SCALAR

Infrastructures Avancées

Stockage distribué avec Ceph

Objectifs

Comprendre la problématique moderne du stockage et la nécessité de Ceph
Comprendre son architecture générale
Comprendre en détails le rôle de CRUSH, des Pools, des PG
Savoir distinguer les 3 services de stockages possibles
Connaître les bonnes pratiques dans l'utilisation de Ceph, et son administration quotidienne

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

Pourquoi Ceph?

Pourquoi on parle de Ceph ?

Ceph résout un problème très simple : **le stockage doit continuer de tourner même quand un disque, un nœud, ou un switch décède !**

- Et idéalement…
  - **Scaler en ajoutant simplement des machines,**
  - **Sans un point central qui bloque tout** (SPOF!)
  - En **pouvant servir du bloc, du fichier, ou de l'objet** selon le besoin **!**
  - En **rééquilibrant tout seul quand on ajoute ou retire du stockage !**

SCALAR

Pourquoi on parle de Ceph ?

Ceph fait tout ça **et le fait BIEN** !

- Il est utilisé **partout en production**
  - OpenStack
  - Kubernetes
  - HPC
  - SI on-prem

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

Voyons son Architecture Générale

CEPH : ARCHITECTURE GÉNÉRALE

Ceph repose sur une couche logicielle appelée RADOS

- **3** types de **démons principaux intégrés à RADOS**
  - **MON** `(ceph-mon)` : **gèrent la carte du Cluster !** *Qui est en vie, ou sont les données, quelles règles de placement utiliser ?*
  - **OSD** `(ceph-osd)` **: Stockent les données**, *en se voyant attribuer du disque, du CPU, de la mémoire de travail…*
  - **MGR** `(ceph-mgr)` **: fournissent surveillance, métriques, orchestrations complémentaires**

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

Ceph : Architecture Générale

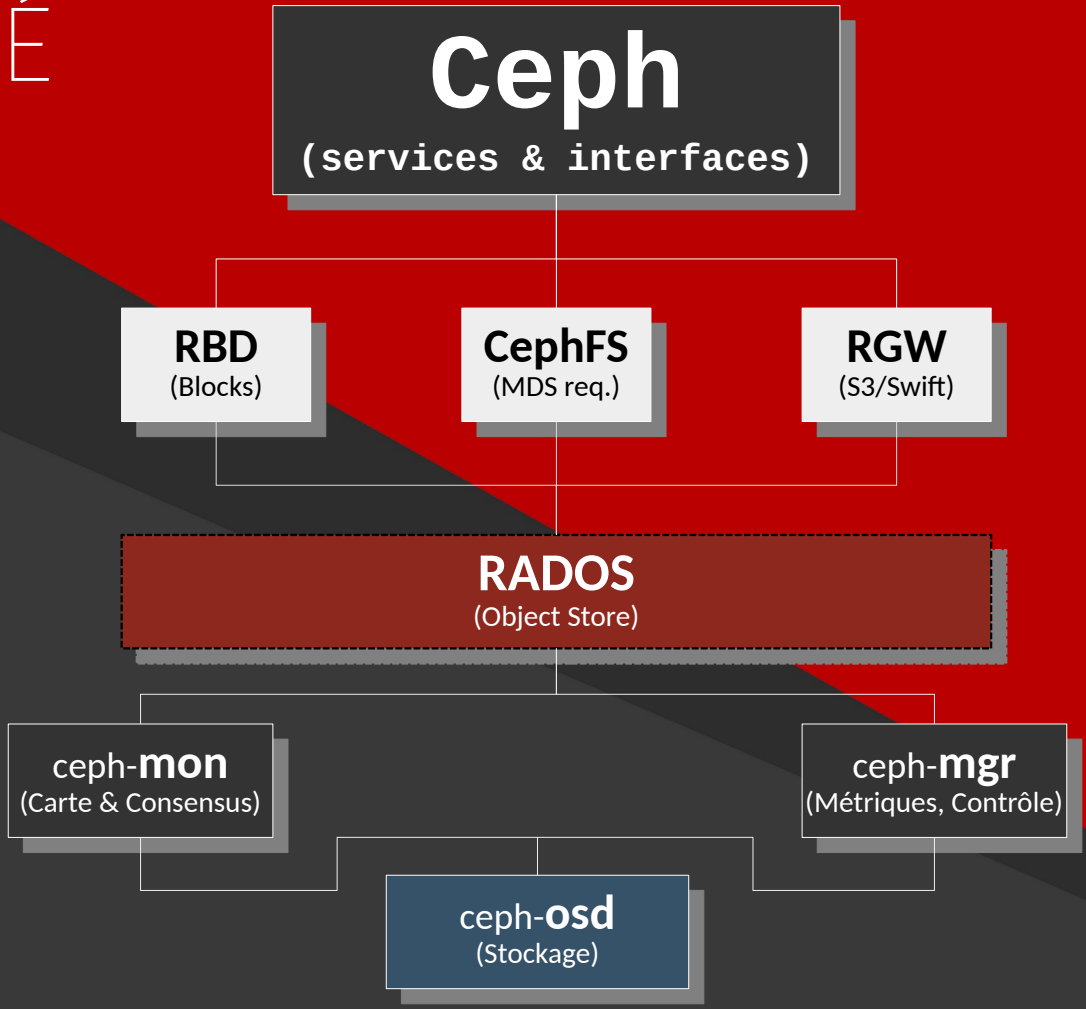
Il existe des démons par dessus RADOS, mais qui n'en font pas partie intégrante

