

## Infrastructure Proxmox au sein de l'OSU Pythéas

Journées Proxmox / CEPH



Julien LECUBIN

Adrien MALGOYRE

AuDACES et ARAMIS - 20/06/2022

# Périmètre de travail

- ➔ 1300 personnes
- ➔ 12 sites
- ➔ 6 UMR et 2 UAR



# Service Informatique Pythéas (SIP)

→ 12 Agents

→ 12 Sites



## Support aux UMR/UAR



## Soutien scientifique





2010

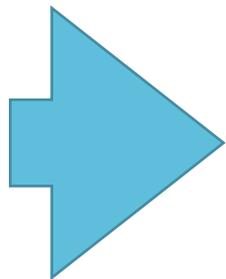
## Contexte de départ

➔ 5 Agents

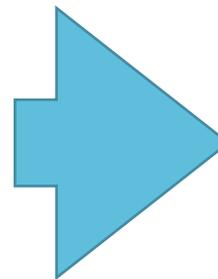
➔ 2 Sites

Serveurs obsolètes  
Pas de virtualisation

OpenSource  
Virtualisation  
Facile à appréhender  
Peu couteux



XEN  
VMWare  
Proxmox



Proxmox

# Choix de virtualisation - OpenVZ / QEMU

OpenVZ devenu LXC (2019)

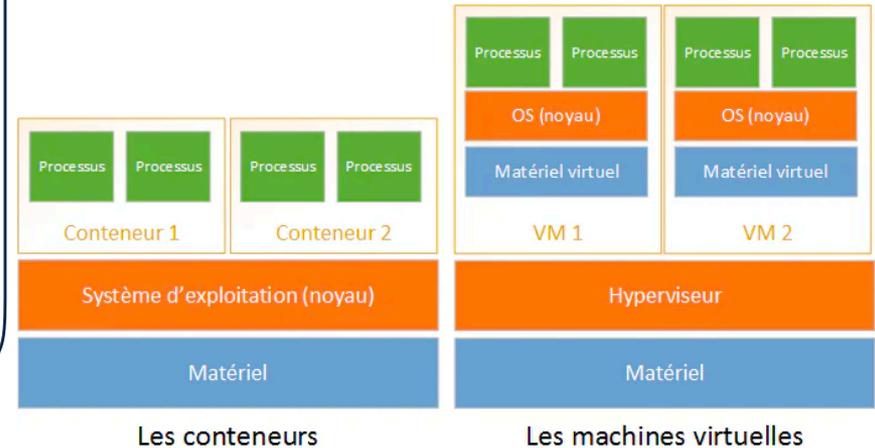
Systeme de fichiers de la VM visible depuis l'hôte

Kernel partagé

Consommation moindre

Moins d'isolation entre la VM et l'hôte

Base d'images disponible



KVM ou LXC? Dépend de l'environnement

# Choix de la virtualisation

## Exemples de services virtualisés sous Proxmox :

### LXC

OS Debian / CentOS

Messagerie, BDD, Cloud, Web,  
DNS, DHCP, Radius, Licences...

### QEMU

OS Linux / Unix

Frontal cluster, applis fortinet (FW),  
anciennes VM Vmware

## Exemples de services non virtualisés sous Proxmox :

**Noeuds de calcul**

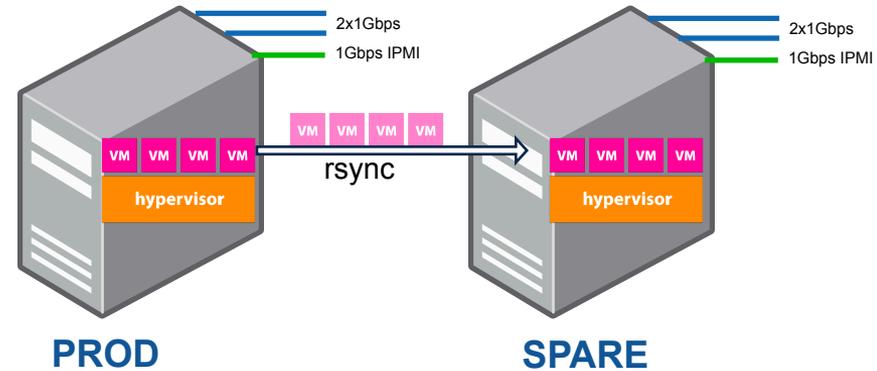
**Serveurs projets (CPU+++)**

**Serveurs Microsoft (hyperV)**

**Stockage**



- 2 serveurs Proxmox v1 identiques
- Filesystem de l'hôte en EXT3
- Deux liens 1Gb/s (LACP)
- ~30 VM / serveur de prod
- VM OpenVZ
- OS Debian uniquement



rsync horaire « -- delete » des VM

Le spare fait une archive par VM, chaque nuit, vers un NAS de backup (NFS)



