PUGNÈRE Denis
CNRS / IN2P3 / IP2I
GT-ProxmoxVE de RESINFO

Stockage PROXMOX VE

Capitoul - 10/04/2025- IRIT - Université de Toulouse

iP2i LES 2 INFINIS LYON
cnrs NUCLÉAIRE & PARTICULES
UCBL Lyon 1
UNIVERSITÉ DE LYON

Plan

- Introduction / contexte
- Stockage Proxmox VE (PVE)
- Ajout d’un stockage dans PVE
- Nos choix de configuration
- Évolution après quelques années d’exploitation
- Conclusion

Introduction / Contexte

- IP2I : Institut de Physique des 2 Infinis de Lyon, UMR CNRS/IN2P3 + Université Lyon 1 (UCBL)
  - Champs d'activité
    - physique subatomique (physique des particules, les neutrinos, la structure nucléaire)
    - astro-particules et cosmologie avec la recherche de la matière noire, de l'énergie noire, ondes gravitationnelles
    - recherches interdisciplinaires liées à l'environnement, à la santé et à l'énergie nucléaire
    - LMA (Laboratoire des Matériaux Avancés) : dépôt de couches optiques ultraminces sur les mirroirs des détecteurs VIRGO / LIGO
  - 250 personnes
- Besoins (2015) :
  - rationaliser le déploiement des serveurs
  - Apporter de la souplesse, de l'agilité dans la gestion des ressources informatiques
  - Être moins impacté par les pannes matérielles (disques, serveurs)
  - Migration des VM pour maintenance...
- Étude de marché : VMware, Xen, PROXMOX VE... => PROXMOX VE (v4)

Stockage des VMs dans PVE (1/2)

- Virtualisation => Où et comment stocker les VM/CT ?
  - **Non hyperconvergé** : besoin d'un système de stockage des VM/CT en dehors de la solution de virtualisation => **2 silos** : 1 Virtualisation + 1 stockage
  - **Hyperconvergé** (unifié) : le stockage des VM est **intégré** à la solution de virtualisation, on gère le stockage des VM à l'intérieur de la solution de virtualisation, les hyperviseurs hébergent aussi les données des VM.
- PVE sait gérer les VM / CT / .iso / … de différentes manières :
  - En stockage local **intégré** (répertoire, LVM, LVM-Thin, ZFS, BTRFS)
  - Stockage réseau distant (NFS, iSCSI, CIFS)
  - Stockage distribué **intégré** ou distant (CEPH, CEPH-fs)
  - Stockage distribué distant (GlusterFS)
- PVE vous laisse le choix du mode d'hébergement des VM (hyperconvergé ou non)
- PVE utilise des plugins de stockage :
  - PVE communique avec les plugins avec les opérations de haut niveau (création, suppression d'images disque, snapshot, …)
  - Les plugins gèrent l'implémentation bas-niveau de chaque type de stockage

Stockage des VMs dans PVE (2/2)

- Les différents stockages sont définis à l'échelle du cluster PVE, on peut en créer autant qu'on veut
- 2 **catégories** de stockage :
  - Stockage de **type « fichiers »** : Les images disques sont stockés sous forme de fichiers (formats qcow2, raw), peut aussi stocker des images ISO, templates de containers, backups. Examples: Local directory, NFS, CIFS, …
  - Stockage de **type « bloc »** : Les images disques sont stockés sous forme de blocs de données. Exemples: ZFS (zvol), Ceph RBD, thin LVM, … Fonctionnalités comme les snapshots sont fournis par ce type de stockage
- 2 **caractéristiques** :
  - **Local** : stockage **implémenté sur chaque nœud** du cluster, **mais disponible seulement(*) depuis le nœud local** du cluster. Le contenu de ce stockage local peut être différent d'un nœud à l'autre. Exemples ; ZFS, répertoire, LVM, BTRFS
  - **Partagé** : Un stockage partagé est considéré comme étant **disponible avec le même contenu sur tous les nœuds**. Certains types de stockage sont partagés par défaut, par exemple tout partage réseau (NFS, iSCSI) ou Ceph RBD. Avantage => migration des machines en cours d'exécution sans interruption de service, sans recopie de l'image de la VM

