Retour sur le stockage Ceph géré par proxmox

Capitoul 29 février 2024

Laurent GUERBY
IMT Mines Albi

Situation initiale 2018

- Chassis Dell M1000
- VMware
- Netapp FAS2552 avec DS2246 24 SAS 1TB et DS4246 24 HDD 4TB, total 100 TB (4 SSD 300 GB)
- Netapp NFS (Vmware) et CIFS
- Dell R730xd avec Atempo TINA et robot bande LTO7 Dell TL2000
- Pas de PCA/PRA

Appliance et obsolescence programmée

- Fin de support du netapp initialement prévu 2022 étendu 2023
- Sans mise à jour de sécurité de l'éditeur difficile de laisser en production un matériel qui pourtant fonctionne encore parfaitement
- Autres appliances comme notre firewall Stormshield et routeur de bordure CISCO qu'il faudrait dupliquer et maintenir a jour pour un PCA/PRA
- Stratégie migration progressive vers de l'hyperconvergé tout virtualisé sans appliance et mise en place d'un PCA/PRA

Matériel 2019-2023

- 5 Dell R640 512G RAM 12 emplacements 2.5 4x10G
- 2 HPE DL385 512/1024G RAM 16 emplacements 2.5 4x10/25G
- Réutilisation 4 lames FX2 pour CPU (28c)/RAM (256G)
- Cluster proxmox et ceph avec achat progressif de SSD
- SSD DC 2023 100 EUR HT/TB Intel puis Samsung PM893 7.68 TB (senetic 692 EUR HT ~ 100 EUR/TB)
- Note : tiroirs compatible Dell « workdone » et tiroirs via les SSD les moins chers du marché matinfo pour HPE

Sauvegarde et PCA/PRA 2022-2023

- 2 HPE DL385 avec 12xSSD 8TB PVE+PBS a Albi
- 1 HPE DL385 2TB RAM 16xNVME/SAS 7.68 TB PVE+PBS envoyé a nos collègues IMT Evry pour PCA/PRA
- ZFS RAIDZ3 9+3 a Albi et 10+3 a Evry
- « Full flash » pas de rotationnel dans l’infra
- Robot de bande testé avec PBS pour faire le offline, attente de cartes PCIe HBA SAS externe HPE pour la production

Logiciel

- Ceph, logiciel libre de stockage distribué auto réparant
- https://telemetry-public.ceph.com/
- 175646 OSD 3005 clusters 1.39 EiB (hiers)
- https://ceph.com/
- Papier académique de Sage Weil « CRUSH »
  https://ceph.com/assets/pdfs/weil-crush-sc06.pdf
- Société indépendante puis achat par Red Hat maintenant IBM
- Packagé pour plusieurs OS incluant Proxmox VE

Concepts ceph

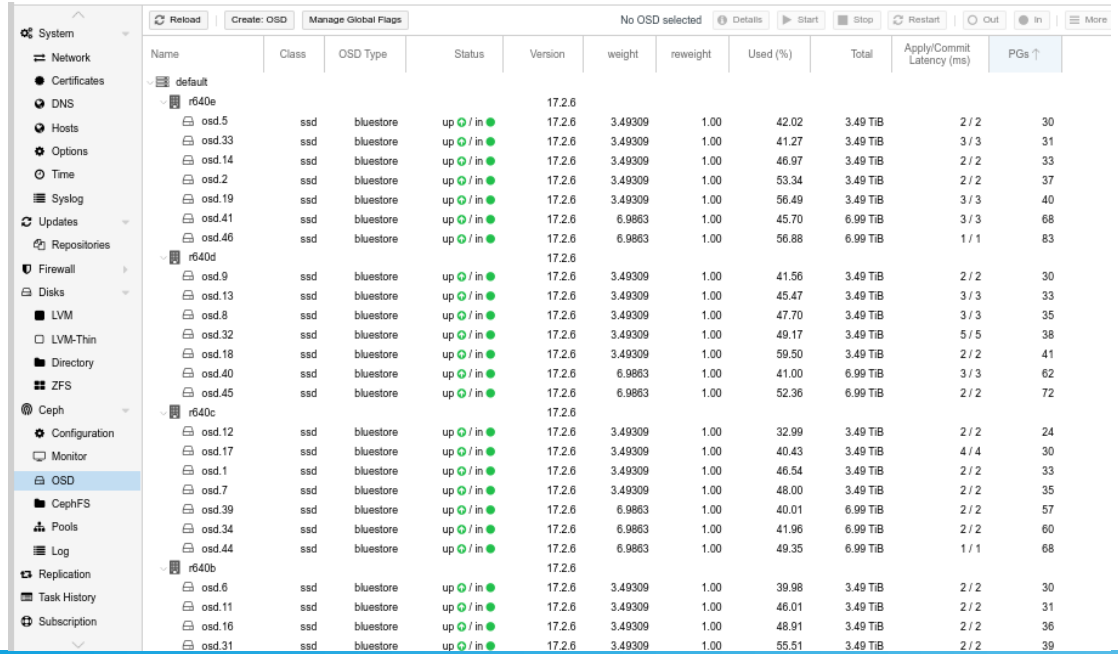
- 3 couches : Rados « object store » clé – valeur, la base
- RBD « rados block device » images disque de VM, construit par dessus rados (utilisé à Albi)
- Cephfs filesystem distribué (présentation suivante)
- Radosgw pour obtenir une API compatible S3
- Pool ceph : description d'une contrainte de réplication à respecter
- En cas de panne disque ou machine ceph constate que des contraintes ne sont plus respectées et régénère des copies

