

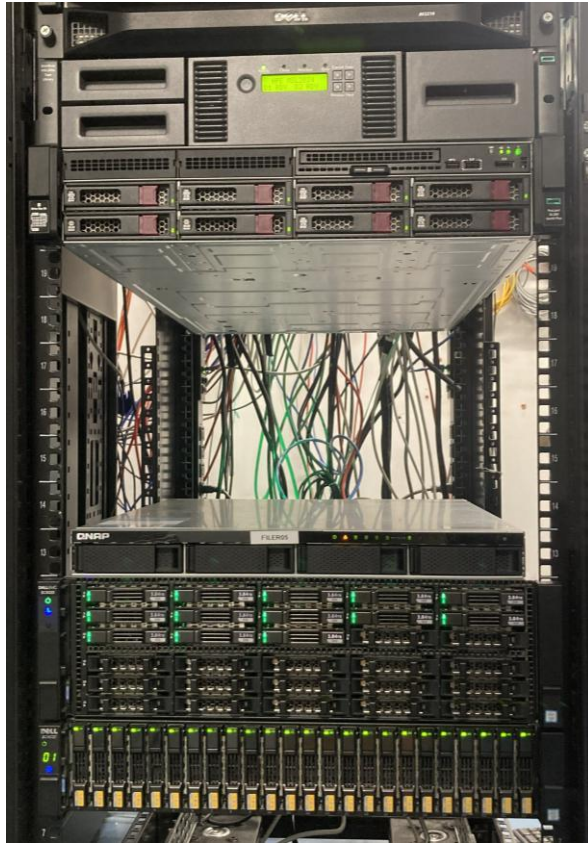
PROXMOX ET CEPH

Vers un système hyperconvergé





PARTIE 1 : ETAT DES LIEUX



SALLE SERVEUR

3 Serveurs HP DL380

Proxmox 8

Récupération de 3 serveurs
DELL R440

Vcenter

Récupérations de 3 serveurs
DELL R630

HyperV

Baie DELL SC4020 et SC5020

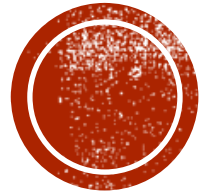
POURQUOI LA SOLUTION CEPH ?

- Fin du SAN : boîte noire, multipath capricieux avec Proxmox

CEPH : flexibilité et de la performance sans payer (licences).

- transforme des disques locaux en un pool géant (intelligent et distribué).
- scalabilité (on ajoute un nœud = on ajoute de la puissance ET du stockage)
- éviter un SPOF





PARTIE 2 : ARCHITECTURE ET DESIGN





TOPOLOGIE MATERIELLE

6 nœuds HP et DELL

256G de RAM/nœud

1 Pool SSD (1,92Tb * 15)

1 Pool HDD 10K tours (1,8Tb * 9)

No RAID

Châssis disque différent

Spare HDD et SSD

Nom	Brand	CPU	RAM	DATA				
Proxnodes01	HP ProLiant DL360 Gen10	2 Intel(R) Xeon(R) Silver 4314 CPU @ 2.40GHz (16cores)	256Gb	SSD 1,92Tb	SSD 1,92Tb	Libre	2xNVMe (OS)	
				SSD 1,92Tb	SSD 1,92Tb	Libre	Libre	Libre
Proxnodes02	HP ProLiant DL360 Gen10	2 Intel(R) Xeon(R) Silver 4314 CPU @ 2.40GHz (16cores)	256Gb	SSD 1,92Tb	SSD 1,92Tb	Libre	2xNVMe (OS)	
				SSD 1,92Tb	SSD 1,92Tb	Libre	Libre	Libre
Proxnodes03	HP ProLiant DL360 Gen10	2 Intel(R) Xeon(R) Silver 4314 CPU @ 2.40GHz (16cores)	256Gb	SSD 1,92Tb	SSD 1,92Tb	Libre	2xNVMe (OS)	
				SSD 1,92Tb	SSD 1,92Tb	Libre	Libre	Libre
Proxnodes04	DELL PowerEdge R630	2 Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz (8 cores)	256Gb	CDROM		SSD 1,6Tb	HDD 1,8T	SSD 1,92Tb
				SSD 480G	SSD 480G	HDD 1,8T	HDD 1,8T	SSD 1,92Tb
Proxnodes05	DELL PowerEdge R630	2 Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz (8 cores)	256Gb	SSD 480G	SSD 1,6Tb	HDD 1,8T	Libre	SSD 1,92Tb
				SSD 480G	HDD 1,8T	HDD 1,8T	Libre	SSD 1,92Tb
Proxnodes06	DELL PowerEdge R630	2 Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz (8 cores)	256Gb	SSD 480G	SSD 1,6Tb	HDD 1,8T	Libre	SSD 1,92Tb
				SSD 480G	HDD 1,8T	HDD 1,8T	Libre	SSD 1,92Tb



Fibre 10G



Fibre 10/25G



Nœud Proxmox/CEPH

eno1	Network Device		Corosync1
eno2	Network Device		Corosync2
idrac	Unknown		
vibr0	Linux Bridge	bond0	PVE trunk B,D trk lacp
vibr1	Linux Bridge	bond1	Flux Backup
vibr2	Linux Bridge	bond2	Flux CEPH
vibr3	Linux Bridge	bond3	Flux SC4020-IP1
vibr4	Linux Bridge	bond4	Flux SC4020-IP2

RJ45 1G



LE RÉSEAU

Chaque nœud :

1 Idrac (RJ45)

Flux LACP PVE (RJ45)

2 Corosync 1 et 2 (RJ45)

Fibre MTU 9000

Flux LACP CEPH

Flux LACP Backup

Flux Baie DELL SCxx DP1

Flux Baie DELL SCxx DP2

Pool #	Name	Size/min	# of Placement Gro...	Optimal # of PGs	Autoscaler Mode	CRUSH Rule (ID)	Used (%)
1	.mgr	3/2	1	1 n/a	on	replicated_rule (0)	72.76 MiB (0.00%)
2	slow_data_pool	3/2	128	1 n/a	on	hdd_rule (1)	6.00 TiB (44.17%)
3	ssd_data_pool	3/2	128	1 n/a	on	ssd_rule (2)	3.58 TiB (36.82%)

9.58 TiB

- Hiérarchie CRUSH (couche logique)
 - CRUSH s'assure qu'une même donnée n'est **jamais** répliquée deux fois sur le même hôte.
 - Grâce aux device classes (HDD et SSD), c'est comme 2 "sous-clusters" virtuels.
Le trafic de boot/système reste sur le pool SSD.
Le trafic de stockage de masse reste sur le pool HDD.
- Le PG Autoscaler (Le "Pilote Automatique")
 - J'ai laissé l'Autoscaler en mode on. Il gère dynamiquement la division des PGs à mesure que nos pools se remplissent, garantissant environ 100 PGs par OSD sans intervention humaine.
- réplication 3/2 sur 6 nœuds
 - **Disponibilité** : Je peux perdre **2 nœuds** simultanément sans perdre de données
 - **Continuité** : Je peux perdre 1 nœud et continuer à écrire (puisque il reste 2 copies, respectant le **min_size 2**).


CONFIGURATION LOGIQUE

Ceph Squid 19.2.3 sur PVE 9.1.9

Crush class

PG Autoscaler

Schéma de réplication 3/2



INSTALLATION CEPH DANS PROXMOX (1/4)

The screenshot shows the Proxmox VE interface for node 'proxmox-03'. The left sidebar contains a tree view of the cluster resources, with 'proxmox-03' selected. The main panel displays the 'Ceph' configuration page, which includes a 'Status' section with a question mark icon and the text 'Ceph is not installed on this node. Would you like to install it now?'. Below this text is a blue 'Install Ceph' button. The 'System' menu on the left is open, and 'Ceph' is highlighted. The 'Ceph Version' is listed as 18.2.2.

The screenshot shows the 'Setup' screen of the Ceph installation wizard. The 'Info' tab is active, displaying the following text:

Ceph?

"Ceph is a unified, distributed storage system, designed for excellent performance, reliability, and scalability."

Ceph is currently not installed on this node. This wizard will guide you through the installation. Click on the next button below to begin. After the initial installation, the wizard will offer to create an initial configuration. This configuration step is only needed once per cluster and will be skipped if a config is already present.

Before starting the installation, please take a look at our documentation, by clicking the help button below. If you want to gain deeper knowledge about Ceph, visit ceph.com.

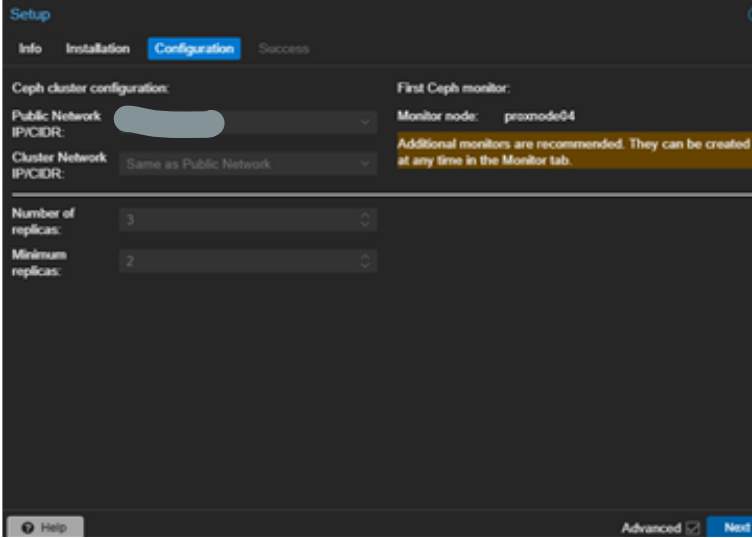
Hint: The no-subscription repository is not the best choice for production setups.

Ceph in the cluster: Could not detect a ceph installation in the cluster

Ceph version to install: squid (19.2) Repository: No-Subscription

Buttons: Help, Start squid installation

INSTALLATION CEPH DANS PROXMOX (2/4)



Recommandations pour 6 nœuds
5 MON -> toujours un chiffre impair
3 MGR suffisent "**1 actif / 2 standby**"

❖ **Séparation du trafic Public** (VM vers Ceph) du trafic Cluster (Réplication /Reconstruction entre OSD) sur deux réseaux distincts

❖ Les MON maintiennent **l'état du cluster** (Quorum, coordination du cluster), tandis que les MGR fournissent des services de **gestion et de monitoring**.

