PRA facile avec Promox

Proxmox Virtual Environment & Backup Server
Sauvegardes sur bandes

Laurent GUERBY

JoSy Proxmox Virtual Environment - 08 octobre 2024

Licence

Ce document est sous licence

Creative Commons Paternité - Partage dans les mêmes conditions 4.0 International

Le texte complet de cette licence est disponible à l'adresse :
https://creativecommons.org/licenses/by-sa/4.0/

Auteur et institutions

- Laurent GUERBY, Administrateur systèmes et réseaux
- 80%, RSSI a mi-temps
- Institut Mines Telecom http://www.imt.fr/
- IMT Mines Albi https://www.imt-mines-albi.fr/
- Membre GT Proxmox et GT ZFS
- Participe a https://capitoul.org (Toulouse)
  - dont plusieurs présentations Proxmox/ceph/PBS

Environnement

- Rapport Annuel IMT Mines Albi 2023
- 362 personnels dont 83 Enseignants chercheurs
- 985 élèves dont 86 doctorants, 184 publications
- Campus à Albi 22 ha, 46200m2 surface plancher, 3 laboratoires
- 3000 prises réseau, 120 bornes wifi, 2 petites salles machines
- DSIN en 4 poles ISR = Ingénieur (2.6), AGP = Gestion de Parc (2 - 3)
  - SIM = SI Métier (5), Reprographie Audiovisuel (3), DSI (1)
  - Un poste CDI Ingénieur ouvert aux candidatures jusqu'au 1er novembre
    - prise de fonction à partir de février 2025

Situation initiale 2018

- Cisco 4400 routeur THdOC, Renater non redondé
- Stormshield SN900 non redondé
- Routage inter VLAN sur une stack de HP 5700
- Stockage NetApp FAS2552 100 TB (74 SATA, 25 SAS, 1 SSD), NFS et CIFS
- Serveurs : 2 chassis Dell M1000 et un FX2
- Virtualisation VMware stockage images VM sur NFS
- Un serveur Dell R730xd avec la solution de sauvegarde Atempo Tina
  - cache disque et robot de bande LTO7 Dell TL4000 en SAS
- Quelques machines physiques pour labo, non sauvegardées
- Tout sur un seul site à Albi
- … Pas de PRA

Juillet 2023

- Lors de la maintenance électrique annuelle du batiment Albi
  - Premier samedi de juillet
  - Mobilisation des ingénieurs du pole ISR avant et après la coupure
  - Documentation éteindre et rallumer le système d'information (SI)
- Tout est éteint electriquement sur notre site pendant une demi journée
- Test de bascule partiel sur notre PRA à Evry (500km)
  - LDAP, CAS, Shib, radius (eduroam), OpenVPN, web (moodle), SIP xivo (interne)
  - (pas les services avec état pour ne pas avoir à gérer le retour)
- Qui marche \o/

Stratégie pour un PRA facile

- Supprimer toutes les « appliances » matérielles
  - Une appliance implique une identique sur le site de PRA
  - Maintenir les versions et configurations a jour sur les deux sites
  - Donc budget EUR et temps humain
  - Et … risque d'erreur
  - Avec surprise lors de l'activation du PRA (murphy rulez)
- Migrer sur des serveurs génériques CPU, RAM, stockage
  - Perdre un serveur à un impact minimum
- Proxmox VE installé systématiquement sur tous les serveurs
- Hyperconvergé

Tout en VM Proxmox

- Passage au « tout VM » sur Proxmox VE
- Routage BGP via debian + BIRD
- Firewall et routage inter vlan via debian + nftables
  - Protection pour attaques PPS par Renater et THdOC
    - Lien gigabit/s pour tout le site Albi …
- Stockage via Proxmox CEPH
  - Replica 3, erasure coding 2+2 et 4+2
  - Suivant taille du cluster et redondance souhaitée
- VM ZFS + NFS + Samba standalone
- Proxmox Backup Server en VM unique sur un Proxmox VE

Mise en place du serveur PRA

- Pour le PRA achat d'une seule grosse machine Proxmox VE pour le site distant
- Merci a nos collègues IMT à Evry pour l'hébergement de la machine
- Envoi par transporteur pré-rempli et chiffré sur le site distant
  - LUKS du zvol et header en fichier séparé via `--header`
- Utilisation du service Renater de livraison de nos IPs sur un deuxième site
  - Avec distinction primaire secondaire via community BGP
  - https://services.renater.fr/services_ip/routage_d_adresses
  - Pas de renumérotation IP a faire lors de la bascule !