Proxmox VE et ceph

- Ceph 12 ajouté dans Proxmox VE version 5.0 en 2017
- Ceph 17 PVE 7.4 mars 2023 (version en production à Albi)
- Ceph 18 PVE 8.1 novembre 2023
- Version ceph N et N+1 supportées par Proxmox pour découpler les migrations PVE et ceph
- Documentation complete des migrations ceph
- Exemple : https://pve.proxmox.com/wiki/Ceph_Pacific_to_Quincy

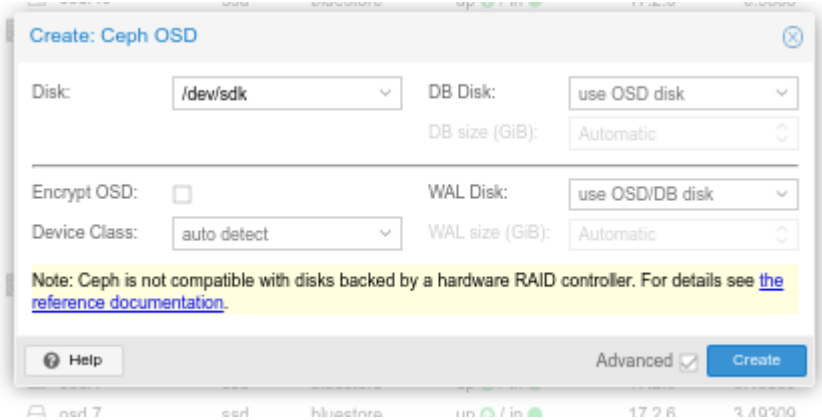
Installation

- Contrainte : cluster PVE avec minimum 3 machines, réseau 10G+
- Installation de ceph via l'interface web de proxmox