    (*) PVE ≥ 8.4 : « Sharing host directories with VM guests using virtiofs : Virtiofs allows sharing directories between the Proxmox VE host and VMs without the overhead of a network filesystem »

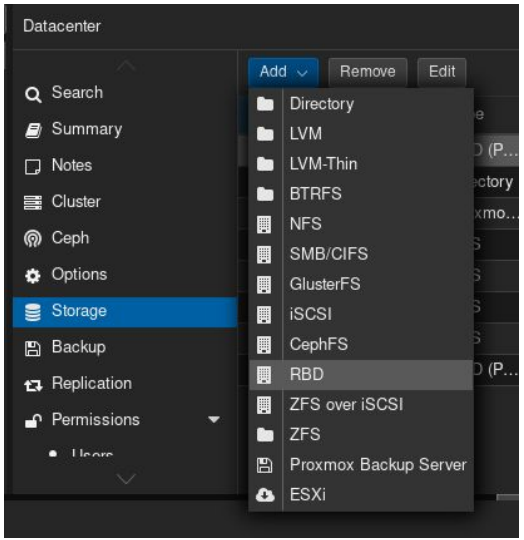
Description	Plugin type	Level	Shared	Snapshots	Stable
ZFS (local)	zfspool	file + block (zvol)	no	yes	yes
Directory	dir	file	no	no (2)	yes
BTRFS	btrfs	file	no	yes	technology preview
NFS	nfs	file	yes	no (2)	yes
CIFS	cifs	file	yes	no (2)	yes
Proxmox Backup	pbs	file + block	yes	n/a	yes
GlusterFS	glusterfs	file	yes	no (2)	yes
CephFS	cephfs	file	yes	yes	yes
LVM	lvm	block	no (3)	no	yes
LVM-thin	lvmthin	block	no	yes	yes
iSCSI/kernel	iscsi	block	yes	no	yes
iSCSI/libiscsi	iscsidirect	block	yes	no	yes
Ceph/RBD	**rbd**	**block**	**yes**	**yes**	**yes**
ZFS over iSCSI	zfs	block	yes	yes	yes

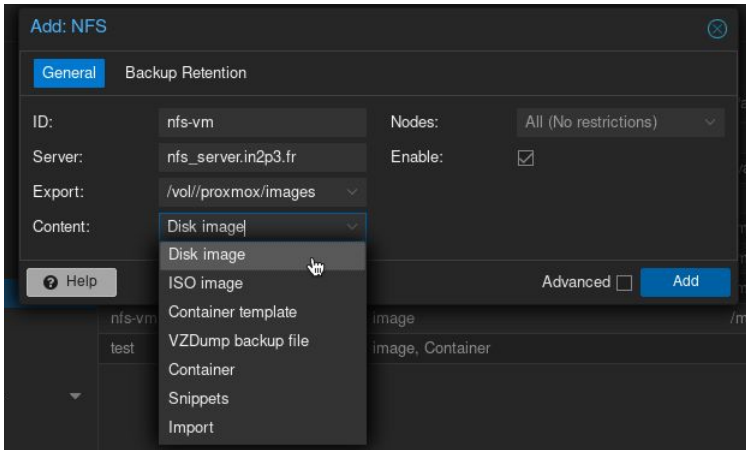
Ref PVE 8.3.1 :
* https://pve.proxmox.com/pve-docs/pve-admin-guide.html#_storage_types
* https://pve.proxmox.com/wiki/Storage

(2) : avec le format **qcow2**
(3) LVM au-dessus d'un système de stockage iSCSI ou FC => stockage LVM partagé