Infrastructure 2024 partie 1

- 2 salles a Albi séparées de 200m, 1 machine PRA a Evry
- CISCO nexus 3064PQ 48x10G + 4x40G, 2 paires (VPC) + 1 spare
- 2 fs.com S3900-24T4S-R pour le RJ45 (IPMI, GTC, divers)
- 4 PC Tour threadripper NVME 2x1TB RAID1 PVE independants
- 2 HPE DL385 mono EPYC 256G RAM 16 emplacements 2.5 RAIDZ3 12x8 TB
  - dont 1 avec 2 cartes HPE 2xSAS externe
  - 3xSAS vers robot Dell TL4000

Infrastructure 2024 partie 2

- Cluster proxmox + ceph
- 5 Dell R640 bi Xeon 512G RAM 12 emplacements 2.5 pouces SATA/SAS
- 3 HPE DL385 bi EPYC 1TB RAM 16 emplacements 2.5 pouces SATA/SAS
- SSD Samsung PM893 7.68 TB et Intel S4510/S4520 3.84 ou 7.68 TB
  - matinfo est un marché d'achat initial
  - liberté d'achat pour les pièces durant la vie des équipements
  - tiroirs
- Réseau serveur 4x10G LACP vers 2 nexus
- Réseau entre les deux salles LACP 8x10G entre les paires de Cisco Nexus

Infrastructure 2024 partie 3

- A Evry HPE DL385 plus musclé :
  - bi EPYC 7443 24C/48T
  - 2 TB RAM
  - 16x7.68 TB NVME over SAS
  - RAIDZ1 2 disques + RAIDZ3 13 disques + 1 spare
  - Un port réseau « out-of-band » sur le réseau IMT Evry
  - Un port réseau vers le NR Renater via la plaque réseau REVE
    - http://www.reve.fr/historique.php

Configuration sauvegardes

- Sur les Proxmox VE le backup configuré en inclusion par défaut
- Toute VM créée est automatiquement incluse dans la liste des VM sauvegardées sans action spéciale
- Seuls cas de non sauvegarde :
  - Ajout manuel d'une VM sur la liste d'exclusion
  - Decochage manuel de la case `backup` d'un disque dans la configuration de la VM dans Proxmox VE
- Objectif : minimiser le risque d'oubli et de mauvaise surprise en situation de PRA
- Sauvegarde mode bloc donc vs agent avec liste de répertoires pas de risque d'oubli non plus
- Deuxième facteur FIDO2/webauthn SSH et Proxmox

Proxmox Backup Server

- Developpé en rust et pas en perl
- https://pbs.proxmox.com/wiki/index.php/Roadmap
  - version 1.0 novembre 2020
  - version 3.2 avril 2024, actuelle
- https://www.proxmox.com/en/proxmox-backup-server/pricing
  - utilisable avec toutes les features gratuitement
  - Niveaux de support pour toutes les bourses
    - 520/1040/2080/4160 EUR HT par serveur par an

Proxmox Backup Server en savoir plus

- Même type d'installeur que Proxmox VE
- Même type d'interface web
- Présentations capitoul de Proxmox Backup Serveur
  - https://capitoul.org/ProgrammeReunion20211014
    - Installation avec copie d'écran
  - https://capitoul.org/ProgrammeReunion20220623
    - Utilisation en forensic
- Si vous avez un Proxmox VE vous pouvez facilement faire un test
  - En créant une VM Proxmox Backup Server

Proxmox Backup serveur sous le capot

- Sur le Proxmox Backup Serveur on configure des `datastore`
- Depuis les Proxmox VE on configure ces datastores comme datatcenter / storage
- Un `datastore` est un simple répertoire avec une arborescence prédéfinie et des fichiers
- Pour augmenter le niveau confiance dans la solution allons voir sous le capot
  - Logiciel libre et donc format ouvert
  - Format simple
  - Outillage