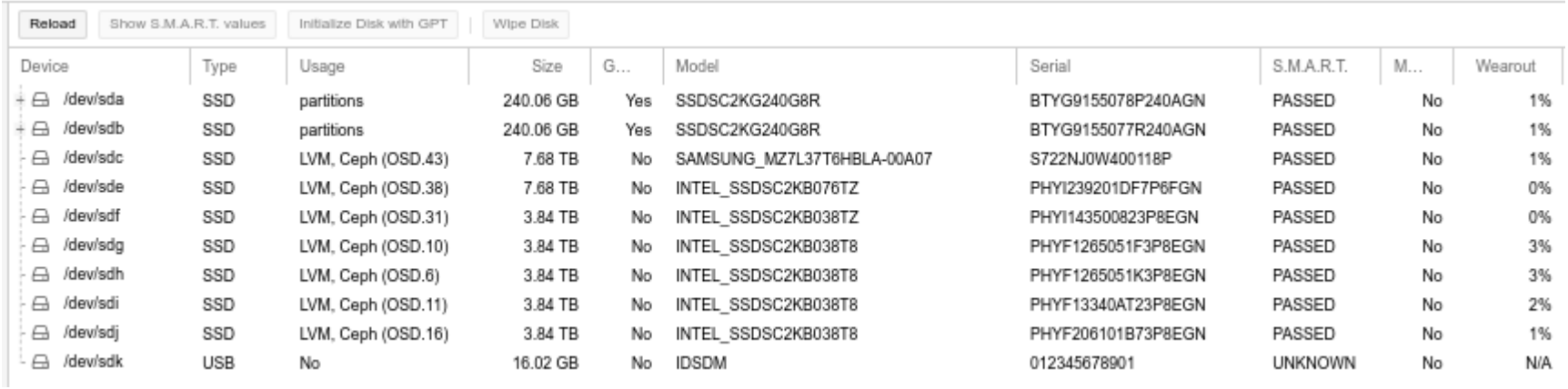
Ceph OSD

- Ceph : carte RAID a proscrire, il faut HBA / JBOD
- Au moins 3 process « monitor » répartis sur 3 machines, 2 « managers » (statistiques)
- Un process « OSD » par disque physique (* ou plus)
- Création via le web



Usage SSD à 3 ans

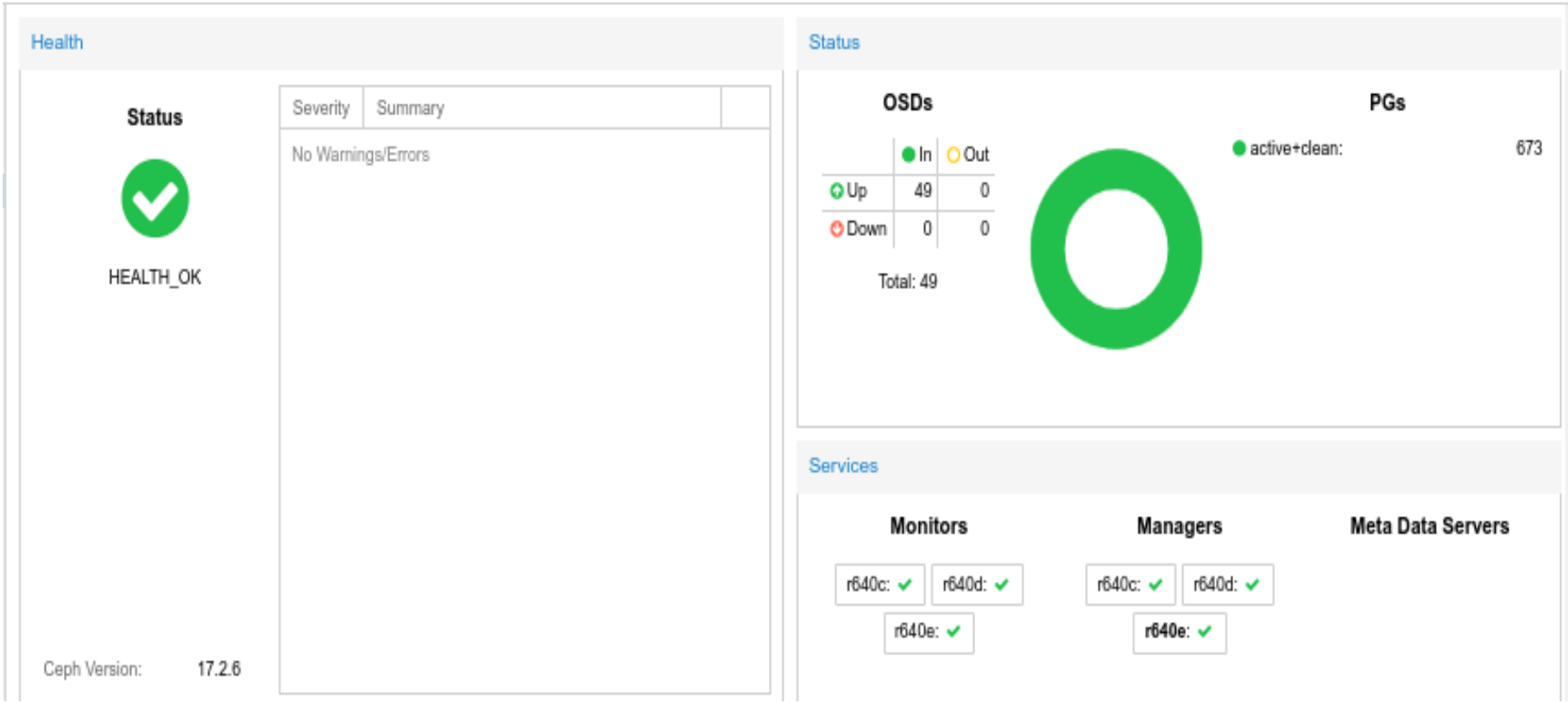
- PVE affiche le « wearout » node / disks
- A Albi les SSD seront obsolètes avant d’être usés :)
- 3 ans =>