- **MDS** (`ceph-mds`) : Uniquement **si on stocke en mode système de fichiers**, c'est à dire avec **CephFS !**
- **RGW** (`radosgw`) **: Passerelle pour stockage objet Amazon S3**
- RBD (`radosblocks`) **: Stockage en mode blocks**

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

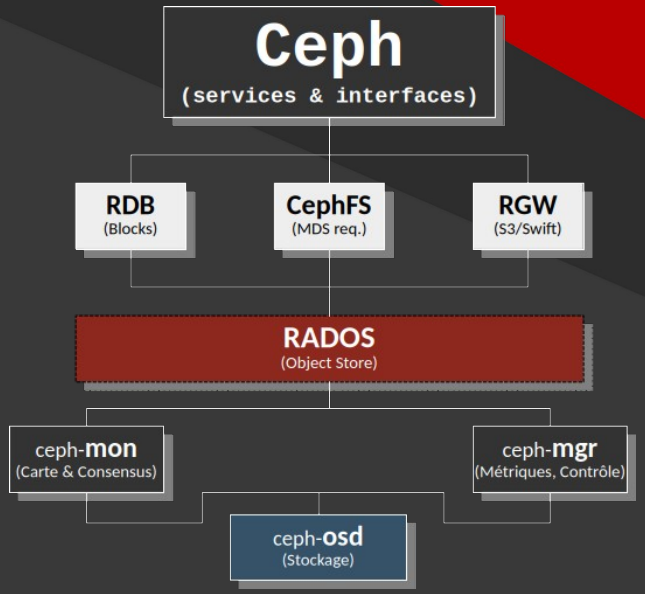
EN RÉSUMÉ

SCALAR



plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

EN RÉSUMÉ



Les clients ne passent pas par un contrôleur central. Ils contactent les MON uniquement pour récupérer des infos, **puis** ils parlent **directement aux OSD.**

Résultat : pas de SPOF, pas de goulot d'étranglement.

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

Et LES DONNÉES se trouvent OU, Concrètement?

CEPH : L'EMPLACEMENT DES DONNÉES

Ceph utilise un algorithme appelé CRUSH

- **Il décide ou mettre les données en fonction d'une carte décrivant la topologie** (racks, nœuds, disques, etc…)
- Placement déterministe : Tout le monde arrive à la même conclusion indépendamment.
- **Pas de base SQL cachée**, pas de lookup massive, **rien**