Monitor					
▶ Start	■ Stop	↻ Restart	➕ Create	➖ Destroy	📄 Syslog
Name ↑	Host	Status	Address	Version	
mon.proxnode02	proxnode02	running		19.2.3	
mon.proxnode03	proxnode03	running		19.2.3	
mon.proxnode04	proxnode04	running		19.2.3	
mon.proxnode05	proxnode05	running		19.2.3	
mon.proxnode06	proxnode06	running		19.2.3	

Manager					
▶ Start	■ Stop	↻ Restart	➕ Create	➖ Destroy	📄 Syslog
Name ↑	Host	Status	Address	Version	
mgr.proxnode04	proxnode04	standby		19.2.3	
mgr.proxnode05	proxnode05	active		19.2.3	
mgr.proxnode06	proxnode06	standby		19.2.3	

INSTALLATION CEPH DANS PROXMOX (3/4)

OSD (Object Storage Daemon)

Name	Class	OSD Type	Status	Version	weight	reweight	Used (%)	Total	Apply/Commit Latency (ms)	PGs
default										
proxnode09										
osd.11	ssd	bluestore	up / in	19.2.3	1.45549	1.00	0.04	1.46 TiB	0 / 0	128
osd.10	hdd	bluestore	up / in	19.2.3	1.80109					
osd.9	hdd	bluestore	up / in	19.2.3	1.80109					
osd.8	hdd	bluestore	up / in	19.2.3	1.80109					
proxnode08										
osd.7	ssd	bluestore	up / in	19.2.3	1.45549					
osd.6	hdd	bluestore	up / in	19.2.3	1.80109					
osd.5	hdd	bluestore	up / in	19.2.3	1.80109					
osd.4	hdd	bluestore	up / in	19.2.3	1.80109					
proxnode07										
osd.3	ssd	bluestore	up / in	19.2.3	1.45549					
osd.2	hdd	bluestore	up / in	19.2.3	1.80109					
osd.1	hdd	bluestore	up / in	19.2.3	1.80109					
osd.0	hdd	bluestore	up / in	19.2.3	1.80109					

PG (Placement Group)

Create: Ceph OSD

Disk: /dev/sdd DB Disk: /dev/sdc

DB size (GiB): Automatic

Encrypt OSD:

WAL Disk: use OSD/DB disk

Device Class: HDD

WAL size (GiB): Automatic

Note: Ceph is not compatible with disks backed by a hardware RAID controller. For details see [the reference documentation](#).

Help Advanced Create

Les métadonnées: si le DB/WAL est sur le HDD, une simple écriture de donnée est freinée par la latence du disque mécanique.

Conseil CEPH:

Le SSD est le cerveau (DB), le HDD est le muscle (Data).

Facteur de risque : Attention ! Si le SSD contenant les DB/WAL de plusieurs OSD tombe, tous les OSD associés tombent aussi

Details: OSD 6

Reload

General Network **Devices**

Device	Type	Physical Device	Size
block	hdd	sdf	1.64 TiB
db	ssd	sdc	167.66 GiB

CDROM	SSD 1,6Tb	HDD 1,8T	SSD 1,92Tb
SSD 480G	SSD 480G	HDD 1,8T	HDD 1,8T
			SSD 1,92Tb

INSTALLATION CEPH DANS PROXMOX (4/4)

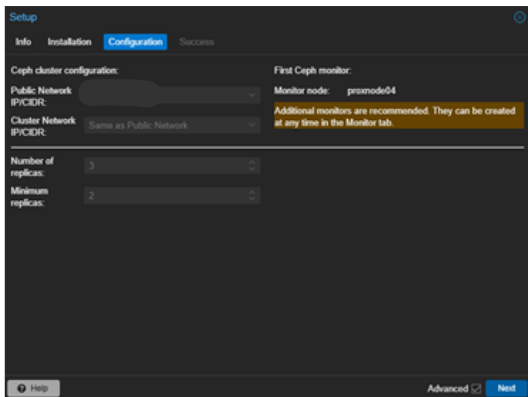
The screenshot displays the Proxmox VE 9.1.7 interface. The left sidebar shows a tree view of the datacenter with nodes proxnode01 through proxnode06. The 'proxnode04' node is selected, and the 'Ceph' menu item is highlighted in the navigation pane. The main content area is divided into three sections:

- Health:** Shows a green checkmark icon and the text 'HEALTH_OK'. Below it, a 'Summary' table indicates 'No Warnings/Errors'. At the bottom, it states 'Ceph Version: 19.2.3'.
- Status:** Features a large green donut chart. To its left, a table shows OSD status: 27 Up, 0 Down, 0 In, and 0 Out. To its right, it shows 'PGs: active+clean: 257' and 'Total: 27'.
- Services:** Lists the status of various services with green checkmarks:
 - Monitors:** proxnode02, proxnode03, proxnode04, proxnode05, proxnode06.
 - Managers:** proxnode04, proxnode05, proxnode06.
 - Metadata Servers:** (No specific entries are visible).



PARTIE 3 : LE RETEX DES PROBLÈMES

PROBLÈME DU RÉSEAU CEPH



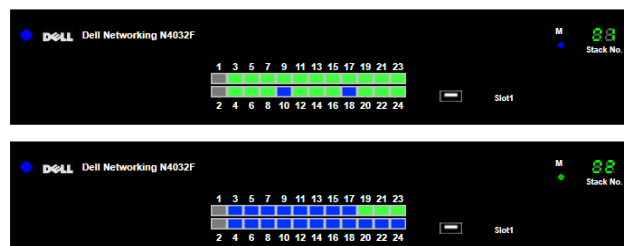
MGR et MON disabled

Le réseau avait des pertes de connexion à cause de mauvaises configurations :

- MTU à 9000
- du LACP sur les switch DELL



Home: Stack View



- Des fibres utilisées

```
GESTIONNAIRE DE TESTS RÉSEAUX
1) Tester CEPH & BACKUP
2) Tester SC4020
3) Tester Nœuds Standards
4) TOUT TESTER (Complet)
5) TEST PERSONNALISÉ (IP unique)
q) Quitter

Sélectionnez une option : 4

>>> LANCEMENT DE LA SÉQUENCE COMPLÈTE <<<

--- Groupe : CEPH ---
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK

--- Groupe : BACKUP ---
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK

--- Groupe : SC4026 ---
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK

--- Groupe : SC4020 ---
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK
[STD 1500] : [x] OK
[MTU 9000] : [x] OK

--- Nœuds Standards (MTU 1500) ---
[STD 1500] : [x] OK
```

PROBLÈMES DE CONFIGURATION CEPH

Si en cas de gros pépin CEPH lors du déploiement on peut revenir à l'état sans CEPH si on exécute le script suivant :

```
#!/bin/bash
systemctl stop ceph-mon.target
systemctl stop ceph-mgr.target
systemctl stop ceph-mds.target
systemctl stop ceph-osd.target
rm -rf /etc/systemd/system/ceph*
killall -9 ceph-mon ceph-mgr ceph-mds
rm -rf /var/lib/ceph/* /var/log/ceph/*
pveceph purge
apt purge ceph-mon ceph-osd ceph-mgr ceph-mds -y
apt purge ceph-base ceph-mgr-modules-core -y
rm -rf /etc/ceph/*
rm -rf /etc/pve/ceph.conf
rm -rf /etc/pve/priv/ceph.*
apt autoremove -y
reboot # OBLIGATOIRE !!!
```

SUPPRESSION D'UN MON DÉFAILLANT

```
echo "=====
echo " 1. RETRAIT LOGIQUE DU MONITEUR $TARGET_MON
echo "=====
echo "Tentative de retrait du moniteur de la Monmap..."
ceph mon remove "$TARGET_MON"
if [ $? -eq 0 ]; then
echo "☑ SUCCÈS : Le moniteur $TARGET_MON a été retiré de la Monmap."
else
echo "✗ ÉCHEC : La commande 'ceph mon remove' a échoué. Poursuite du nettoyage local."
fi
echo "--- Vérification du nouveau quorum ---"
ceph -s | grep "mon:"
echo "-----"
---
## 2. NETTOYAGE CHIRURGICAL ET TENTATIVE DE RELANCE
echo ""
echo "=====
echo "2. NETTOYAGE CHIRURGICAL SUR $TARGET_MON "
echo "=====
echo "⚠ Connexion SSH vers $NODE_IP_TO_CLEANUP requise (mot de passe ou clé root)."
# Nettoyage des fichiers locaux sur le nœud ciblé via SSH
ssh "root@$NODE_IP_TO_CLEANUP" << EOF
echo "Arrêt du service ceph-mon@$TARGET_MON ..."
systemctl stop ceph-mon@$TARGET_MON.service
echo "1. Suppression des répertoires de données Ceph MON locaux (/var/lib/ceph/mon/ceph-$TARGET_MON) ..."
rm -rf "/var/lib/ceph/mon/ceph-$TARGET_MON"
echo "2. Nettoyage CHIRURGICAL du fichier de configuration $CEPH_CONF_PATH ..."
if [ ! -f "$CEPH_CONF_PATH" ]; then
echo "✗ ERREUR : Le fichier $CEPH_CONF_PATH n'existe pas. Nettoyage de la configuration locale ignoré."
else
CEPH_IP_ESC=$(echo "$CEPH_MON_IP_TO_REMOVE" | sed 's/\./\\./g')
# 2a. Supprimer la section [mon.proxnode09] (et l'IP associée)
sed -i "/^\[mon\\.\\$TARGET_MON\\]/,/^public_addr = $CEPH_IP_ESC/d" $CEPH_CONF_PATH
```

LE WIPE QUI NE WIPE PLUS... 1/2

The screenshot shows a disk management interface with a sidebar on the left and a main table of disks. The 'Wipe Disk' tab is active. An error dialog box is overlaid on the table, indicating that disk /dev/sdf has a holder (500).