**Objectif -> on réduit au maximum la volumétrie du container pour :**

- **Optimiser la durée de sauvegarde et de restauration**
- **Réduire l'espace disque utilisé par la sauvegarde de l'ensemble**
- **Contenir la taille du système de fichiers de nos hyperviseurs**

**Montage NFS sur l'hyperviseur des jeux de données**

**La VM peut monter un/des dossier(s) de l'hyperviseur**



## Hyperviseurs en ZFS (depuis Proxmox 3)

### 1 pool ZFS (RaidZ2 sur 6 HDD SATA)

- 1 dataset pour l'hôte
- 1 dataset par VM

Pas de fsck au reboot

Compression

Copie en mode bloc (ZFS send / recv)

Snapshots horaire des VM

Restauration 1 fichier / toute la VM



## Bascule prod => spare

- Couper les VM sur le serveur de prod
- Relancer script de synchro horaire manuellement
- Démarrer les VM sur le serveur de spare

## Panne hardware

- Démarrage VM sur le spare (perte max 1h de données)
- Échange de disques entre serveur de prod et de spare (HBA / ZFS)

# Infrastructure 2018



2018

8 sites

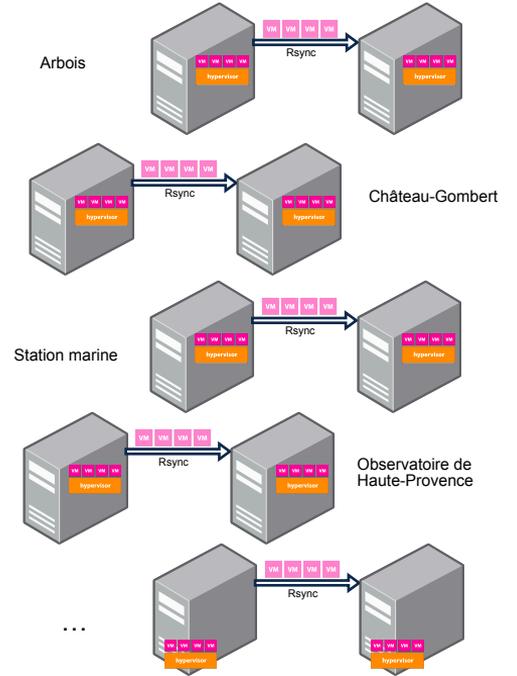
Solution identique quel que soit le site

Full ZFS (Virtualisation Proxmox et stockage FreeNAS)

Matériels en 10Gb/s

Stable et performant

Le personnel du service info sait intervenir via les GUI



~ 20 serveurs proxmox pour 150 VM

## Infrastructure proxmox multi-sites en 2018

### ➔ Le périmètre et les besoins augmentent à moyens humains et financiers constants

- Beaucoup de serveurs à administrer et à renouveler
- Problématique de redémarrage d'un hyperviseur
- Trop de scripts maison
- Pas de HA

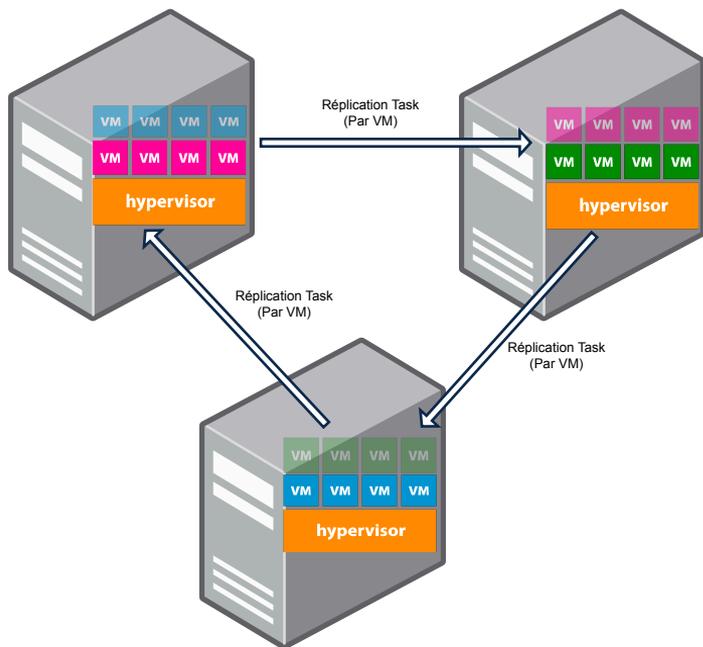
### ➔ Rationaliser l'infrastructure en conservant l'essentiel

