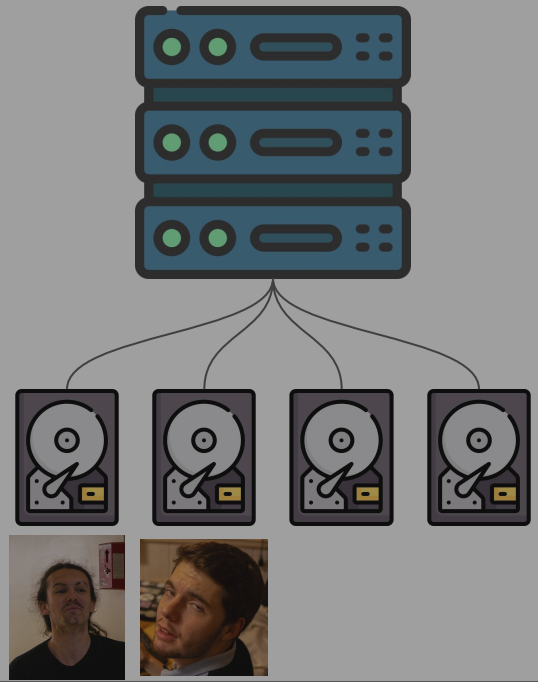


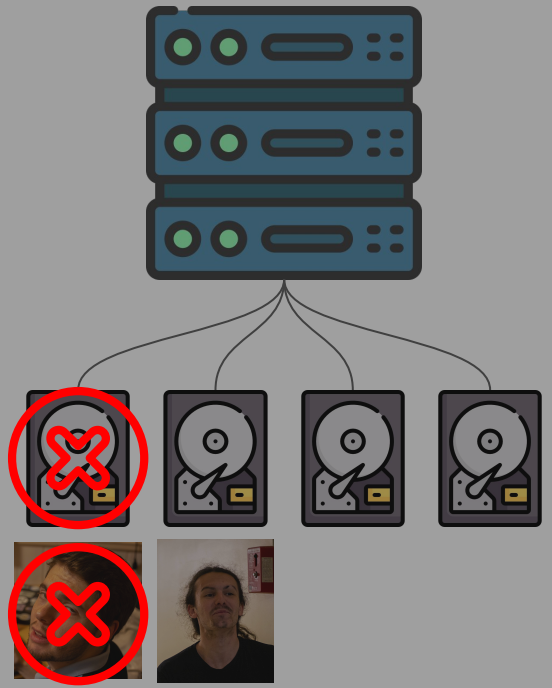
Réplication ×3

Perdre 2 OSDs



✅

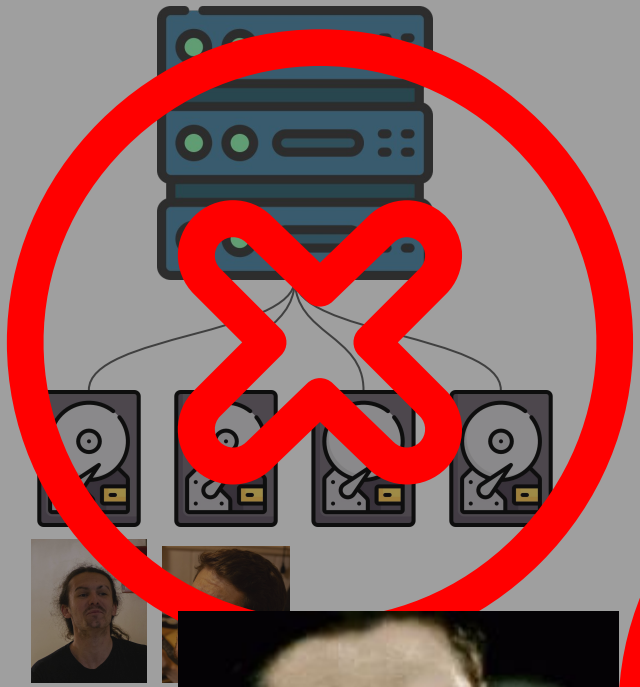




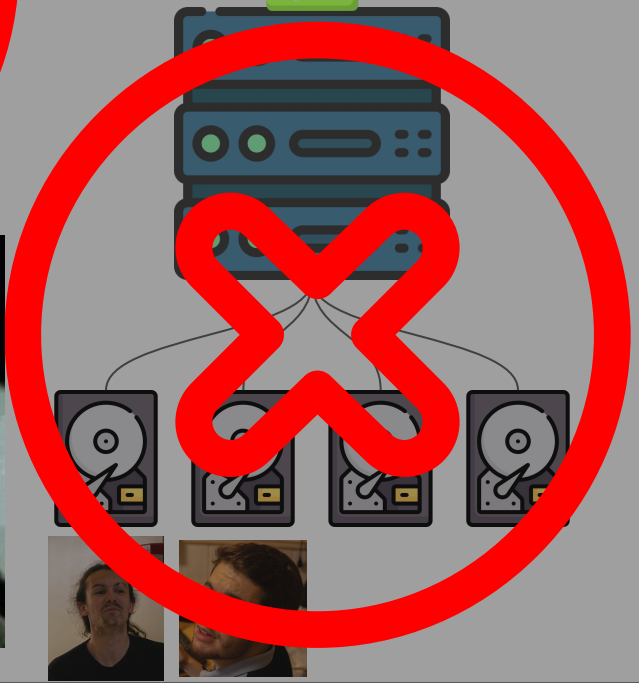
Réplication ×3

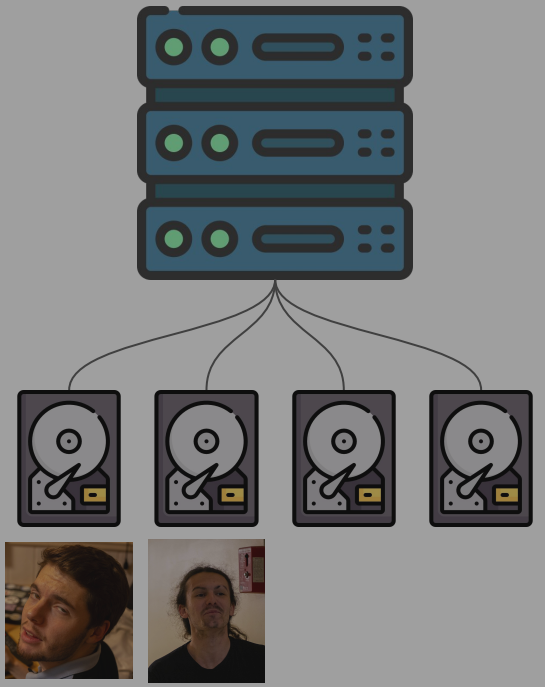
APOCALYPSES

Perdre 2 serveurs



✅





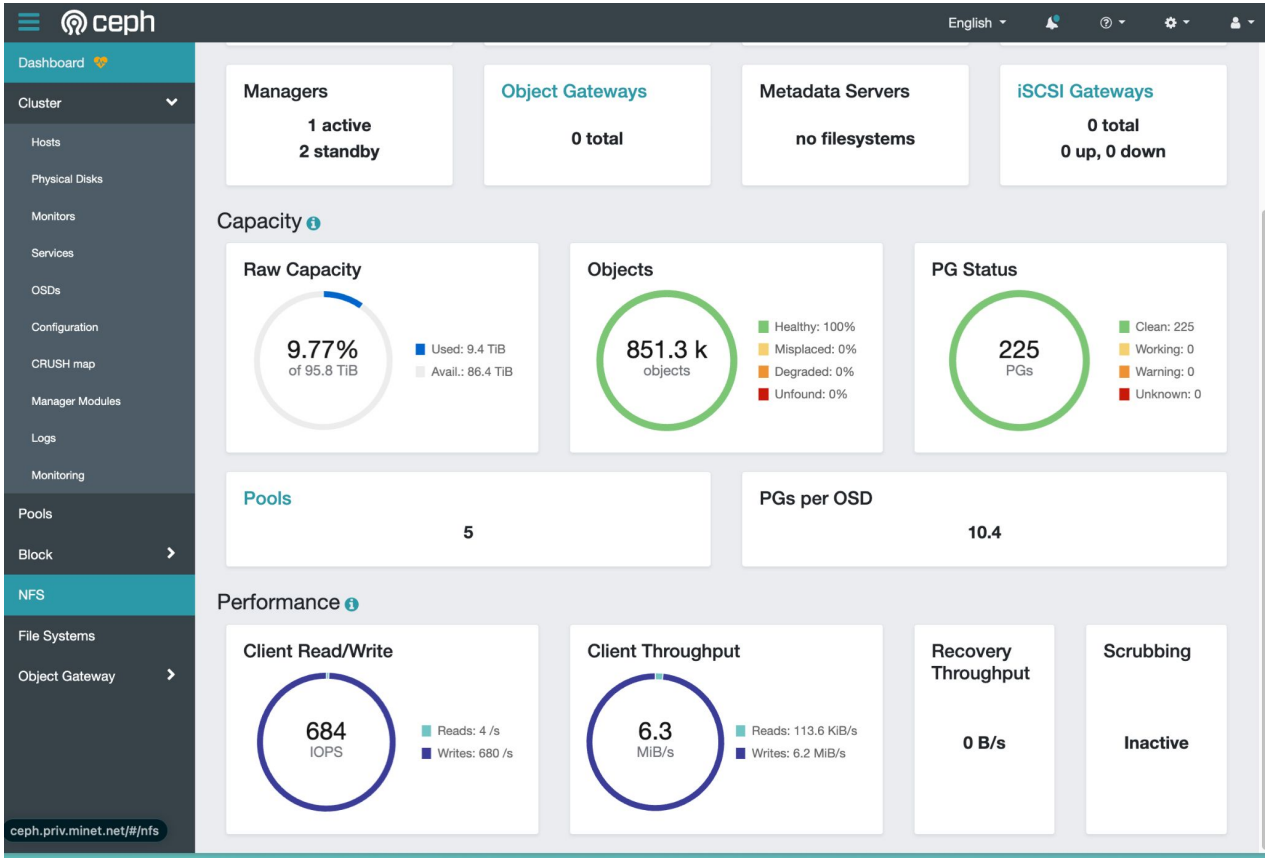
4. Monitoring et troubleshoot

Maman j’ai cassé le stockage

Va sur le wiki petit chenapan

https://wiki.minet.net/fr/cluster/ceph

1. Monitoring avec CEPH

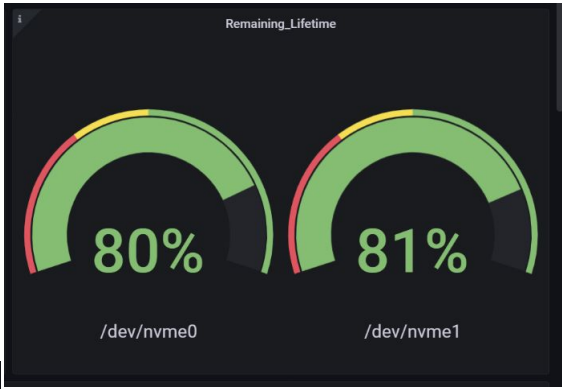