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

CEPH : L'EMPLACEMENT DES DONNÉES

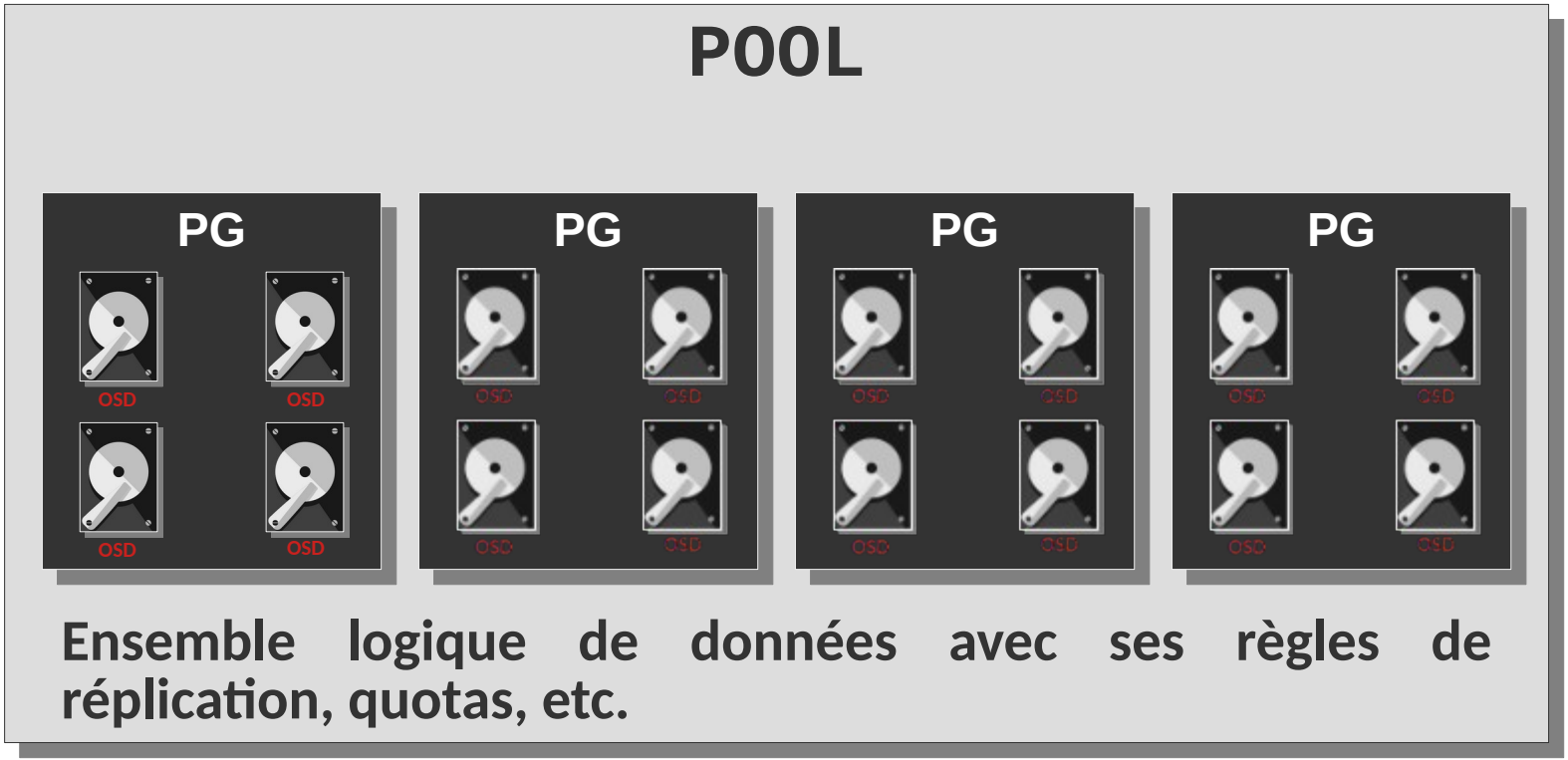
Grâce à CRUSH :

- **Une réplication des données est garantie !**
- **Un disque meurt ? CRUSH recalcule ou recréer les copies**
- **Un disque s'ajoute ? Il redistribue automatiquement, la charge est équilibrée !**

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

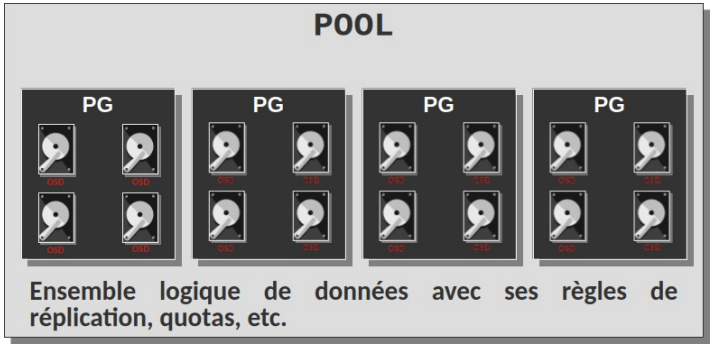
Nomenclature & éléments du stockage de Ceph

Pools & PG dans Ceph :



plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

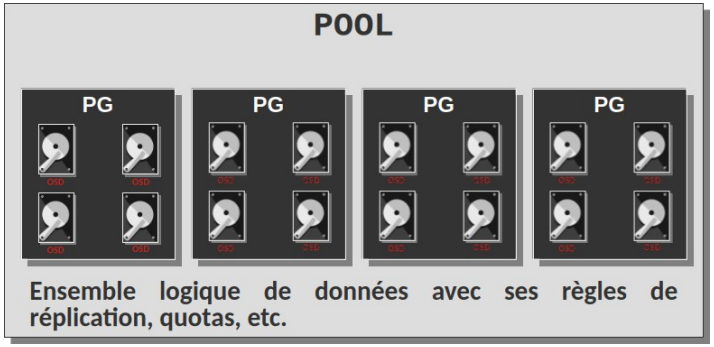
Pools & PG dans Ceph :