- Exploiter toutes les ressources (plus de notion prod/spare)
- Centraliser en un point les services les plus critiques
- Centraliser la sauvegarde des VM
- Plus de scalabilité sur les sites
- Automatiser la reprise d'un service lors d'un incident (HA)



2020

## Mode cluster ZFS / HA



- ✓ Interface de gestion unifiée (sauvegardes, stockage...)
  - ✓ Souples des mises à jour et interventions techniques
  - ✓ Reconversion des serveurs de spare en production.
  - ✓ Bascule simplifiée, quelque soit le sens.
  - ✓ Bascule automatisée dans le cadre du HA
  - ✓ Réduction du délai entre 2 copies (par défaut à 15min).
- 3 Noeuds minimum recommandés surtout pour du HA

## Mode cluster ZFS / HA

Réplication des VM en mode bloc via les outils internes à Proxmox = pas d'erreurs humaines

Avec un même nombre de machines physiques nous accueillons plus de VM

Respecter le QUORUM (primordial en HA) = 3 Noeuds

Liens dédiés pour CoroSync (latence)

Pour chaque VM ne pas oublier le « Replication Job »

Rédiger un PRA!

Enabled	Guest ↑	Job ↑	Target	Status	Last Sync	Dur...	Next Sync	Sched...	Comment
<input checked="" type="checkbox"/>	345	0	sipvirt4c	✓ OK	2022-06-02 15:45:04	4.7s	2022-06-02 16:00:00	*/15	
<input checked="" type="checkbox"/>	392	0	sipvirt4c	✓ OK	2022-06-02 15:45:09	6.1s	2022-06-02 16:00:00	*/15	
<input checked="" type="checkbox"/>	395	1	sipvirt4c	✓ OK	2022-06-02 15:45:15	3.2s	2022-06-02 16:00:00	*/15	

Create: Replication Job

CT/VM ID:

Target: sipvirt1s

Schedule: \*/15 - Every 15 minutes

Rate limit (MB/s): unlimited

Comment:

Enabled:

[Help](#) [Create](#)

## Mode cluster ZFS / HA

Qu'avons nous fait de notre modèle basé sur 2 Noeuds?

Selon les sites nous avons:

- Ajouté un 3ème noeud pour permettre la mise en place du HA
- Conservé 2 noeuds sans HA et rédigé une procédure de secours (actions manuelles)
- Conservé notre ancien modèle sans le mode cluster (Rsync)
- ~~Ajouté un Raspberry avec Debian + corosync-qdevice pour faire un Quorum avec seulement 2 Proxmox~~



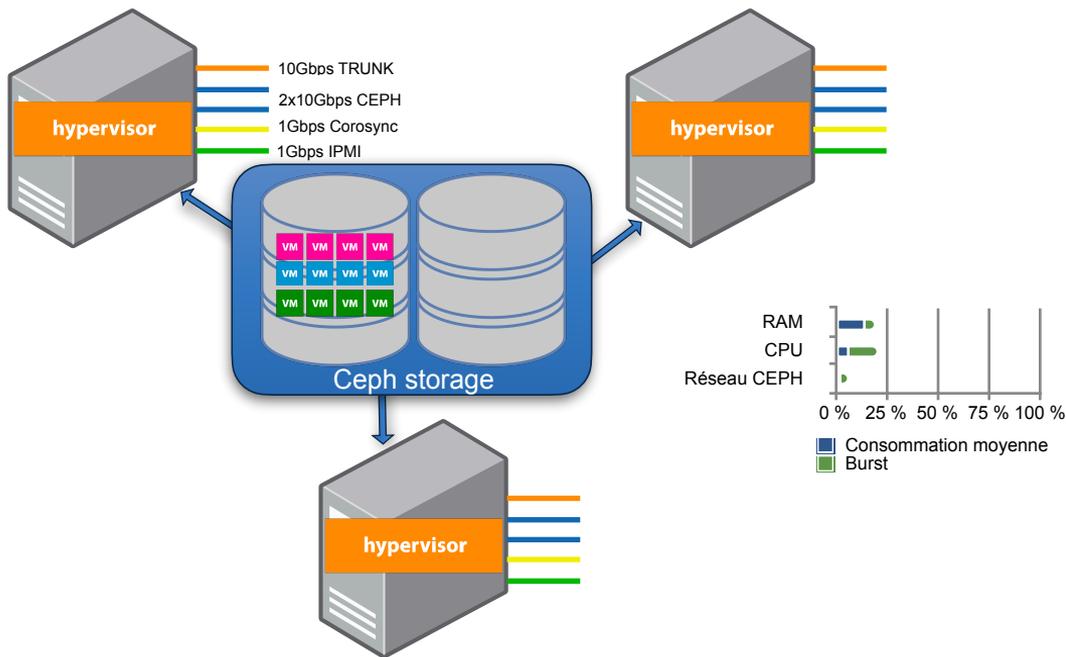
Dans le cadre de la centralisation de nos services critiques nous voulions