Device	Type	Usage	Size	GPT	Model	Serial	S.M.A.R.T.	M
/dev/sda	SSD	partitions	480.10 GB	Yes	SS		PASSED	
/dev/sda1	partition	BIOS boot	1.03 MB	Yes				
/dev/sda2	partition	EFI	1.07 GB	Yes				
/dev/sda3	partition	ZFS	479.03 GB	Yes				
/dev/sdb	SSD	partitions	480.10 GB	Yes	SSC		PASSED	
/dev/sdc	SSD	LVM, Ceph (DB)	1.60 TB	No	INTEL_S...		PASSED	
/dev/sdd	unknown	LVM, Ceph (OSD.0)	1.80 TB	No	AL15		OK	
/dev/sde	unknown	LVM, Ceph (OSD.1)	1.80 TB	No	AL14SFB18FQY		OK	
/dev/sdf	unknown	LVM, Ceph (OSD.2)	1.80 TB	No	AL14SFB18FQY		OK	
/dev/sdg	SSD	LVM, Ceph (OSD.3)	1.80 TB	No	AL14SFB18FQY		PASSED	
/dev/sdh	SSD	LVM, Ceph (OSD.4)	1.80 TB	No	AL14SFB18FQY		PASSED	

Error dialog box: Error disk/partition '/dev/sdf' has a holder (500) OK

```
root@proxnode04:~/scripts# ./reset_disks.sh

=== CEPH RESET TOOL ===
1) [LISTER] Voir l'etat des disques
2) [PURGE] Supprimer proprement un OSD
3) [WIPE] Forcer le nettoyage (Anti-Holder)
q) Quitter / h) Aide
Votre choix : 1

--- ETAT DES DISQUES LOCAUX (Node: proxnode04) ---
NOM          TYPE      TAILLE   OSD ID(s)  SERIAL          STATUT
-----
/dev/sda     SSD       447.1G   -          [REDACTED]     LIBRE
/dev/sdb     SSD       447.1G   -          [REDACTED]     LIBRE
/dev/sdc     SSD       1.5T    0,1,2     [REDACTED]     PARTAGE
/dev/sdd     HDD       1.6T    0         [REDACTED]     OSD UNIQUE
/dev/sde     HDD       1.6T    1         [REDACTED]     OSD UNIQUE
/dev/sdf     HDD       1.6T    2         [REDACTED]     OSD UNIQUE
/dev/sdg     SSD       1.7T    3         [REDACTED]     OSD UNIQUE
/dev/sdh     SSD       1.7T    4         [REDACTED]     OSD UNIQUE

=== CEPH RESET TOOL ===
1) [LISTER] Voir l'etat des disques
2) [PURGE] Supprimer proprement un OSD
3) [WIPE] Forcer le nettoyage (Anti-Holder)
q) Quitter / h) Aide
Votre choix :
```

LE WIPE QUI NE WIPE PLUS...2/2

```
if [[ "$confirm" = "CONFIRMER" ]]; then
    log_info "1/4 - Liberation des verrous dmsetup ..."
    for map in $(lsblk -ln -o NAME "/dev/$disk" | grep "ceph"); do
        log_warn "Suppression du mapping : $map"
        dmsetup remove -f "$map" 2>/dev/null
    done

    log_info "2/4 - Suppression des signatures (wipefs) ..."
    wipefs -a "/dev/$disk"

    log_info "3/4 - Destruction des tables GPT (sgdisk) ..."
    sgdisk --zap-all "/dev/$disk"

    log_info "4/4 - Notification du noyau (partprobe) ..."
    partprobe "/dev/$disk"

    log_info "SUCCES : /dev/$disk est vierge."
else
    log_info "Operation annulee."
fi
```

```
root@proxnode04:~/scripts# dmsetup ls
A6onIO-23bx-WusN-l0nL-ihdZ-gR4k-w1pBgB (252:8)
GxwRBX-FyBl-Go3p-K622-Br08-cMLN-2E5hof (252:13)
IdJl4X-CCiu-oJym-2ANr-7dm5-vBLG-Z3PIwk (252:11)
Sw0gFc-A0qJ-VD7z-KvIY-ufNE-TNuR-1HeQc4 (252:14)
XtDKsm-VRIy-6Fug-qc73-nZ0N-owIX-mo0p16 (252:9)
ceph--1a1d450c--19df--400c--bb1f--3d1212f9f8d7-osd--block--c6fcd3e7--0a2f--46ec--9fb3--565be7cd4882 (252:5)
ceph--4c9992fd--cd0c--442c--b9de--51368946e8b9-osd--block--594a7eb5--6546--42cc--9980--3491e5aac627 (252:7)
ceph--6fcf6a6c--25e8--42fa--ae70--aede8a8a2a1e-osd--block--eachf7c3--76ce--430a--951b--cb0ed241556e (252:6)
ceph--932c44fa--9961--4405--91c8--2c319a06522a-osd--db--4c3485b7--ff94--48a0--a00a--239c3e2148dd (252:4)
ceph--932c44fa--9961--4405--91c8--2c319a06522a-osd--db--73d90589--3770--47bb--b4b8--5c3efb0b67fa (252:2)
ceph--932c44fa--9961--4405--91c8--2c319a06522a-osd--db--f4dc19e9--465d--4d7e--bbdc--5cacd2aff733 (252:3)
ceph--db669564--48e4--4fa0--a679--dfe900bbcc31-osd--block--0f7db511--d41f--4a6b--83f3--f2b7b82e16bf (252:1)
ceph--fee3a7ae--7013--498a--b0f4--1674e6741be4-osd--block--eabd58f1--d99f--4270--82b9--83703c6eb827 (252:0)
h1qSzI-UgLn-FODj-KSHe-PKbf-euSz-dy0LLz (252:12)
i0oVbC-GdJQ-WSA0-d1JQ-MYwX-030B-6ceSNW (252:15)
mpatha (252:16)
sI7cdQ-BeQU-iksZ-GBJa-OKmz-Lywt-4dHD8u (252:10)
vg_shared_dell-test_bench_fio (252:17)
vg_shared_dell-vm--100--disk--0 (252:23)
vg_shared_dell-vm--100--disk--0.qcow2 (252:24)
vg_shared_dell-vm--102--disk--0.qcow2 (252:19)
vg_shared_dell-vm--102--disk--1 (252:20)
vg_shared_dell-vm--102--disk--2 (252:21)
vg_shared_dell-vm--106--disk--0.qcow2 (252:22)
vg_shared_dell-vm--109--disk--0.qcow2 (252:25)
root@proxnode04:~/scripts# dmsetup remove -f ceph--4c9992fd--cd0c--442c--b9de--51368946e8b9-osd--block--594a7eb5--6546--42cc--9980--3491e5aac627
```

SORTIE DE 3 OSD... VERSION KAMIKAZ

default											
proxnode09				19.2.3							
osd.11	ssd	bluestore	up / in	19.2.3	1.45549	1.00	0.04	1.46 TIB	0 / 0	128	
osd.10	hdd	bluestore	down / in	19.2.3	1.80109	1.00	9.14	1.80 TIB	0 / 0	18	
osd.9	hdd	bluestore	up / in	19.2.3	1.80109	1.00	9.14	1.80 TIB	0 / 0	175	
osd.8	hdd	bluestore	up / in	19.2.3	1.80109	1.00	9.14	1.80 TIB	0 / 0	170	
proxnode08				19.2.3							
osd.7	ssd	bluestore	up / in	19.2.3	1.45549	1.00	0.08	1.46 TIB	0 / 0	128	
osd.6	hdd	bluestore	down / in	19.2.3	1.80109	1.00	9.13	1.80 TIB	0 / 0	0	
osd.5	hdd	bluestore	up / in	19.2.3	1.80109	1.00	9.14	1.80 TIB	0 / 0	175	
osd.4	hdd	bluestore	up / in	19.2.3	1.80109	1.00	9.14	1.80 TIB	0 / 0	172	
proxnode07				19.2.3							
osd.3	ssd	bluestore	up / in	19.2.3	1.45549	1.00	0.01	1.46 TIB	0 / 0	128	
osd.2	hdd	bluestore	up / in	19.2.3	1.80109	1.00	9.14	1.80 TIB	0 / 0	167	
osd.1	hdd	bluestore	down / in	19.2.3	1.80109	1.00	9.14	1.80 TIB	0 / 0	0	
osd.0	hdd	bluestore	up / in	19.2.3	1.80109	1.00	9.14	1.80 TIB	0 / 0	175	

Health

Status



HEALTH_WARN

	Summary	
	3 osds down	
	Reduced data availability: 63 pgs inactive	
	Degraded data redundancy: 893/2799 objects degra...	

Status

OSDs

	In	Out
Up	9	0
Down	3	0

Total: 12



PGs

	active+clean:	284
	active+undersized:	43
	active+undersized+degraded:	184
	stale+undersized+degraded+peered:	15
	stale+undersized+peered:	3
	undersized+degraded+peered:	55
	undersized+peered:	13