Proxmox Backup Server - datastore

```
root@backup:~# ls -la /mnt/datastore/datastore4
total 17520
drwxr-xr-x   6 backup backup     4096 Oct  6 10:33 .
drwxr-xr-x   6 root   root       4096 Mar 19  2024 ..
drwxr-x---   1 backup backup 17879040 Sep  2  2023 .chunks
drwxr-xr-x   9 backup backup     4096 Jan 11  2023 host
drwxr-xr-x 461 backup backup    12288 Oct  4 05:37 vm
```

Proxmox Backup Server - datastore - VM

```
root@backup4:~# ls -la .../datastore4/vm/433/2024-10-05T23:33:06Z
total 40
drwxr-xr-x  2 backup backup  4096 Oct  6 04:18 .
drwxr-xr-x 37 backup backup  4096 Oct  6 07:25 ..
-rw-r--r--  1 backup backup   555 Oct  6 04:18 client.log.blob
-rw-r--r--  1 backup backup 20480 Oct  6 04:18 drive-scsi0.img.fidx
-rw-r--r--  1 backup backup   412 Oct  6 04:18 index.json.blob
-rw-r--r--  1 backup backup   389 Oct  6 04:18 qemu-server.conf.blob
```

Proxmox Backup Server - datastore - debug

```
root@backup:~# proxmox-backup-debug
proxmox-backup-debug api create <api-path> [OPTIONS]
proxmox-backup-debug api delete <api-path> [OPTIONS]
proxmox-backup-debug api get <api-path> [OPTIONS]
proxmox-backup-debug api ls [<path>] [OPTIONS]
proxmox-backup-debug api set <api-path> [OPTIONS]
proxmox-backup-debug api usage <path> [OPTIONS]
proxmox-backup-debug diff archive <prev-snapshot> <snapshot> <archive-name> [OPTIONS]
proxmox-backup-debug help [{<command>}] [OPTIONS]
proxmox-backup-debug inspect chunk <chunk> [OPTIONS]
proxmox-backup-debug inspect file <file> [OPTIONS]
proxmox-backup-debug recover index <file> <chunks> [OPTIONS]
```

Proxmox Backup Server - datastore - debug inspect blob

```
root@backup4:~# proxmox-backup-debug inspect file \
 .../datastore4/vm/433/2024-10-05T23:33:06Z/qemu-server.conf.blob --decode -
agent: 1
boot: c
bootdisk: scsi0
cores: 2
...
```

- On retrouve le fichier `VMID.conf` de Proxmox VE

Proxmox Backup Server - datastore - debug inspect fidx

```
root@backup4:~# proxmox-backup-debug inspect file \
  .../datastore4/vm/433/2024-10-05T23:33:06Z/drive-scsi0.img.fidx | head
size: 2147483648
creation time: Sun Oct  6 01:33:08 2024
chunks:
  "b45158c650a5f5cc8715d549fa186c71d0126e74c2d686efb347ce6c9eb1149f"
  "07c655438daeadce6f56e8a7bde31cf2a499ccfe6b78381f84040a5e0e2f12a2"
  "e34e2267b8d0c0ff8c5811d8c130ba04845faebe93cdef02ba04717fc693ef8e"
  "41cce1daa40d353c64ebab894a902d2177528b3e9192a2e1a5e97c4014e6e8b6"
  "e4a8c8924fd6a2cebf04795919450ece0f9b8ec4e434469d5e8c3438a3e50c88"
  "d2466ba108651e9ba60efc7f2fe9ceba380aeee644c53096d86bebda1037e5a8"
  "53e8c96388116146a4278cb893e93a036c69e7aa7226d14a7adb608b28c4da9f"
...
```

Proxmox Backup Server - datastore - debug inspect chunk

```
root@backup4:~# proxmox-backup-debug inspect chunk .../datastore4/.chunks/b451/\
b45158c650a5f5cc8715d549fa186c71d0126e74c2d686efb347ce6c9eb1149f
CRC: "3919924461(OK)"
encryption: "none"
is-compressed: true
size: 1039837
```

- In fine un garbage collect
  - parser les .fidx et noter les chunks référencés
  - une fois fait supprimer les chunks non référencés

Proxmox Backup Server - datastore - statistiques

```
root@backup4:~# df -k /mnt/datastore/datastore4/
Filesystem       1K-blocks        Used  Available Use% Mounted on
/dev/sdb1      67885432772 66371197940 1514218448  98% /mnt/datastore/datastore4
root@backup4:~# du -ks vm
17927736  vm
root@backup4:~# du -ks .chunks
66401426252     .chunks
root@backup4:~# find .chunks -type f | wc -l
29294550
```