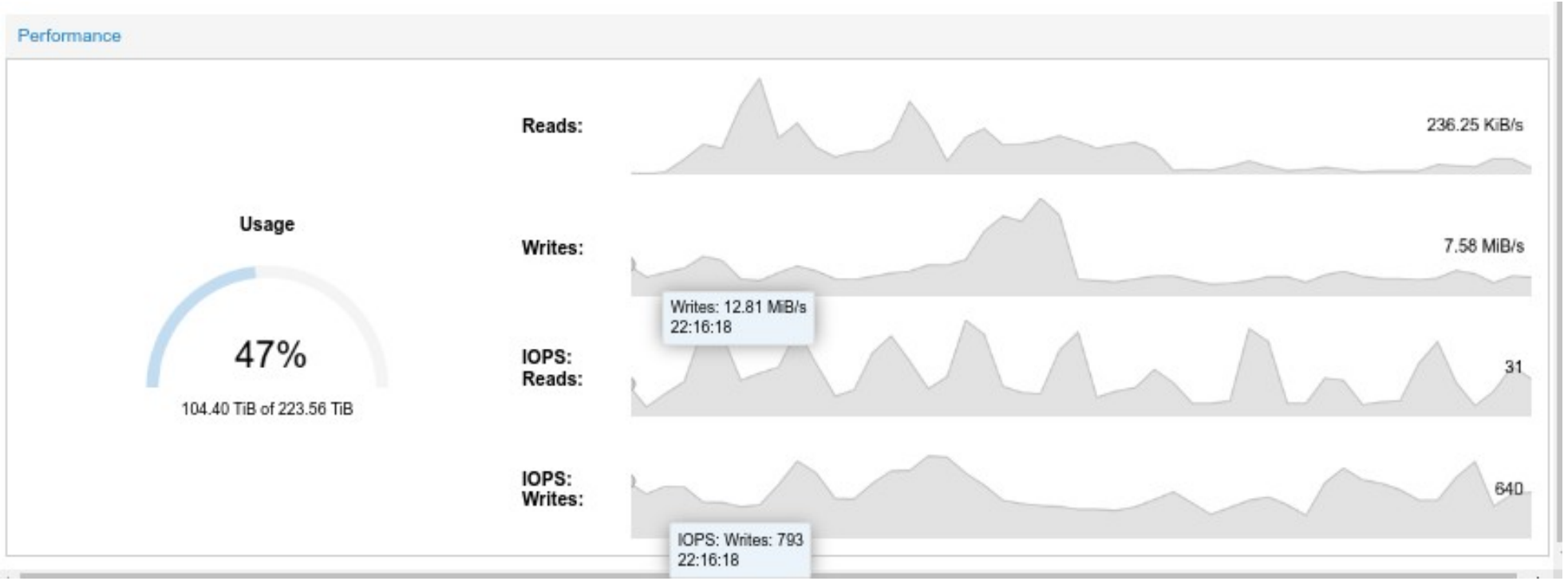
Reload | Show S.M.A.R.T. values | Initialize Disk with GPT | Wipe Disk

Device	Type	Usage	Size	G…	Model	Serial	S.M.A.R.T.	M…	Wearout
/dev/sda	SSD	partitions	240.06 GB	Yes	SSDSC2KG240G8R	BTYG9155078P240AGN	PASSED	No	1%
/dev/sdb	SSD	partitions	240.06 GB	Yes	SSDSC2KG240G8R	BTYG9155077R240AGN	PASSED	No	1%
/dev/sdc	SSD	LVM, Ceph (OSD.43)	7.68 TB	No	SAMSUNG_MZ7L37T6HBLA-00A07	S722NJ0W400118P	PASSED	No	1%
/dev/sde	SSD	LVM, Ceph (OSD.38)	7.68 TB	No	INTEL_SSDSC2KB076TZ	PHYI239201DF7P6FGN	PASSED	No	0%
/dev/sdf	SSD	LVM, Ceph (OSD.31)	3.84 TB	No	INTEL_SSDSC2KB038TZ	PHYI143500823P8EGN	PASSED	No	0%
/dev/sdg	SSD	LVM, Ceph (OSD.10)	3.84 TB	No	INTEL_SSDSC2KB038T8	PHYF1265051F3P8EGN	PASSED	No	3%
/dev/sdh	SSD	LVM, Ceph (OSD.6)	3.84 TB	No	INTEL_SSDSC2KB038T8	PHYF1265051K3P8EGN	PASSED	No	3%
/dev/sdi	SSD	LVM, Ceph (OSD.11)	3.84 TB	No	INTEL_SSDSC2KB038T8	PHYF13340AT23P8EGN	PASSED	No	2%
/dev/sdj	SSD	LVM, Ceph (OSD.16)	3.84 TB	No	INTEL_SSDSC2KB038T8	PHYF206101B73P8EGN	PASSED	No	1%
/dev/sdk	USB	No	16.02 GB	No	IDSDM	012345678901	UNKNOWN	No	N/A