Services

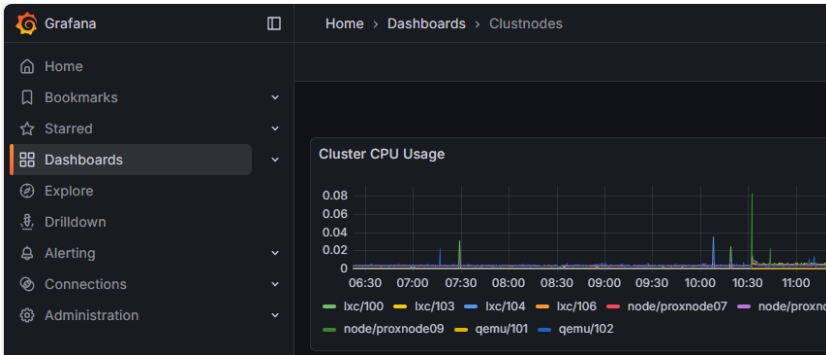
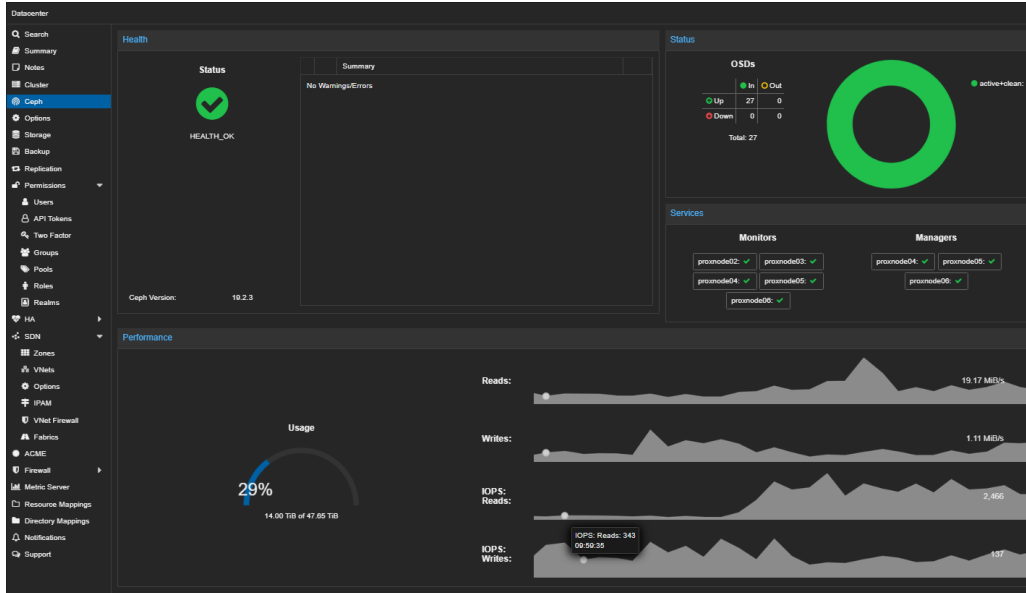
Monitors

Managers

Metadata Servers

BIEN MONITORER

- Dashboards



- Scripts et crontab

```
--- 1. ANALYSE & SURVEILLANCE ---
11) AUDIT COMPLET & RÉPARATION INTERACTIVE
12) Monitoring Live (Standard)
13) Monitoring Live (Latences)
14) Aide Mémoire de Crise

--- 2. URGENCE & SÉCURITÉ ---
21) MODE PANIQUE (Relancer TOUS les OSDs)
22) Geler le Recovery (Set NoOut/NoBackfill)
23) Libérer le Recovery (Unset flags)

--- 3. SIMULATIONS (TESTS) ---
31) Simuler Panne Complète d'OSD (Stop + Out)
32) Forcer un OSD à passer DOWN (Manuel - Stop Service)
33) Forcer un OSD à passer OUT (Manuel - Exclusion)
34) Simuler Panne NOEUD (Arrêt d'un hôte)

--- 4. ACTIONS DE RÉPARATION ---
41) Réintégration Complète (Start + In)
42) Forcer un OSD à passer START (Manuel)
43) Forcer un OSD à passer IN (Manuel)
44) Forcer Deep-Scrub (Réparer PGs)
45) Ajuster Vitesse Recovery (QoS)

--- 5. ÉTAT & DIAGNOSTIC ---
51) Test de Cohérence & Résilience (Détail)
52) Analyser Equilibrage (OSD DF)

q) Quitter
Choix : █
```

```
root@proxmoted04:~/scripts# ./SMART_monitor.sh

== GESTIONNAIRE DE DISQUES PROXMOX/CEPH ==
1) Afficher la liste des disques (lsblk)
2) TESTER TOUS les disques (Scan + Log)
3) ANALYSER 1 seul disque (SMART instantané)
4) TEST ECRITURE/STRESS (Vérifier CRC/Vitesse)
5) VOIR l'historique des résultats
6) LANCER un test SMART complet (Long Self-test)
7) ANALYSE FINE (Journal d'erreurs)
q) Quitter
Choix : 3
Disque : /dev/sde
{
  "date": "2026-04-21 10:09",
  "disk": "/dev/sde",
  "realloc": 0,
  "health": "P:0",
  "temp": "-C",
  "crc": 0,
  "status": "[M]OK"
}
```

```
# Verification état de CEPH
* * * * /bin/bash /root/scripts/check_ceph.sh > /dev/null 2>&1

# Parcourt les disques, extrait le CRC actuel, le compare à la référence, et envoie un mail
03 9 * * * /bin/bash /root/scripts/SMART_monitor.sh --cron
```

- Logs

```
2026-01-30T16:16:55.699197+01:00 proxmoted09.priv.iem.fr kernel sd 0:0:1:0: Power-on or device reset occurred
2026-01-30T16:21:21.701032+01:00 proxmoted09.priv.iem.fr kernel sd 0:0:1:0: Power-on or device reset occurred
2026-01-30T16:23:14.452187+01:00 proxmoted09.priv.iem.fr kernel sd 0:0:1:0: Power-on or device reset occurred
2026-01-30T16:47:07.962114+01:00 proxmoted09.priv.iem.fr kernel sd 0:0:0:0: Power-on or device reset occurred
2026-01-30T16:48:20.448434+01:00 proxmoted09.priv.iem.fr kernel sd 0:0:0:0: Power-on or device reset occurred
2026-02-02T10:32:15.637778+01:00 proxmoted09.priv.iem.fr kernel sd 12:0:0:1: Power-on or device reset occurred
2026-02-02T10:32:15.664765+01:00 proxmoted09.priv.iem.fr kernel sd 14:0:0:1: Power-on or device reset occurred
info@prd-lci-collect2:~$ █
```

BIEN MONITORER

Dell Tech Support
technical_support@help.dell.com

Répondre Répondre à la liste Transférer Archiver Indésirable Supprimer Autres

Pour 28/01/2026, 16:01

Réponse à Dell Tech Support

Support Technique DelleMC - Incident n° / Legacy case n° / Service Tag : / NoPhone
[OTHER] - Multiples Hard Drive/PERC/BP/reporting issue [thread:]

List-ID

[Mon Compte](#)

[Centre de Support](#)

[Pilotes et Téléchargements](#)

[Forums d'aide](#)

Bonjour,

Nous vous remercions d'avoir contacté le Pro Support Dell.

Les éléments collectés confirment la présence d'événements répétés de réinitialisation interne (INTERNAL_DEVICE_RESET) ainsi que de multiples TASK_ABORT_INTERNAL affectant plusieurs disques simultanément. Ces événements entraînent une réduction forcée de la profondeur de file (Queue Depth), suivie de tentatives de restauration automatiques par le contrôleur.

Bien que les disques eux-mêmes ne rapportent pas de défaut matériel individuel, ce comportement indique une instabilité sur la chaîne de communication entre le contrôleur, le backplane et/ou le câblage.

[Ouvrir son incident en ligne](#)

Vous souhaitez ouvrir votre incident exclusivement en ligne ?
C'est désormais possible!

[Demande de Support en Ligne](#)

[Webinaires Techniques](#)

Le support vous propose de participer gratuitement à des webinaires techniques en ligne.

[Consultez le calendrier](#)

LES CH'TIS POOLS

Pool #	Name	Size/min	# of Placement Groups	Optimal # of PGs	Autoscaler Mode	CRUSH Rule (ID)	Used (%)
1	.mgr	3/2	1	n/a	on	replicated_rule (0)	121.14 MiB (0.00%)
2	slow_data_pool	3/2	128	n/a	on	hdd_rule (1)	6.09 TB (44.78%)
3	ssd_data_pool	3/2	128	n/a	on	ssd_rule (2)	6.41 TB (22.33%)
							12.50 TB

Storage 'slow_data_pool' on node 'proxnode04'

Summary | Hour | Maximum | Average

VM Disks | CT Volumes | Permissions

Status

Enabled	Yes
Active	Yes
Content	Disk image, Container
Type	RBD
Usage	47.23% (2.46 TB of 5.21 TB)

HDD: $9 * 1.8\text{Tb} \Rightarrow 5.21\text{ Tb}$
au lieu de **16.2 Tb**

Storage 'ssd_data_pool' on node 'proxnode04'

Summary | Hour | Maximum | Average

VM Disks | CT Volumes | Permissions

Status

Enabled	Yes
Active	Yes
Content	Disk image, Container
Type	RBD
Usage	22.33% (2.35 TB of 10.53 TB)

SSD: $18 * 1.75\text{Tb} \Rightarrow 10.53\text{ Tb}$
au lieu de **31.5 Tb**



PARTIE 4 : RÉSILIENCE ET CRASH TESTS



SCRIPT RÉSILIENCE - MENU

```
--- 1. ANALYSE & SURVEILLANCE ---
11) AUDIT COMPLET & RÉPARATION INTERACTIVE
12) Monitoring Live (Standard)
13) Monitoring Live (Latences)
14) Aide Mémoire de Crise

--- 2. URGENCE & SÉCURITÉ ---
21) MODE PANIQUE (Relancer TOUS les OSDs)
22) Geler le Recovery (Set NoOut/NoBackfill)
23) Libérer le Recovery (Unset flags)

--- 3. SIMULATIONS (TESTS) ---
31) Simuler Panne Complète d'OSD (Stop + Out)
32) Forcer un OSD à passer DOWN (Manuel - Stop Service)
33) Forcer un OSD à passer OUT (Manuel - Exclusion)
34) Simuler Panne NOEUD (Arrêt d'un hôte)

--- 4. ACTIONS DE RÉPARATION ---
41) Réintégration Complète (Start + In)
42) Forcer un OSD à passer START (Manuel)
43) Forcer un OSD à passer IN (Manuel)
44) Forcer Deep-Scrub (Réparer PGs)
45) Ajuster Vitesse Recovery (QoS)

--- 5. ÉTAT & DIAGNOSTIC ---
51) Test de Cohérence & Résilience (Détail)
52) Analyser Équilibrage (OSD DF)

q) Quitter
```

1. Analyse de l'état du Cluster

- `ceph health` : Affiche un résumé rapide de la santé du cluster (HEALTH_OK, WARN, ERR).
- `ceph -s` (ou `status`) : Donne une vue d'ensemble (état du quorum, des PGs, usage data).
- `ceph pg stat` : Affiche les statistiques des Placement Groups (utilisé ici pour le monitoring et la récupération des IDs de PGs).