Ajout d'une cible de stockage

- GUI (sur un des noeuds) :
  Datacenter => Storage => Add => NFS





- **Content** : type de données à stocker (virtual disk images, cdrom iso images...) :
  - images : QEMU/KVM VM images.
  - rootdir : données persistantes des containers .
  - vztmpl : Container templates.
  - backup : Backup des VM/CT (vzdump).
  - iso : CDROM ISO images
  - snippets : scripts, guest hook scripts

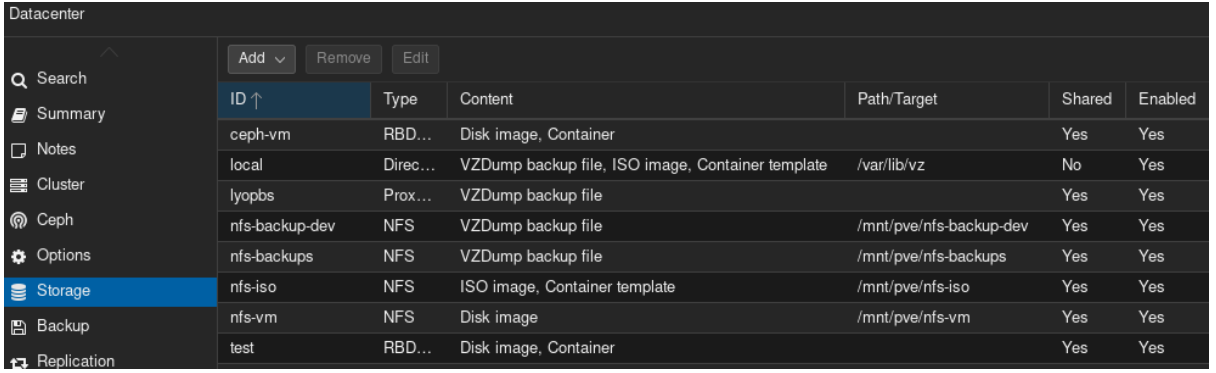
```
# cat /etc/pve/storage.cfg
nfs: nfs-vm
      export /vol/proxmox/images
      path /mnt/pve/nfs-vm
      server nfs_server.in2p3.fr
      content images
```

- CLI (sur un des noeuds) :

```
# pvesm add nfs nfs-vm --path /mnt/pve/nfs-vm --server nfs_server.in2p3.fr --export /vol/proxmox/images
# pvesm set nfs-vm --content images
```

Plusieurs cibles de stockage (y compris backups)

```
# pvesm status
Name                  Type     Status           Total            Used       Available        %
ceph-vm                rbd     active      69922868233      12447800329      57475067904   17.80%
local                  dir     active         57225328         20692084         33593956   36.16%
pbs                    pbs     active                0                0                0    0.00%
nfs-iso                nfs     active       8789177408         57907712       8731269696    0.66%
nfs-vm                 nfs     active       8789177408         57907712       8731269696    0.66%
test                   rbd     active      57475067910                6      57475067904    0.00%
```




```
# cat /etc/pve/storage.cfg

nfs: nfs-iso
        export /vol/proxmox/iso
        path /mnt/pve/nfs-iso
        server nfs_server.in2p3.fr
        content vztmpl,iso

nfs: nfs-vm
        export /vol/proxmox/images
        path /mnt/pve/nfs-vm
        server nfs_server.in2p3.fr
        content images

rbd: ceph-vm
        content images,rootdir
        krbd 0
        pool ceph-vm
```

Diverses images ISO

Virtualisation de serveurs physiques (dd if=/dev/sda of=/mnt/pve/nfs-vm)

Stockage des images des VM / CT

Notre choix : PVE + CEPH

Quelques caractéristiques de Ceph sur Proxmox VE qui ont guidé nos choix :

- Facilité d'installation et de gestion via CLI et GUI
- Stockage de blocs, de systèmes de fichiers et d'objets
- Prise en charge des snapshots
- Tolérant à la panne disque (les données sont répliquées) ou hyperviseur (HA, migration à chaud des VM), migration des données d'un stockage à un autre
- Pools avec différentes caractéristiques de performance et de redondance (replicas, Erasure Coding, HDD, SSD / NVME, tiering...)
- Fonctionne sur du matériel de base (debian)
- Pas besoin de contrôleurs RAID matériels
- Open Source