Vue cluster ceph 1



Vue cluster ceph 2



Ligne de commande

- Quelques opérations ceph ne sont pas disponibles dans l'interface web
- Exemple : création de ruleset par type de stockage hdd et ssd
- Section 8.9. Ceph CRUSH & device classes
- ceph osd crush rule create-replicated replicated_hdd default host hdd
- ceph osd crush rule create-replicated replicated_ssd default host ssd
- Outil ceph direct, reconnu ensuite par proxmox

Ligne de commande 2

- Création de « pool » avec « erasure coding » qui est la généralisation de la notion de RAID « K+M » avec K disques de donnée et M disques de parité
- pveceph pool create MonPool22 --erasure-coding k=2,m=2
- Utilisation d’un outil PVE « pveceph » qui est un peu plus simple que l’outil ceph correspondant
- Section 8.8.2. Erasure Coded Pools
- Albi : utilisation replica 3 (défaut), EC 2+2 et bientôt EC 4+2

Cluster ceph externe

- Section 7.15. Ceph RADOS Block Devices (RBD)
- Configuration Example for a external Ceph cluster (/etc/pve/storage.cfg)

  rbd: ceph-external

  monhost 10.1.1.20 10.1.1.21 10.1.1.22

  pool ceph-external

  content images

  username admin

Synchronisation de deux clusters ceph

- https://docs.ceph.com/en/latest/rbd/rbd-mirroring/
- Journal-based: This mode uses the RBD journaling image feature to ensure point-in-time, crash-consistent replication between clusters. Every write to the RBD image is first recorded to the associated journal before modifying the actual image.
- Snapshot-based: This mode uses periodically scheduled or manually created RBD image mirror-snapshots to replicate crash-consistent RBD images between clusters. The remote cluster will determine any data or metadata updates between two mirror-snapshots and copy the deltas to its local copy of the image

Conseils

- Avec du replica 3 et de l’EC K+2 ceph permet perte de deux machines et de tous leurs disques sans perte de donnée
- Et de continuer la production… à condition d’avoir assez d’espace sur les disques des N-2 machines restantes
- Sur un cluster a 7 machines, 2/7 = 28.6 % donc ne pas dépasser 70 % d’usage
- Comme les répartitions ne sont pas parfaite plutot 60 % en pratique

Conseils 2

- Attention aux disques de tailles significativement différentes
- Pas évident a paramétrer, possibilité de diviser en deux les « gros » disques avec 2 OSD par disque
- Division aussi pour des raisons de performance
- https://ceph.io/en/news/blog/2023/reef-osds-per-nvme/
- https://ceph.io/en/news/blog/2024/ceph-a-journey-to-1tibps/
- 68 x Dell PowerEdge R6615 avec chacun 192 GB DDR5 2x100Gb 10x 15.36TB NVME

ZFS sur ceph pour NFS et CIFS

- Pour remplacer le NFS et CIFS du netapp « home » et partages
- VM debian 12 avec disque systeme classique ext4 et un ou plusieurs disques additionnels pour les données, sur ceph
- Chacun des disques donnée est un pool indépendant ZFS
- Actuellement 5 VM NFS et 1 VM SAMBA mode « standalone » qui monte les NFS et reexporte en CIFS
- Permet de répartir la charge, incluant les durées de sauvegarde PBS : 11 disques pour un total de 21 TB
- Redondance confiée a ceph et snapshot & cie à ZFS

Conclusion

- Pas de soucis en production ni lors des mises à jour
- En juillet 2023 après coupure électrique planifiée un des serveurs du cluster n’a pas voulu booter, proxmox+ceph => aucun problème avec une machine de moins
- Réintégration de la machine KO après réparation => idem
- Croissance progressive en machine et disque au fur et à mesure des besoins & budgets (PGD des projets de recherche => budget pour les disques)
- Pas de cycle de renouvellement imposé par un vendeur