Disponibilité sur le service / la donnée

Scalabilité

- ✓ Plus de décalage de la donnée des VM entre les noeuds
- ✓ Pas (ou peu) d'interruption de service en cas de migration d'un noeud a un autre
- ✓ Pas (ou peu) d'interruption de service en cas de crash d'un noeud
- ✓ Augmentation simple de la volumétrie et/ou des ressources au cours du temps
- ✓ Première experience en production du stockage distribué

## Mode cluster CEPH / HA



Consommation moyenne du cluster : 800Wh

### • Hardware

- ▶ 3 serveurs DELL R640:
  - ▶ 2 x Intel Xeon 20 coeurs @ 2.10GHz
  - ▶ 384 Go de RAM
  - ▶ 256Go pour l'OS (Carte BOSS)
  - ▶ Réseau 4x10Gbps SFP+ Intel
  - ▶ 1 Sata 1To + 2 NVME Intel 3.2 To

### • Ceph Pacific

- ▶ 3 OSD par noeud: 1 Sata, 2 nvme
- ▶ 3 Monitor, 3 Metadata, 3 Manager
- ▶ 2 Pools RBD ( 1 sata / 1 nvme )
  - NVME: Réplication 3/2
  - 128 PG
  - 6,7To utiles (~1/3 du brut)
- ▶ 1 Pool dédiés CEPHFS (sata+nvme)

### • Hyperviseurs Proxmox 7.1

- ▶ 55 VM LXC
- ▶ 3 VM QEMU

## Infra CEPH

Migration temps réel

Pas de décalage des données

Plus besoin de « replication job » par vm

Stockage extensible

CEPH graphique avec Proxmox

Accès aux fichiers du conteneur depuis l'hôte

Accès aux snapshots

Complexité de Ceph

Mises à jour (attention!)



Le volume augmentait et les performances se dégradait

✓ Réglé après ajout d'OSD

MAJ Ceph Nautilus vers Octopus = on avait tout cassé!

Réinstallation from scratch du cluster entier via IDRAC

✓ Lié à un bug de migration

RBD I/O Error sur un des hyperviseurs ?

# Sauvegardes



- Les snapshots utilisés dans le cadre des « replication job » ne sont pas visibles dans l'interface graphique.

- Des snapshots fait manuellement avec les commandes ZFS/RBD ne le sont pas non plus.

Le seul moyen de les rendre accessible est d'utiliser « pct snapshot »

Malheureusement, il n'est pas possible de définir des rétentions spécifiques pour les snap. Par exemple horaire sur 1 journée, journalier sur 30 jours...

—> proxmox-autosnap (Exemple: <https://github.com/apprell/proxmox-autosnap>)

Network	automonthly220501000023	2022-05-01 00:00:23	autosnap
DNS	autodaily220508000024	2022-05-08 00:00:24	autosnap
Options	autodaily220509000021	2022-05-09 00:00:21	autosnap
Task History	autohourly220509170021	2022-05-09 17:00:21	autosnap
Backup	autohourly220509180025	2022-05-09 18:00:25	autosnap
Replication	autohourly220509190026	2022-05-09 19:00:25	autosnap
Snapshots	autohourly220509200021	2022-05-09 20:00:21	autosnap
	autodaily220510000022	2022-05-10 00:00:21	autosnap

# Proxmox Backup Server



2019

- Interface dédiée aux sauvegardes proxmox
  - Deduplication ZFS
  - Snapshots Proxmox File Archive Format (.pxar)
- Politique de sauvegarde et de rétention commune
- Stockage local et/ou distant + archivage sur bandes
- Modèle client/serveur (API)