- 66 TB
- Seulement 18 GB de métadonnées hors chunks
- Moyenne de taille des chunks autour de 2 MB (algorithme « rolling hash »)

Fonctionnement horaire

- VM PBS backup1 sert pour le backup initial 00h-02h30
  - Les Proxmox VE ne voient que backup1
  - Pas le droit d'effacer sur PBS
  - Un compte PBS par Proxmox VE (un seul pour tout le cluster)
- VM PBS backup4 réplication locale à Albi
- backup-evry (PRA) et backup4 lancent remote sync avec backup1
  - début 3h30 fin 4h30
- Sur la machine de PRA à Evry le PBS est aussi en VM sur le Proxmox VE
- Le matin trois copies identiques, deux sites

Sauvegarde sur bandes

- Le matin backup4 lance 3 `tape jobs` incrémentaux en parallèle
  - 6h30-6h50
- Le vendredi matin actions supplémentaires :
  - Sortie d'un jeu (6 bandes) du robot, posé sur étagère
    - Quatrième copie « hors ligne »
  - Insertion dans le robot d'un nouveau jeu de 6 bandes, formatage
  - Lancement d'une `full` des derniers backups (« latest ») sur les 6 bandes en 3 jobs
  - Durée 13 heures environ, presque 1 GB/s cumulé sur les 3 lecteurs
- Du samedi matin au vendredi matin suivant incrémental

Sauvegarde sur bandes - Restauration

- Dans PBS les bandes ont une copie du datastore
- Pas de restauration de VM directement depuis les bandes
  - Il faut passer par un datastore sur disque
- Volumétrie disque insuffisante sur la machine de backup
- Solution :
  - Sur le cluster créer pool dédié au test de restaure
  - Le Proxmox VE qui porte le PBS devient client du stockage CEPH du cluster
  - Ajout a la VM d'un disque scratch sur ce nouveau stockage
  - Restaure !

PCIe Passthrough

- VM backup4 avec PCIe passthrough
  - Deux cartes HPE PCIe SAS externe vers robot LTO7 Dell TL4000
- Ajout a la VM backup4 de deux devices PCIe, via web UI
- Dans VMID.conf cela donne :

```
hostpci0: 0000:03:00,pcie=1
hostpci1: 0000:04:00,pcie=1
```

- Robot reconnu directement par Proxmox Backup Server

PRA - Réseau initial

- Dans les sauvegardes les VMID.conf sont enregistrés tel quel
  - incluant le `vmbr0`
- Parfait pour un restore en local dans le même Proxmox VE
- Exemple :

```
net0: virtio=52:54:ff:00:01:02,bridge=vmbr0,tag=10
net0: virtio=52:54:ff:00:03:04,bridge=vmbr0,trunks=2;3;4;5;6;7;8
```

PRA - Réseau cible

- Sur le Proxmox VE du site de PRA
  - Les interfaces physiques sont dans vmbr1 et vmbr2
- vmbr0 est libre et purement virtuel
- Donc quand on restaure rien à faire de spécial
- Machine virtuelle qui émule le routeur Renater/THdOC d'Albi
  - Même IPs d'interconnection coté interne vmbr0
  - BGP avec Renater via REVE coté externe vmbr1

PRA - divers

- En cas de PRA avoir ses docs facilement accessibles
  - pas sur une VM qui est KO …
  - Choix de mkdocs dans git https://www.mkdocs.org/
    - commande `mkdocs serve` et web sur localhost
    - avec fonction recherche qui marche en offline
    - ne pas oublier un `git pull` de temps en temps
  - Migration progressive des pages pertinentes depuis redmine
- Déploiement EDR Harfanglab incluant sur hyperviseur proxmox
  - et VM de stockage
  - achat SOC Exaprobe 24x7x365 via marché groupe logiciel

Futur

- Test automatisé des sauvegardes
  - Restauration dans un bac a sable
  - Démarrer chaque VM ping et quelques tests à scripter
  - Faisable avec les API et scripting proxmox mais pas présent par défaut
- Politique de prune par catégorie de VM
- Visibilité du cout stockage de backup par VM
- Parallelisation sur un hote des backups, optimisation niveau cluster

Questions