How to : https://pve.proxmox.com/wiki/Deploy_Hyper-Converged_Ceph_Cluster

Nos choix techniques initiaux et configuration

- Choix techniques initiaux
  - 8 hyperviseurs en 1 cluster de 4 hyperviseurs / DC, sur 2 DC (distance 200m)
    - 1 HV = 384 Go RAM, 2 * Intel Gold 5220R, 1 SSD 960 GB + 6 * 8TB,
      2 eth 25Gb/s (eth0 = trafic VM, eth1 = trafic CEPH)
    - Interco DC 1 <-> 10 Gb/s <-> DC 2
    - DC 1 (local) : 4 hyperviseurs + Corosync External Vote
    - DC 2 (distant) : 4 hyperviseurs
    - Stockage des VM dans CEPH hyperconvergé (RBD : RADOS Block Device) :
      - 6 OSD / HV, meta-datas sur SSD = 48 OSD
      - min 4, max 6 replicas, Crush rule = 2 replicas placés dans chaque DC, puis sur 2 hosts
  - Backup des VM sur NFS
  - PVE + CEPH hyperconvergé = 2 sous-ensembles à considérer indépendamment
    - PVE : calcul du quorum PVE (votes en cas d'arrêt d'HV, cf : **`pvecm status`**)
    - CEPH : Placement des données par CEPH en cas d'arrêt d'OSD ou d'un HV, cf : **`ceph osd tree`**)
- Exploitation / Tests / bascule / fencing

Évolutions après quelques années d’exploitation

- Changement géométrie cluster PVE
  - 1 cluster de 8 hyperviseurs => 2 clusters : 5 hyperviseurs / DC principal + 3 hyperviseurs / DC secondaire (reprise)
  - Stockage des VM dans CEPH hyperconvergé (toujours) :
    - min 2, max 3 replicas, Crush rule standard
  - En cours synchronisation stockage des 2 clusters (Ceph RBD Mirroring)
- Backup NFS => migration backup sur Proxmox Backup Server (PBS)
  - Espace NFS occupé par les sauvegardes trop grand => 10TB en croissance forte
  - Diminuer profondeur des sauvegardes, sélection des VM …. => recherche autre solution
  - Bascule des sauvegardes sur PBS
    - Plus efficace (déduplication coté serveur),
    - protection anti ransomware (user PVE avec droits « Datastore.Backup » seulement)
    - Rétention paramétrée coté PBS

Conclusion

- PVE :
  - Accessible à toute l’équipe ASR
  - + quelques « power users » pour gérer leurs VMs
- CEPH :
  - Solide : crash tests multiples réalisés sur cluster de test (sortie disques, arrêt brutal…)
  - A été résistant à la disparition (temporaire) d’un des 2 DC
  - Changement de la géométrie sans arrêter la production
  - Bien documenter la procédure changement disque HS / pre-fail
- PVE sait utiliser beaucoup de types de stockages :
  - On utilise NFS pour import de VM .qcow2 ou .raw ou les appliances
- PVE Devenu une brique d’infrastructure critique du laboratoire
- Très satisfait, on ne regrette pas
- Achat support « Community » pour supporter le projet

Merci pour votre attention !
Questions ?

Biblio :
- Wiki : https://pve.proxmox.com/wiki/Main_Page
- Docs : https://pve.proxmox.com/pve-docs/
- Roadmap : https://pve.proxmox.com/wiki/Roadmap
- Clusters répartis (sur 2 ou 3 DC) :
  - https://ceph.io/en/news/blog/2025/stretch-cluuuuuuuuusters-part1/ (02/04/2025)
  - https://ceph.io/en/news/blog/2025/stretch-cluuuuuuuuusters-part2/ (03/04/2025)