Name ↑	Size	Modified	Type
root.pxar.didx			Directory
pxarexclude-cli	11 B		File
bin			Directory
boot			Directory
dev			Directory
etc			Directory
.pwd.lock	0 B	Mon Jul 08 2019 07:33:25 G...	File
X11			Directory
adduser.conf	2.91 KIB	Mon Jul 08 2019 07:33:25 G...	File
aliases	3.74 KIB	Tue Jul 07 2020 11:15:38 G...	File
aliases.db	48.00 KIB	Tue Jul 07 2020 11:28:16 G...	File

PROXMOX Backup Server 2.1-1

Dashboard  
Configuration  
Access Control  
Remotes  
Traffic Control  
Certificates  
Subscription  
Administration  
Shell  
Storage / Disks  
Tape Backup  
Datastore  
zpool  
Add Datastore

Datastore: zpool

Summary Content Prune & GC Sync

Reload Verify All Prune All

Backup Group ↑

- ct/100120
- ct/100150
  - ct/100150/2022-03-31T22:08:49Z
    - catalog.pcat1.didx
    - client.log.blob
    - index.json.blob
    - pct.conf.blob
    - root.pxar.didx
  - ct/100150/2022-04-04T22:07:26Z
  - ct/100150/2022-04-05T22:07:20Z
  - ct/100150/2022-04-06T22:10:16Z
  - ct/100150/2022-04-07T22:08:26Z
  - ct/100150/2022-04-10T22:07:42Z
  - ct/100150/2022-04-11T22:08:06Z

- Restauration Fichier/VM via hyperviseurs
- Contrôle d'intégrité

# Proxmox Backup Server

Storage	local-zfs	ZFS	Disk image
Backup	mdt	NFS	Disk image /mnt/pve/mdt
Replication	nvme	RBD (PVE)	Disk image, Container
Permissions	pbs	Proxmox Ba...	VZDump backup file
Users	sata	RBD (PVE)	Disk image
	srvbcp1d	NFS	VZDump backup file /mnt/pve/srvbcp1d

Edit: Proxmox Backup Server

General Backup Retention Encryption

ID: pbs Nodes: All (No restrictions)

Server: pbs.ospytheas.fr Enable:

Username: root@pam Content: backup

Password: \*\*\*\*\* Datastore: zpool

Fingerprint: de:20:77:19:21:db:4b:ee:69:6c:2f:ab:72:f4:4d:fa:cc:cc:c9:2e:45:f1:95:2f:28:32:

Help OK Reset

Edit: Proxmox Backup Server

General Backup Retention Encryption

Keep all backups

Keep Last: [ ] Keep Hourly: [ ]

Keep Daily: [ ] Keep Weekly: [ ]

Keep Monthly: [ ] Keep Yearly: [ ]

It's preferred to configure backup retention directly on the Proxmox Backup Server.

Help OK Reset

Encryption

Encryption Key: None

Do not encrypt backups

Auto-generate a client encryption key

Upload an existing client encryption key

Help OK Reset

16

VM QEMU

20 Go

Espace moyen VM QEMU

10Go

Vol. PBS QEMU /jour

410

Snapshots (Mai 2022)

220

Conteneur LXC

3,9 Go

Espace moyen VM LXC

50Go

Vol. PBS LXC /jour

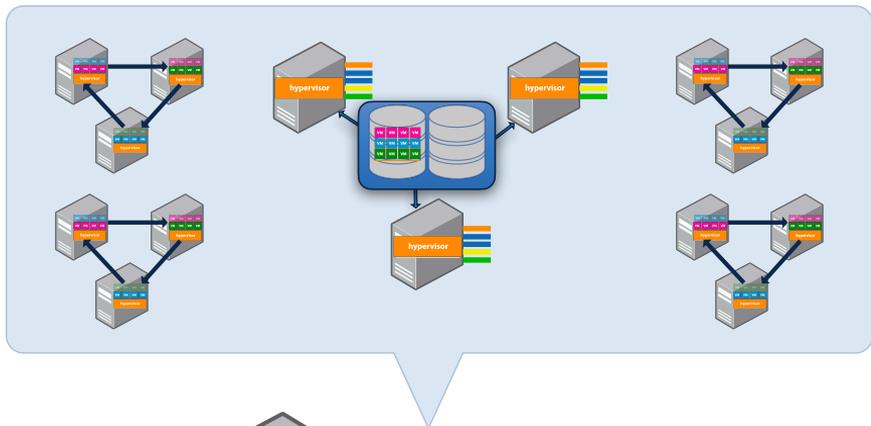
7300

Snapshots (Mai 2022)