Rôle des **PG** :

- **Éviter d'avoir un mapping « objet → OSD » gigantesque**
- **Équilibrer la charge plus proprement**
- **Simplifier la réplication & la réparation**

Tolérance aux pannes dans Ceph :



Ceph assure la **résilience via** :

- **Réplication (x2,x3)**
- **Ou `erasure coding`,** qui optimise la capacité mais plus lourd en CPU

Quand un OSD tombe, Ceph… :

- **Marque la zone inactive**
- **Recalcule via CRUSH ou reconstituer les copies**
- **Les OSD restants rééquilibrent automatiquement**

Choisir son type de Stockage

CEPH : CHOISIR SON TYPE DE STOCKAGE

RBD : RADOS Block Devices, **typiquement adapté pour les cas suivants :**

- **`Libvirt + KVM`**
- **`OpenStack Cinder`**
- **`Kubernetes via CSI RDB`**

Fournis des fonctionnalités assez utiles de snapshots, de clones, de thin provisionning…

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

CEPH : CHOISIR SON TYPE DE STOCKAGE

CephFS : **système de fichiers distribué** conforme POSIX **:**

- **Nécessite des MDS** *(MetaData Servers)*
- **Scalable sur les métadonnées**, et meilleur que NFS pour les charges lourdes !

Utilisé en HPC, big data, clusters de calculs, etc...

SCALAR

CEPH : CHOISIR SON TYPE DE STOCKAGE

RGW : Interface objet compatible Amazon S3/Swift

Utilisé pour les applications cloud-native, buckets pour apps web, backups, etc...

SCALAR

Mécanismes Internes & vocabulaire technique

CEPH : MÉCANISMES & VOCABULAIRE

- **Scrub** : vérification de cohérence
- **Backfill** : Reconstruction après panne !
- **Recovery** : rétablissement

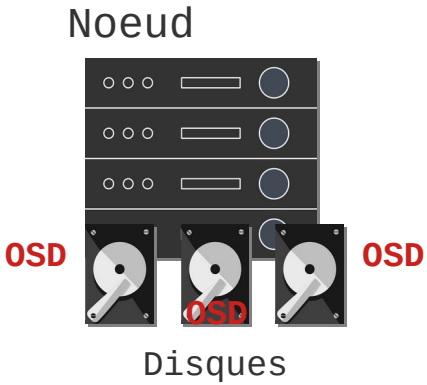
plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

Bonnes Pratiques pour un déploiement Ceph

Ceph : Bonnes Pratiques de déploiement

Ceph est **sensible à 2 choses :**

- **La Latence réseau**
- **L'équilibrage de la topologie**



Donc généralement :

- **3 MON, ou 5, mais jamais pair !**
- **Plusieurs OSD par nœud, et un OSD par disque**
- **Un réseau dédié pour le trafic interne** (cluster network)
- Des disques rapides pour les journaux, comme des SSD ou SSD NVME

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

SCALAR

ATTENTION!

Il faut éviter les machines fourre-tout

(OSD + MON + RGW + MDS dans tous les sens sur la même machine)

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

Okai mais... concrètement,

LES COMMANDES

Ceph : Déploiement

- La majorité du déploiement se fait par **la commande centrale :**

```
# cephadm
```

- La **sous-commande permettant de générer le 1er nœud :**

```
# cephadm bootstrap --nom-node IP_Noeud1
```

Le bootstrap installe :

- **1 MON** sur le nœud en question
- **1 MGR**
- **Un shell Ceph**

plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

Ceph : Commandes Essentielles

- **Vérifier la santé** du Cluster Ceph :

```
# ceph status
```

- **Vérifier la topologie** du Cluster Ceph :

```
# ceph osd tree
```

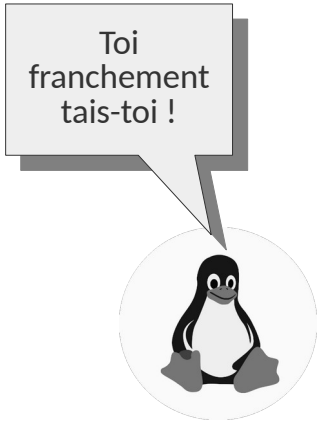
- **Vérifier l'espace :**

```
# ceph df
```


plus d'info sur www.linkedin.com/in/surquincedric
csurquin.pro@pm.me 07.82.69.95.25