2. Gestion et Inventaire des OSD (Object Storage Daemons)

- `ceph osd dump` : Récupère l'état détaillé de tous les OSD (utilisé avec `-f json` pour filtrer les états UP/DOWN et IN/OUT).
- `ceph osd ls` : Liste simplement les IDs de tous les OSD du cluster.
- `ceph osd tree` : Affiche la structure hiérarchique du cluster (Hôtes > OSD) et leur statut.
- `ceph osd find <ID>` : Localise sur quel hôte physique se trouve un OSD spécifique.
- `ceph osd perf` : Affiche les performances de latence (commit et apply) pour chaque OSD.
- `ceph osd df tree` : Affiche l'utilisation de l'espace disque et l'équilibrage des données par OSD et par hôte.

SCRIPT RÉSILIENCE - MENU

```
--- 1. ANALYSE & SURVEILLANCE ---
11) AUDIT COMPLET & RÉPARATION INTERACTIVE
12) Monitoring Live (Standard)
13) Monitoring Live (Latences)
14) Aide Mémoire de Crise

--- 2. URGENCE & SÉCURITÉ ---
21) MODE PANIQUE (Relancer TOUS les OSDs)
22) Geler le Recovery (Set NoOut/NoBackfill)
23) Libérer le Recovery (Unset flags)

--- 3. SIMULATIONS (TESTS) ---
31) Simuler Panne Complète d'OSD (Stop + Out)
32) Forcer un OSD à passer DOWN (Manuel - Stop Service)
33) Forcer un OSD à passer OUT (Manuel - Exclusion)
34) Simuler Panne NOEUD (Arrêt d'un hôte)

--- 4. ACTIONS DE RÉPARATION ---
41) Réintégration Complète (Start + In)
42) Forcer un OSD à passer START (Manuel)
43) Forcer un OSD à passer IN (Manuel)
44) Forcer Deep-Scrub (Réparer PGs)
45) Ajuster Vitesse Recovery (QoS)

--- 5. ÉTAT & DIAGNOSTIC ---
51) Test de Cohérence & Résilience (Détail)
52) Analyser Équilibrage (OSD DF)

q) Quitter
```

3. Actions de Maintenance et Résilience

- `ceph osd in <ID>` : Réintègre un OSD dans le cluster pour qu'il recommence à stocker des données.
- `ceph osd out <ID>` : Sort logiquement un OSD du cluster (déclenche la migration des données vers d'autres disques).
- `ceph osd set <flag>` : Active des drapeaux de cluster (utilisé pour `noout` et `nobackfill` afin de geler les mouvements de données).
- `ceph osd unset <flag>` : Désactive les drapeaux précédemment cités pour reprendre un fonctionnement normal.
- `ceph pg deep-scrub <PG_ID>` : Force une vérification bit à bit en profondeur des données sur un groupe de placement spécifique.

Le script ne se contente pas de parler à Ceph, il utilise aussi le protocole SSH pour agir sur les services distants via :

- `systemctl start/stop/restart ceph-osd@<ID>` (gestion individuelle).
- `systemctl stop/restart ceph-osd.target` (gestion de tous les OSD d'un nœud d'un coup).

CAS CONCRET : CHANGER UN DISQUE 1/4

☆	ALERTE : Etat Cluster Ceph (ALERTE - 1/10)	○	ceph-ale	🔔	12/04/2026, 12:52
☆	ALERTE : Etat Cluster Ceph (ALERTE - 2/10)	○	ceph-ale	🔔	12/04/2026, 12:53
☆	ALERTE : Etat Cluster Ceph (ALERTE - 3/10)	○	ceph-ale	🔔	12/04/2026, 12:54
☆	ALERTE : Etat Cluster Ceph (ALERTE - 4/10)	○	ceph-ale	🔔	12/04/2026, 12:55
☆	ALERTE : Etat Cluster Ceph (ALERTE - 5/10)	○	ceph-ale	🔔	12/04/2026, 12:56
☆	ALERTE : Etat Cluster Ceph (ALERTE - 6/10)	○	ceph-ale	🔔	12/04/2026, 12:57

 ceph-alert@ceph-alert@
Pour

ALERTE : Etat Cluster Ceph (ALERTE - 1/10)

[Répondre](#)

Type d'envoi : ALERTE

Attention, le cluster Ceph est en etat : HEALTH_WARN 1 osds down; Degraded data redundancy: 158569/2542770 objects degraded (6.236%), 64 pgs degraded
Tentative de detection n°1 sur 10

```
-----
cluster:
  id: 8a503189-0e86-4862-9e3a-39101f829a59
  health: HEALTH_WARN
         1 osds down
         Degraded data redundancy: 158569/2542770 objects degraded (6.236%), 64 pgs degraded
```

```
services:
  mon: 3 daemons, quorum proxnode04,proxnode05,proxnode06 (age 3w)
  mgr: proxnode04(active, since 3d), standbys: proxnode05, proxnode06
  osd: 15 osds: 14 up (since 8s), 15 in (since 5w)
```

```
data:
  pools: 3 pools, 257 pgs
  objects: 847.59k objects, 3.2 TiB
  usage: 11 TiB used, 16 TiB / 27 TiB avail
  pgs: 158569/2542770 objects degraded (6.236%)
      193 active+clean
      64 active+undersized+degraded
```

```
io:
  client: 34 MiB/s rd, 2.3 MiB/s wr, 4.02k op/s rd, 261 op/s wr
```

CAS CONCRET : CHANGER UN DISQUE 2/4

The image shows two panels from the Ceph dashboard. The left panel, titled 'Health', shows a yellow warning icon and the text 'HEALTH_WARN'. A red box highlights a summary message: '1 daemons have recently crashed' with a sub-message 'osd.14 crashed on host proxnode06 at 2026-04-12T10:51:49.601720z'. The right panel, titled 'Status', shows OSDs and PGs. A red box highlights the 'Down' count for OSDs, which is '1'. A large green donut chart indicates that 14 OSDs are 'Up' and 0 are 'Down'.

A window titled 'Disque 1' showing disk information: 'Inconnu', '960 Mo', and 'Non initialisé'. A red box highlights the text '960 Mo' and 'Non alloué'.


A Windows 'Propriétés de' dialog box for a 'SAMSUNG MZ7LM1T9HMJP-000 SCSI Disk Device'. The 'Pilote' tab is selected, showing details like 'Fournisseur du pilote : Microsoft', 'Date du pilote : 21/06/2006', and 'Version du pilote : 10.0.26100.7705'. Buttons for 'Détails du pilote', 'Mettre à jour le pilote', 'Restaurer le pilote', 'Désactiver l'appareil', and 'Désinstaller l'appareil' are visible.

Name	Class	OSD Type	Status	Version	weight	reweight	Used (%)	Total	Apply/Commit Latency (ms)	PGs	
default				19.2.3							
proxnode06	19.2.3							19.2.3			
osd.14	ssd	bluestore	down / out	19.2.3	1.7466	0.00	0.00	1.00 KiB	0 / 0	0	
osd.13	ssd	bluestore	up / in	19.2.3	1.7466	1.00	68.52	1.75 TiB	1 / 1	128	
osd.12	hdd	bluestore	up / in	19.2.3	1.80109	1.00	44.62	1.80 TiB	1 / 1	41	
osd.11	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.85	1.80 TiB	1 / 1	45	
osd.10	hdd	bluestore	up / in	19.2.3	1.80109	1.00	45.45	1.80 TiB	1 / 1	43	
proxnode05	19.2.3							19.2.3			
osd.9	ssd	bluestore	up / in	19.2.3	1.7466	1.00	32.18	1.75 TiB	0 / 0	60	
osd.8	ssd	bluestore	up / in	19.2.3	1.7466	1.00	36.38	1.75 TiB	1 / 1	68	
osd.7	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.90	1.80 TiB	1 / 1	45	
osd.6	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.08	1.80 TiB	3 / 3	44	
osd.5	hdd	bluestore	up / in	19.2.3	1.80109	1.00	42.94	1.80 TiB	1 / 1	40	
proxnode04	19.2.3							19.2.3			
osd.4	ssd	bluestore	up / in	19.2.3	1.7466	1.00	36.37	1.75 TiB	0 / 0	68	
osd.3	ssd	bluestore	up / in	19.2.3	1.7466	1.00	32.20	1.75 TiB	0 / 0	60	
osd.2	hdd	bluestore	up / in	19.2.3	1.80109	1.00	41.88	1.80 TiB	1 / 1	38	
osd.1	hdd	bluestore	up / in	19.2.3	1.80109	1.00	48.05	1.80 TiB	1 / 1	46	
osd.0	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.99	1.80 TiB	1 / 1	45	

CAS CONCRET : CHANGER UN DISQUE 3/4

default										
proxnode06										
osd.13	ssd	bluestore	up ● / in ●	19.2.3	1.7466	1.00	68.51	1.75 TiB	0 / 0	128
osd.12	hdd	bluestore	up ● / in ●	19.2.3	1.80109	1.00	44.62	1.80 TiB	1 / 1	41
osd.11	hdd	bluestore	up ● / in ●	19.2.3	1.80109	1.00	47.85	1.80 TiB	0 / 0	45
osd.10	hdd	bluestore	up ● / in ●	19.2.3	1.80109	1.00	45.45	1.80 TiB	1 / 1	43

Health

Status

HEALTH_WARN

Summary
1 daemons have recently crashed
osd.14 crashed on host proxnode06 at 2026-04-12T10:51:49.681728Z

Ceph Version: 19.2.3

Status

OSDs

	● In	● Out
● Up	14	0
● Down	0	0
Total: 14		

PGs

● active+clean:	252
● active+remapped+backfill_wait:	4
● active+remapped+backfilling:	1

Services

Monitors	Managers	Metadata Servers
proxnode04: ✓ proxnode05: ✓ proxnode06: ✓	proxnode04: ✓ proxnode05: ✓ proxnode06: ✓	

Performance

Usage
44%
11.06 TiB of 24.84 TiB

Recovery/Rebalance: 49.69 MiB/s
99.54% (49.89 MiB/s - 15m 42.5s left)

Reads: 481.21 KiB/s

Writes: 3.36 MiB/s

IOPS: Reads: 49

IOPS: Writes: 172

CAS CONCRET : CHANGER UN DISQUE 4/4

Node 'proxnode06'