## Politique de sauvegarde

	Type	Granularité	Emplacement	Pas de temps	Cout stockage	Rétention
<b>PBS</b>	Snapshots .pxar	VM/Fichier	Distant (PBS)	1/jour	- - (dedup)	720h/60d/30w/ 12m/2y
<b>BackupJob</b>	.tar.gz	VM	Distant (NAS)	1/jour	+ +	15d/5w/6m
<b>Réplication task</b> (hors ceph)	Snapshots ZFS	-	Cluster	15 min	- -	Aucune
<b>Pct Snapshots</b>	Snapshots ZFS ou RBD	VM/Fichier/ Dossier	Local (ou pool ceph)	Horaire	+	36h/5d/2m

# Topologie / Bilan 2022



PBS  
centralisé

Backup  
centralisé



2022

1

Cluster Proxmox  
CEPH (3 nodes)

10

Clusters Proxmox  
ZFS (2-3-4 nodes)

6

Proxmox ZFS  
(Ancien modèle)

20

VM QEMU

~ 250

Conteneur LXC

20 Go

Espace moyen VM  
QEMU

~ 4 Go

Espace moyen VM  
LXC

420Go

Vol. Archives  
QEMU /jour

~ 1To

Vol. Archives LXC /  
jour

10Go

Vol. PBS QEMU /jour

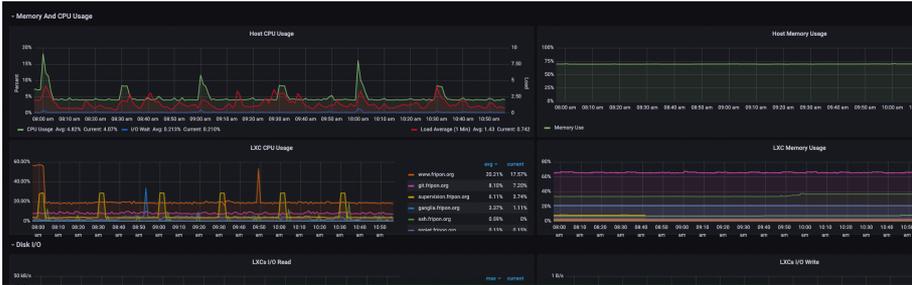
50Go

Vol. PBS LXC /jour

# Supervision



2021

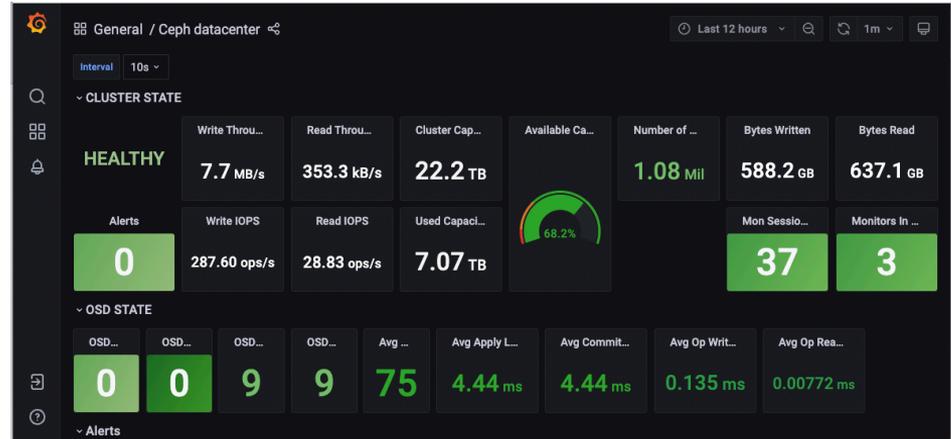


- **PROXMOX** cluster dashboard (InfluxDB)

- **GRAFANA** Ceph dashboard

- + Outils classiques:

- Icinga2 / Ganglia
- Alertes E-mail
- Syslog....



## Perspectives

Automatisation de la réplication-task (via API)

Supervision de tous les cluster (via API)

CEPH sur notre infrastructure de stockage?

Réseau 40 ou 100 Gbps ?

Passage de 3 à 5 serveurs dans l'infra ceph?

Autoconfiguration des noeuds via ansible