Ceph : Documentation

- **La meilleure source d'info pour le déploiement** et la gestion d'un cluster Ceph reste… sa source **officielle !**

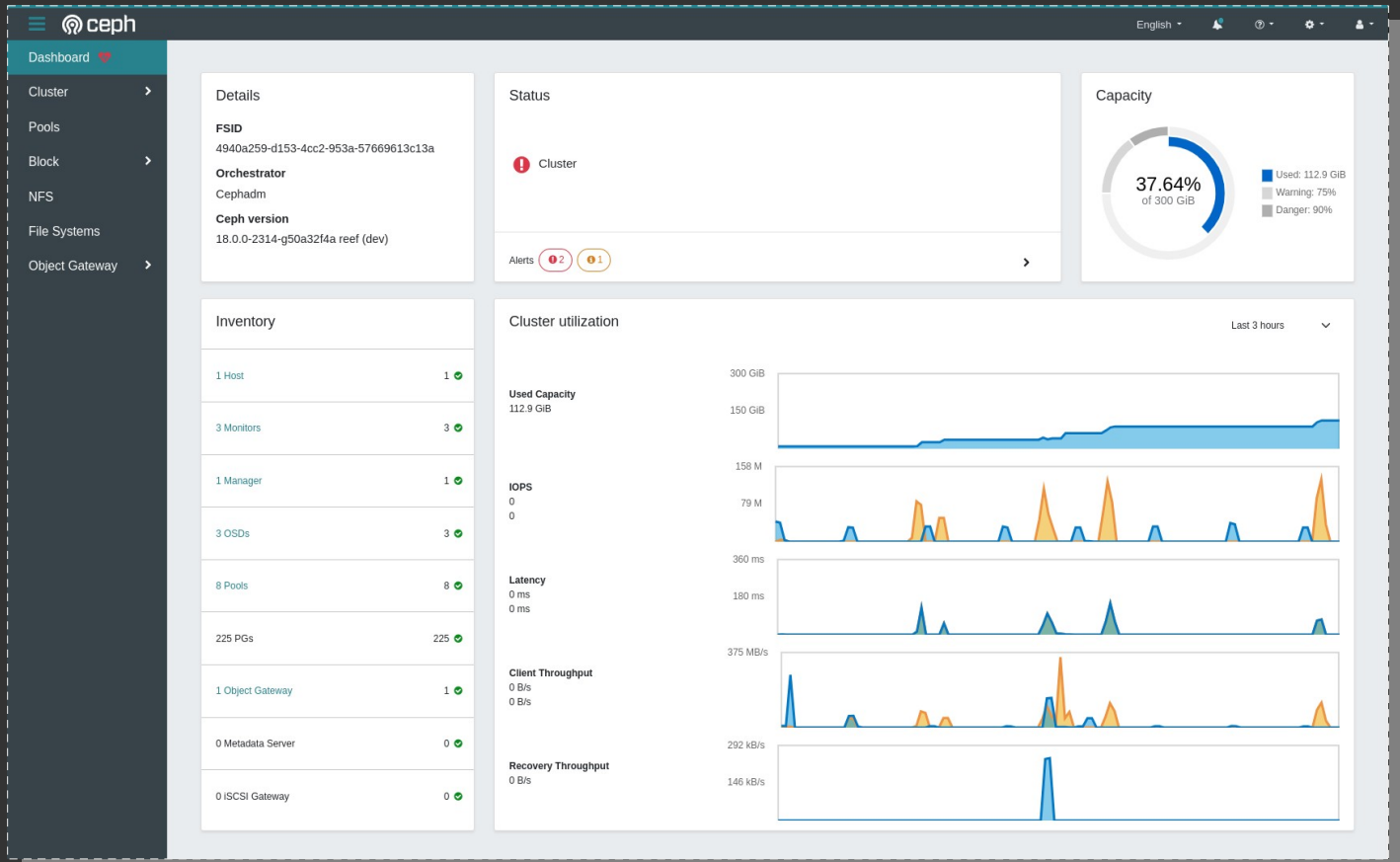


Docs.ceph.com

Ceph Web UI

```
# ceph mgr module enable dashboard
```

Activera la WebUI d'un Cluster Ceph déjà mis en service



Ce cours et les supports d'exercices qui y sont liés sont la propriété exclusive de son auteur

Merci de respecter le temps et le soin y ayant été accordés en ne diffusant pas son contenu sans l'autorisation explicite!

Des inexactitudes ? Des outils dépréciés ? Des fautes ?
contact@scalar-formation.fr

Cédric Surquin
Admin Système Linux & Réseaux Cisco
Consultant & Formateur, auteur du cours.