CEPH dispose d'un module très utile permettant de voir beaucoup d'informations sur l'état depuis internet.

Au jour où je fait cette formation, le dashboard CEPH n'es pas accessible à cause de bugs suite au passage à debian 12 =(

2. Monitoring avec Grafana

⚠️ Il faut bien faire attention à l'état des SSD, sans eux tout CEPH est perdu et ils vieillissent très vite.



Quelques commandes utiles:

```
ceph status
```

Permet de voir le status du cluster. (la commande la plus utile de toutes)

```
ceph health detail
```

Permet de voir le détail, c'est cette commande qui permet d'orienter vos recherches quand quelquechose ne va pas

```
ceph -w
```

Permet de voir le status + les événement arrivant sur le cluster

```
ceph osd tree
```

Pour voir tous les OSDs

```
ceph mon stat
```

Permet de voir le cluster des mons

```
ceph osd lspools
```

Pour voir toutes les pools

```
rbd ls nom_d'une_pool
```

Pour voir toutes les block devices d'une pool

```
rbd snap ls nom_d'une_pool/nom_d'un_block_device
```

A savoir:

- Ne SURTOUT jamais débrancher un disque qui est encore visible sur la machine
- Suivre la page dédiée -> https://wiki.minet.net/cluster/ceph/remplacement_disque

MERCI DE NOUS AVOIR ÉCOUTÉ