Search

Reload Show S.M.A.R.T. values Initialize Disk with GPT Wipe Disk

Device	Type	Usage	Size	GPT	Model	Serial
/dev/sda	SSD	partitions	480.10 GB	Yes	[REDACTED]	[REDACTED]
/dev/sda1	partition	BIOS boot	1.03 MB	Yes	[REDACTED]	[REDACTED]
/dev/sda2	partition	EFI	1.07 GB	Yes	[REDACTED]	[REDACTED]
/dev/sda3	partition	ZFS	479.03 GB	Yes	[REDACTED]	[REDACTED]
/dev/sdb	SSD	partitions	480.10 GB	Yes	[REDACTED]	[REDACTED]
/dev/sdb1	partition	BIOS boot	1.03 MB	Yes	[REDACTED]	[REDACTED]
/dev/sdb2	partition	EFI	1.07 GB	Yes	[REDACTED]	[REDACTED]
/dev/sdb3	partition	ZFS	479.03 GB	Yes	[REDACTED]	[REDACTED]
/dev/sdc	SSD	LVM, Ceph (DB)	1.60 TB	No	INTERT	[REDACTED]
/dev/sdd	unknown	LVM, Ceph (OSD.10)	1.80 TB	No	[REDACTED]	[REDACTED]
/dev/sde	unknown	LVM, Ceph (OSD.11)	1.80 TB	No	AL14SEB16EQY	[REDACTED]
/dev/sdf	unknown	LVM, Ceph (OSD.12)	1.80 TB	No	[REDACTED]	[REDACTED]
/dev/sdg	SSD	LVM, Ceph (OSD.13)	1.80 TB	No	[REDACTED]	[REDACTED]
/dev/sdm	SSD	No	1.80 TB	No	[REDACTED]	[REDACTED]

Confirm

Are you sure you want to wipe /dev/sdm? All data on the device will be lost!

Type: SSD
Usage: No
Size: 1.80 TB
Serial: [REDACTED]

Yes No

Reload Create OSD Manage Global Flags

Name	Class	OSD Type	Status	Version	weight	reweight	Used (%)	Total
default				19.2.3				
proxnode06				19.2.3				
osd.13	ssd	bluestore	up / in	19.2.3	1.7466	1.00	68.52	1.75 TB
osd.12	hdd	bluestore	up / in	19.2.3	1.80109	1.00	44.62	1.80 TB
osd.11	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.85	1.80 TB
osd.10	hdd	bluestore	up / in	19.2.3	1.80109	1.00	45.45	1.80 TB
proxnode05				19.2.3				
osd.9	ssd	bluestore	up / in	19.2.3	1.7466	1.00	33.28	1.75 TB
osd.8	ssd	bluestore	up / in	19.2.3	1.7466	1.00	35.29	1.75 TB
osd.7	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.90	1.80 TB
osd.6	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.90	1.80 TB
osd.5	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.90	1.80 TB
proxnode04				19.2.3				
osd.4	ssd	bluestore	up / in	19.2.3	1.7466	1.00	33.28	1.75 TB
osd.3	ssd	bluestore	up / in	19.2.3	1.7466	1.00	35.29	1.75 TB
osd.2	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.90	1.80 TB
osd.1	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.90	1.80 TB
osd.0	hdd	bluestore	up / in	19.2.3	1.80109	1.00	47.90	1.80 TB

Create: Ceph OSD

Disk: /dev/sdm DB Disk: use OSD disk

DB size (GiB): Automatic

Encrypt OSD: WAL Disk: use OSD/DB disk

Device Class: auto detect WAL size (GiB): Automatic

Note: Ceph is not compatible with disks backed by a hardware RAID controller. For details see the [reference documentation](#).

Help Advanced Create

LVM-Thin Directory ZFS Ceph

Configuration Monitor OSD CephFS Pools Log Replication Task History Subscription

Performance

Usage

41%

11.05 TiB of 28.00 TiB

Recovery/Rebalance: 147.72 MiB/s

91.00% (147.72 MiB/s - 1h 31m 57.5s left)

Reads: Writes: IOPS: Reads: IOPS: Writes:

1h48 après :

Health Status

Summary No Warnings/Errors

HEALTH_OK

OSDs

Summary

● In ● Out

● Up 15 0

● Down 0 0

Total: 15

PGs

● active+clean. 257

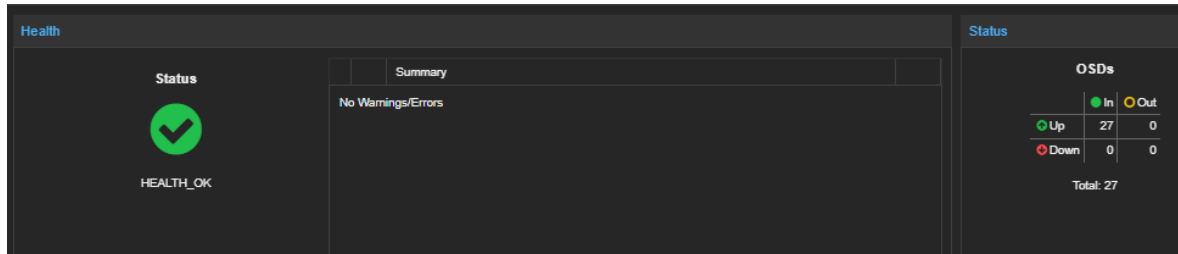
CAS CONCRET : REBOOT D'UN NŒUD 1/5

```
root@proxnode04:~/scripts# ./prep-reboot.sh --help
≡ AIDE : GESTION REBOOT CEPH SUR PROXMOX ≡
Ce script sécurise et automatise la maintenance d'un nœud Ceph.

ÉTAPES POUR UN REBOOT RÉUSSI :
1. Lancer --prepare : Pose les drapeaux de sécurité pour éviter le
   rebalancement des données et arrête proprement les services locaux.
2. Lancer --reboot : Redémarre physiquement le serveur.
3. Lancer --finish : Vérifie le réseau, relance les services et
   retire les drapeaux de sécurité une fois que tout est stable.
4. Lancer --repair : (Optionnel) Si un OSD reste DOWN, force le
   reset de Systemd et réintègre l'OSD dans le cluster.

COMMANDES CLI DISPONIBLES :
--check      Vérifie la visibilité du réseau public Ceph.
--prepare    Prépare le nœud avant l'arrêt.
--reboot     Exécute 'reboot' après confirmation.
--finish     Finalise la maintenance (automatique après reboot).
--repair     Lance la RÉPARATION EXTRÊME (force le réveil des OSD).
--help      Affiche cette aide.
```

```
root@proxnode04:~/scripts# ./prep-reboot.sh --check
>>> Vérification du réseau public Ceph...
[OK] Réseau Ceph (172.17.5.4/24) joignable.
```



1. Mode Préparation (--prepare)

Le but est de figer le cluster pour qu'il ne s'excite pas pendant que le nœud s'éteint.

- Bloquer les mouvements de données :

```
ceph osd set noout
ceph osd set noscrub
ceph osd set nodeep-scrub
```
- Arrêter les services proprement (sur le nœud local) :

```
systemctl stop ceph-osd.target
systemctl stop ceph-mon.target
systemctl stop ceph-mgr.target
```

2. Redémarrage (--reboot)

- `reboot` : Redémarre physiquement la machine.

3. Finalisation après redémarrage (--finish)

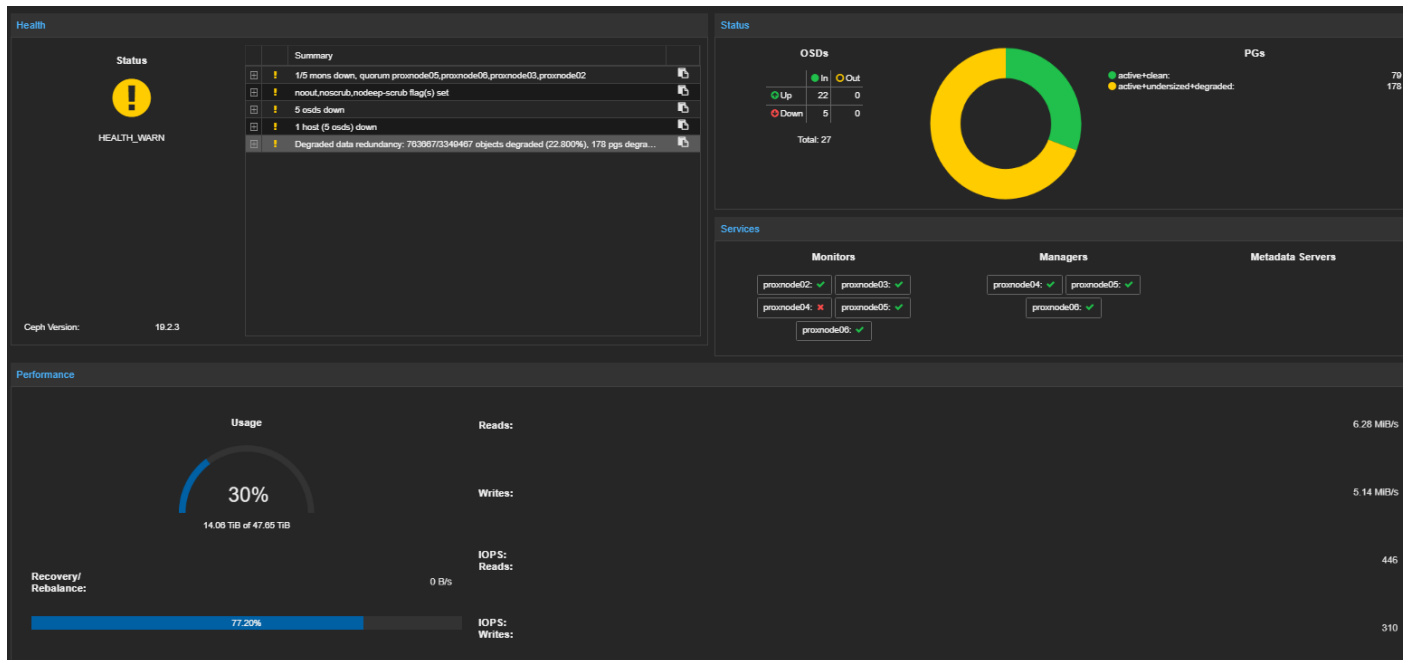
L'objectif est de s'assurer que tout est stable avant de réactiver les mouvements de données.

- Redémarrage des services :

```
systemctl start ceph-mon.target ceph-mgr.target ceph-osd.target.
```
- Surveillance : * Utilise `ceph osd tree` pour isoler les OSD du serveur local et vérifier leur statut (`up` ou `down`).
- Retrait des drapeaux :
 - `ceph osd unset noout, noscrub, nodeep-scrub` (rend au cluster son comportement normal).
- Vérification finale :
 - `ceph -s` (affiche l'état de santé global).

CAS CONCRET : REBOOT D'UN NŒUD 2/5

```
root@proxnode04:~/scripts# ./prep-reboot.sh --prepare
>>> ETAPE 1 : Préparation du nœud (Sécurisation)
Santé actuelle : HEALTH_OK
Activation des drapeaux de sécurité (noout, noscrub, nodeep-scrub)...
noout is set
noscrub is set
nodeep-scrub is set
Arrêt propre des services locaux (OSD, MON, MGR)...
[SUCCES] Nœud prêt. Le cluster ne déclenchera pas de rebalancement.
```



Node	OSD	Type	Storage	Status
proxnode06	osd.14	ssd	bluestore	up
	osd.13	ssd	bluestore	up
	osd.12	hdd	bluestore	up
	osd.11	hdd	bluestore	up
	osd.10	hdd	bluestore	up
proxnode05	osd.9	ssd	bluestore	up
	osd.8	ssd	bluestore	up
	osd.7	hdd	bluestore	up
	osd.6	hdd	bluestore	up
	osd.5	hdd	bluestore	up
proxnode04	osd.4	ssd	bluestore	down
	osd.3	ssd	bluestore	down
	osd.2	hdd	bluestore	down
	osd.1	hdd	bluestore	down
	osd.0	hdd	bluestore	down
proxnode03	osd.18	ssd	bluestore	up
	osd.17	ssd	bluestore	up
	osd.16	ssd	bluestore	up
	osd.15	ssd	bluestore	up
	proxnode02	osd.22	ssd	bluestore
osd.21		ssd	bluestore	up
osd.20		ssd	bluestore	up
osd.19		ssd	bluestore	up
proxnode01		osd.26	ssd	bluestore
	osd.25	ssd	bluestore	up
	osd.24	ssd	bluestore	up
	osd.23	ssd	bluestore	up

CAS CONCRET : REBOOT D'UN NŒUD 3/5

```
root@proxnode04:~/scripts# ./prep-reboot.sh --reboot
Reboot en cours ...
root@proxnode04:~/scripts#
Remote side unexpectedly closed network connection
```

PROXMOX Virtual Environment 9.1.7

Server View [Settings] Node 'proxnode05'

- Datacenter (clustnodes)
 - proxnode01
 - proxnode02
 - proxnode03
 - proxnode04
 - proxnode05**
 - proxnode06
 - Ad_jern-fr-Linux-Serveur
 - Ad_jern_fr-WINServer
 - Serveurs_apache
 - Serveurs_nginx

Search, Summary, Notes, Shell, System, Network, Certificates, DNS, Hosts

Health

Status: HEALTH_WARN

Summary

- 1/5 mons down, quorum proxnode05,proxnode06,proxnode03,proxnode02
- noout_noscrub,nodeep-scrub flag(s) set
- 5 osds down
- 1 host (5 osds) down
- Degraded data redundancy: 763672/3340482 objects degraded (22.800%), 178 pgs degra...

Status

OSDs

	In	Out
Up	22	0
Down	5	0

Total: 27

PGs: 79 / 178

- active+clean: 79
- active+undersized+degraded: 0

Services

Monitors: proxnode02: ✓, proxnode03: ✓, proxnode04: ✗, proxnode05: ✓, proxnode06: ✓

Managers: proxnode04: ?, proxnode05: ✓, proxnode06: ✓

Metadata Servers: (empty)

Performance

Usage: 30% (14.06 TiB of 47.85 TiB)

Recovery/Rebalance: 77.20%

Reads: 74.80 MiB/s

Writes: 3.25 MiB/s

IOPS: Reads: 628, Writes: 161

CAS CONCRET : REBOOT D'UN NŒUD 4/5

- ALERTE : Etat Cluster Ceph (ALERTE - 1/10) ceph-aler 10:30
- ALERTE : Etat Cluster Ceph (ALERTE - 2/10) ceph-aler 10:31
- ALERTE : Etat Cluster Ceph (ALERTE - 3/10) ceph-aler 10:32
- ALERTE : Etat Cluster Ceph (ALERTE - 4/10) ceph-aler 10:33
- ALERTE : Etat Cluster Ceph (ALERTE - 5/10) ceph-aler 10:34

ALERTE : Etat Cluster Ceph (ALERTE - 6/10) ceph-aler 10:35

ceph-aler

Pour

ALERTE : Etat Cluster Ceph (ALERTE - 6/10)

Type d'envoi : ALERTE
Attention, le cluster Ceph est en etat : HEALTH_WARN 1/5 mons down, quorum proxnode05,proxnode06,proxnode03,proxnode02; noout,noscrub,nodeep-scrub flag(s) set; 5 osds down; 1 host (5 osds) down; Degraded data redundancy: 763672/3349482 objects degraded (22.800%), 178 pgs degraded, 178 pgs undersized
Tentative de detection n°6 sur 10

cluster:
id: 8a503189-0e86-4862-9e3a-39101f829a59
health: HEALTH_WARN
1/5 mons down, quorum proxnode05,proxnode06,proxnode03,proxnode02;
noout,noscrub,nodeep-scrub flag(s) set;
5 osds down;
1 host (5 osds) down;
Degraded data redundancy: 763672/3349482 objects degraded (22.800%), 178 pgs degraded, 178 pgs undersized

Node 'proxnode05'

Health

Status: HEALTH_WARN

Summary

- noout,noscrub,nodeep-scrub flag(s) set
- 5 osds down
- 1 host (5 osds) down
- Degraded data redundancy: 763672/3349482 objects degraded (22.800%), 178 pgs degraded, 178 pgs undersized
- 1/5 mons down, quorum proxnode05,proxnode06,proxnode03,proxnode02

Ceph Version: 10.2.3

got timeout (500)

Reboot Shutdown

Status

OSDs

	In	Out
Up	22	0
Down	5	0

Total: 27

active+clean: (green)
active+undersized+degraded: (yellow)

Services

Monitors

proxnode02: ✖	proxnode03: ✔
proxnode04: ✔	proxnode05: ✔
proxnode06: ✖	

Managers

proxnode04: ?	proxnode05: ✔
proxnode06: ✔	

CAS CONCRET : REBOOT D'UN NŒUD 5/5

```
root@proxnode04:~/scripts# ./prep-reboot.sh --finish
>>> ÉTAPE 3 : Finalisation et stabilisation ...
>>> Vérification du réseau public Ceph ...
[OK] Réseau Ceph [redacted] joignable.
Redémarrage des services Ceph ...

>>> ATTENTE CRITIQUE : Stabilisation des OSD locaux ...

[!] BLOCAGE : 4 OSD(s) encore DOWN : osd.0 osd.2 osd.3 osd.4
Attente de l'initialisation (Tentative 1/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 4 OSD(s) encore DOWN : osd.0 osd.2 osd.3 osd.4
Attente de l'initialisation (Tentative 2/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 1 OSD(s) encore DOWN : osd.0
Attente de l'initialisation (Tentative 3/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 1 OSD(s) encore DOWN : osd.0
Attente de l'initialisation (Tentative 4/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 1 OSD(s) encore DOWN : osd.0
Attente de l'initialisation (Tentative 5/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 1 OSD(s) encore DOWN : osd.0
Attente de l'initialisation (Tentative 6/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 1 OSD(s) encore DOWN : osd.0
Attente de l'initialisation (Tentative 7/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 1 OSD(s) encore DOWN : osd.0
Attente de l'initialisation (Tentative 8/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 1 OSD(s) encore DOWN : osd.0
Attente de l'initialisation (Tentative 9/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 1 OSD(s) encore DOWN : osd.0
Attente de l'initialisation (Tentative 10/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 1 OSD(s) encore DOWN : osd.0
Attente de l'initialisation (Tentative 11/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

[!] BLOCAGE : 1 OSD(s) encore DOWN : osd.0
Attente de l'initialisation (Tentative 12/12) ...
10 ... 9 ... 8 ... 7 ... 6 ... 5 ... 4 ... 3 ... 2 ... 1 ...

ALERTE : Délai dépassé. OSD toujours bloqués : osd.0
CONSEIL : Lancez l'option 5 (Repair) dans un autre terminal.
Voulez-vous quand même libérer les flags (noout) ? (y/N)
```

```
root@proxnode04:~/scripts# ./prep-reboot.sh --repair
>>> ÉTAPE 4 : Lancement de la RÉPARATION EXTRÊME ...
Traitement OSD.0 en cours ...
- OSD.0 : Reset effectué.
>>> Tentatives terminées.
```

Health

Status: HEALTH_WARN

Summary:

- 1/5 mons down, quorum proxnode04,proxnode05,proxnode03,proxnode02
- 1 daemons have recently crashed
- 33 slow ops, oldest one blocked for 402 sec, mon.proxnode00 has slow ops

Ceph Version: 19.2.3

Performance

Usage: 30% (14.08 TiB of 47.95 TiB)

Reads: 4.80 MB/s

Writes: 64.93 MB/s

IOPS: 74

IOPS: 521

Status

OSDs: 27 Up, 0 In, 0 Out, 0 Down

PGs: 253

Services:

- Monitors: proxnode02, proxnode03, proxnode04, proxnode05, proxnode06
- Managers: proxnode04, proxnode05, proxnode06
- Metadata Servers: proxnode06

Health

Status: HEALTH_OK

Summary: No Warnings/Errors

Ceph Version: 19.2.3

Status

OSDs: 27 Up, 0 In, 0 Out, 0 Down

PGs: 257

Services:

- Monitors: proxnode02, proxnode03, proxnode04, proxnode05, proxnode06
- Managers: proxnode04, proxnode05, proxnode06
- Metadata Servers: proxnode06

```

root@proxnode04:~/scripts# ./benchmark_disk.sh --help
Usage: ./benchmark_disk.sh [OPTION]

Actions Globales :
-a, --auto          LANCE TOUT : Initialisation + Tous les Benchmarks + Tableau final
-c, --compare       Affiche le tableau comparatif des derniers résultats enregistrés
-i, --init          Prépare et mappe les volumes de test (ZFS, LVM, RBD)
-v, --view          Affiche l'état de présence/mapping des volumes de test
-clean, --clean     Supprime tous les volumes de test et réinitialise les logs
-h, --help         Affiche cette aide

Benchmarks Individuels :
-z, --zfs          Test ZFS Local (ZVOL)
-s, --san          Test SAN Dell (LVM sur Multipath)
-cs, --ceph-ssd    Test Ceph SSD Pool (Chiffré)
-ch, --ceph-hdd    Test Ceph HDD Pool (Slow)
root@proxnode04:~/scripts# ./benchmark_disk.sh --auto
>>> Préparation des stockages ...
/dev/rbd0
/dev/rbd1
Terminé.
Test ZFS_LOCAL (/dev/zvol/rpool/data/test_bench_fio) ... OK
Test SAN_DELL (/dev/vg_shared_dell/test_bench_fio) ... OK
Test CEPH_SSD (/dev/rbd0) ... OK
Test CEPH_HDD (/dev/rbd1) ... OK

===== TABLEAU COMPARATIF FINAL =====
STOCKAGE      | IOPS READ   | IOPS WRITE  | LATENCE AVG | DÉBIT
-----
ZFS_LOCAL     | 159545      | 53320       | 0.30ms      | 831.5MB/s
CEPH_SSD      | 11634       | 3892        | 3.40ms      | 60.6MB/s
CEPH_HDD      | 4042        | 1345        | 7.34ms      | 21.0MB/s
SAN_DELL      | 4032        | 1343        | 12.18ms     | 21.0MB/s
=====

```

TEST RAPIDITÉ DES STORAGES

Local-ZFS (nvme) – NON crypté

Pool CEPH HDD 10000K - crypté

Pool CEPH SDD - crypté

Baie DELL SC4020 (multipath iscsi) –
NON crypté

Proxmox 9 + Ceph Squid :

- Reprise de contrôle et gain financier
- Gagner en résilience, 1 nœud OUT sans problème
- Eliminer le SPOF
- Simple pour rajouter de la puissance et de l'espace partagé
- Scripts, logs, dashboard et mail pour administrer simplement

Futur :

- Amélioration de performance (règle CRUSH)
- Amélioration de surveillance plus graphique et centralisée (actuellement il y a 18 scripts bash)
- Regarder l'erasure coding, ceph et S3...

Caractéristique	Réplication (x3)	Erasure Coding (ex: 4+2)
Utilisation Espace	33% efficace (200% de surplus)	66% efficace (50% de surplus)
Coût Stockage	Très élevé	Réduit significativement
Performance CPU	Faible (simple copie)	Élevée (calculs mathématiques)
Performance IOPS	Excellente	Plus lente (pénalité en écriture)

PROXMOX AUTOINSTALL



CERTIFICAT DE COMPÉTENCE

Grade : Installeur Proxmox Confirmé

Ce certificat est décerné à :

SAMUEL TENON

En reconnaissance de son travail exceptionnel dans l'installation et la configuration du système Proxmox VE. Grâce à sa maîtrise technique et à son implication, **Samuel Tenon** a procédé avec succès à la réinstallation de **40 serveurs Proxmox**, démontrant ainsi ses compétences avancées en gestion d'infrastructures virtuelles et en automatisation système.

Proxmox Installer – Niveau Expert

24 avril 2024

Signature :

Responsable technique / Supérieur hiérarchique



Sur un nœud Proxmox

- 1) Télécharger l'iso proxmox
- 2) Installer les paquet nécessaire:
apt install proxmox-auto-install-assistant
- 3) Créer 2 fichiers de conf: *answer.toml* et *firstboot.sh*

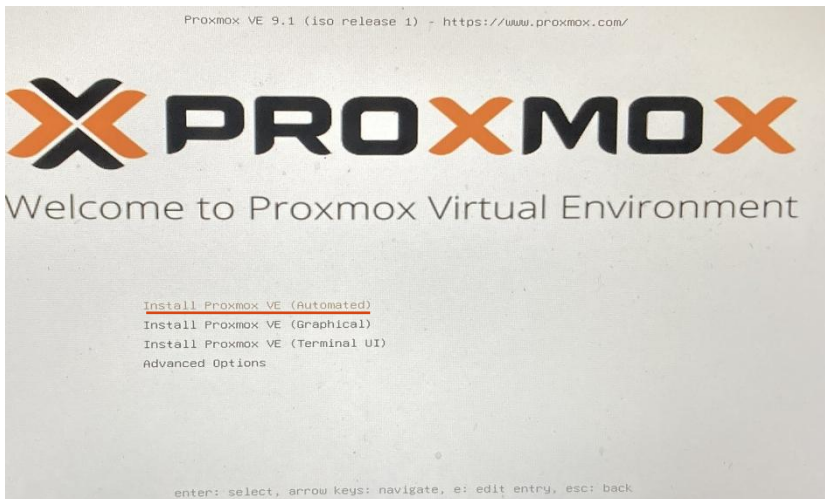
4) Créer le nouvel iso avec ces 2 fichiers:

proxmox-auto-install-assistant prepare-iso proxmox-ve_9.1-1.iso --fetch-from iso --answer-file answer.toml --on-first-boot firstboot.sh

5) Copier le fichier résultat proxmox-ve_9.1-1-auto-from-iso.iso vers le disque USB

6) Démarrer un serveur avec cette iso et 5-7 minutes après le serveur est opérationnel sans intervention

-> bien pour le PRA !!!



Exemples answer.toml

Serveurs HP

```
[global]
keyboard = "fr"
country = "fr"
fqdn = "proxnode01"
mailto = " "
timezone = "Europe/Paris"
root-password-hashed = "$y$j9"
root-ssh-keys = [
  "ssh-rsa AAAAB3NzaC"
  "ssh-rsa AAAAB"
]

[network]
source = "from-answer"
cidr = "/24"
dns = " "
gateway = " "
filter.ID_NET_NAME_MAC = "*5"

[disk-setup]
filesystem = "ext4"
disk-list = ["nvme0n1"]

[first-boot]
source = "from-iso"
```

Serveurs DELL

```
[global]
keyboard = "fr"
country = "fr"
fqdn = "proxnode04"
mailto = " "
timezone = "Europe/Paris"
root-password-hashed = "$y$j9"
root-ssh-keys = [
  "ssh-rsa AAAAB3Nza"
  "ssh-rsa AAAAB3N"
]

[network]
source = "from-answer"
cidr = "/24"
dns = " "
gateway = " "
filter.ID_NET_NAME_MAC = "*d"

[disk-setup]
filesystem = "zfs"
zfs.raid = "raid1"
zfs.ashift = 12
zfs.checksum = "on"
zfs.compress = "lz4"
zfs.copies = 1
zfs.arc_max = 16384
disk_list = ["sda", "sdb"]

[first-boot]
source = "from-iso"
```

Exemples firstboot.sh

```
#!/bin/bash

# Arrêt immédiat en cas d'erreur
set -e

# 1. RÉSEAU (BONDING ET BRIDGES)
NETWORK_FILE="/etc/network/interfaces"
cp "$NETWORK_FILE" "${NETWORK_FILE}.bak"

cat > "$NETWORK_FILE" << EOL
##### CONFIG AUTO FOUR IEMV #####
auto lo
iface lo inet loopback

iface idrac inet manual

auto eno1
iface eno1 inet static
    address [REDACTED]/24
#Corosync1

auto eno2
iface eno2 inet static
    address [REDACTED]/24
#Corosync2
```

```
auto eno3
iface eno3 inet manual

auto eno4
iface eno4 inet manual

auto enp3s0f0np0
iface enp3s0f0np0 inet manual

auto enp3s0f1np1
iface enp3s0f1np1 inet manual

auto enp129s0f0np0
iface enp129s0f0np0 inet manual

auto enp129s0f1np1
iface enp129s0f1np1 inet manual

auto enp130s0f0np0
iface enp130s0f0np0 inet manual

auto enp130s0f1np1
iface enp130s0f1np1 inet manual

auto bond0
iface bond0 inet manual
    bond-slaves eno3 eno4
    bond-miimon 100
    bond-mode 802.3ad
    bond-xmit-hash-policy layer2+3
```

```
auto vubr2
iface vubr2 inet static
    address [REDACTED]/24
    bridge-ports bond2
    bridge-stp off
    bridge-fd 0
    mtu 9000
#Flux CEPH
auto vubr3
iface vubr3 inet static
    address [REDACTED]/24
    bridge-ports bond3
    bridge-stp off
    bridge-fd 0
    mtu 9000
#Flux SC4020-IP1

auto vubr4
iface vubr4 inet static
    address [REDACTED]/24
    bridge-ports bond4
    bridge-stp off
    bridge-fd 0
    mtu 9000
#Flux SC4020-IP2

source /etc/network/interfaces.d/*
EOL

systemctl restart networking

# 2. CHRONY (NTP)
cat > /etc/chrony/chrony.conf << 'EOL'
server nto [REDACTED] iburst
server [REDACTED] iburst
driftfile /var/lib/chrony/chrony.drift
makestep 1.0 3
rtcsync
EOL
systemctl restart chrony

# 3. BASHRC ROOT
BASHRC="/root/.bashrc"
{
    echo "export LS_OPTIONS='--color=auto'"
    echo "eval \"\${dircolors}\""
    echo "alias ls='ls \${LS_OPTIONS}'"
    echo "alias ll='ls \${LS_OPTIONS} -l'"
    echo "alias l='ls \${LS_OPTIONS} -lA'"
} >> "$BASHRC"

# 6. RESTAURATION DU RÉPERTOIRE /root/scripts
mkdir -p /root/scripts
```