D’un crash à l’autre

Action Audace/ARAMIS : Proxmox VE et Ceph, mercredi 8 juin 2022

Sommaire

- Architecture avant le crash
- 1er crash
- Réparation
- Stabilisation relative et « scories »
- Étude nouveau cluster
- 2ème crash
- Réparation
- Conclusion

Architecture avant crash

- Utilisation
  - 1 pool RBD pour Proxmox
  - 1 pool en SSD pour les Metadonnées de CephFS
  - 1 pool en HDD pour les données pour les stockage des données scientifiques des labos LAPP et LAPTh, des homes des labos Université, des homes « grille »,...

- 4 serveurs en version Mimic
  - 1 MDS actif + 2 en « standby »
  - Pool SSD en « Replica 3 » pour les métadonnées
  - Pool HDD en « Erasure Coding 6+3 »

- Mise à jour du cluster en 2020
  - Passage en octopus (fin du support Mimic)
  - Ajout de 2 nouveaux serveurs
  - Maquette pour tester la procédure et bien la documenter (plusieurs semaines)

1er crash

- Pour la mise à jour effective de Mimic vers Octopus, application de la procédure, mais dès la mise à jour des « monitors », 6 OSD SSD sur 8 tombent :
  - ➔ Pool de métadonnées HS
  - ➔ Tout CephFS inaccessible

- La mise à jour implique un changement sur la structure des OSD et les 6 OSD « down » restent dans un état intermédiaire
  - ➔ Impossible de remonter les OSD « down » et de redémarrer les MDS

Réparation

- Tentative de reconfiguration des OSD « down » un par un et attente de reconstruction → Métadonnées toujours HS
- Tentative de reconstruction des métadonnées
  - Phase 1 « cephfs-data-scan scan_extents »
    - Analyse de tous les objets pour reconstruire les fichiers et calculer leur taille
    - Très, très...très long sur ~60 To → 2 jours complets
  - Phase 2 « cephfs-data-scan scan_inode »
    - Lecture du premier objet de chaque fichier pour récupérer les métadonnées (droits, heure de modif…)
    - Très, très long sur plusieurs millions de fichiers → + 1 jour
  - Phase 3 « cephfs-data-scan scan_link »
    - ***Lancée par script à la suite de la phase 2***
    - Récréation des liens et correction des erreurs
    - Rapide sans erreur !

    ➔ MDS ne veut toujours pas redémarrer

Réparation, suite

- Phase 4 « Réinitialisation »
  - des journaux
  - des tables d'inodes
  - des snapshots
  - du répertoire racine

    ➔ Ceux-ci conservaient leurs anciennes données

- Phase 5 « cephfs-data-scan scan_links »
  - Reconstruction des liens et récupération des erreurs
  - Une vingtaine d'erreurs rapportées (en particulier plusieurs liens pointant vers le même inode)

    ➔ Relance MDS OK

Stabilisation relative et « scories »

- Cluster remonté et semble stable sans perte de données

- Mais quelques scories :
  - Pas moyen de passer en multi-MDS : remplissage du pool de métadonnées puis plantage
  - Certains répertoires ne sont plus effaçables (vus comme répertoires non vide)
  - Plus moyen de faire évoluer le cluster

Étude d’un nouveau cluster

- Fin 2021 : Audit de la situation avec société extérieure et Sébastien Geiger
- Choix d’une nouvelle architecture avec les objectifs suivants :
  - Stabilité et performances accrues

- Passage en Pacific
- 9 serveurs
- Pool Métadonnées SSD en « Replica 3 »
- Pool Data SSD en « Replica 3 »
- Pool Data HDD en « Replica 3 » (avec WAL et RockDB sur SSD)
- Multi-MDS avec « Pinning » dynamique et statique (4 MDS actifs, 4 en « Standby-relay » et 1 en « Standby »)

2ème crash

- A partir d'avril 2021 : mise en place du nouveau cluster avec 3 nouvelles machines et transfert des plus petites zones sans soucis
- Bascule d'une première machine de l'ancien vers le nouveau cluster
- Il ne reste qu'une zone (données scientifiques LAPP et LAPTh) de 42 To à basculer mais il faut reconfigurer une autre machine de l'ancien vers le nouveau cluster
- Pour le basculement du serveur nécessaire (un peu en « urgence » car DELL nous demande un changement proactif d'une barrette de RAM) :
  - ✔ mise en « out » des OSD concernés
  - ✔ reconstruction OK
- « rm » de la machine :
  - ✗ le cluster repart en reconstruction
  - ✗ dans la nuit, 2 OSD d'une autre machine passent « down »

→ tout s'écroule

Réparation

- Zones « grille » et « université » → nouveau cluster, donc OK
- Zone du « labo » impactée
- Démontage de la zone sur les machines clientes afin d'éviter une aggravation
- Analyse de la situation avec l'entreprise extérieure
- Récupération des PGs impactés
- Réintégration de la machine sortie :
  - mais les OSD ne remontent pas, les « keyring » des OSD ayant été effacés par le « host rm »
- Réintégration des « keyring » et relance des OSD hébergeant des PGs impactés.
- Il reste un OSD qui était passé « down » automatiquement qui ne veut pas remonter (« failed to found label ») :
  - une zone de 1 ko en début de disque contenant le label de bluestore est absente
- Les outils de bluestore permettent de modifier des champs du label mais pas de recréer un label vierge
- Donc, copie avec « dd » du label d'un OSD OK, puis modification des champs nécessaires avec « /usr/bin/ceph-bluestore-tool set-label-key »

Réparation

- Reste 68 objets unfound sur (53,5M) :
  - ➔ Blocage de la reconstruction
- Si le pool avait été en « Replica » : possibilté de « revert » sur ces objets afin de récupérer une ancienne version
- Mais en EC : « delete » des objets → max 68 fichiers impactés.
- Sur les 53,5M de fichiers :
  - deux utilisateurs en totalisent + de 37M peu importants
    - ➔ suppression ce qui facilitera la synchro vers le nouveau cluster.
- Sur les 16M restants :
  - énormément d'environnements « conda » reconstructibles
    - ➔ peu de fichiers réellement perdus (<10)
- A l'issue de la réparation :
  - Sortie un à un des OSD (« out » puis « down »)
  - Surveillance des autres OSD (positionner les flags « noout » et « nodown » en cas de « flapping »)
- « rm » de la machine → OK
- Reconfiguration de la machine sur le nouveau cluster et synchro de la zone → OK
- Sauvegarde de la zone avant mise en production
- Ancien cluster conservé pour tests de restauration des sauvegardes

Conclusion

- Après un gros crash sur un cluster hébergeant du CephFS, il vaut mieux repartir sur un nouveau cluster propre assez rapidement
- Lors de la sortie d'un serveur, sortie manuelle des OSD (« out » puis « down ») en surveillant l'état des autres OSD
- Pour du CephFS :
  - Utiliser du « Replica » :
    - meilleures performances
    - plus de possibilités de récupération en cas de crash
  - avoir une redondance des compétences ou souscrire à une aide extérieure :
    - permet de mieux analyser les problèmes et les solutions à appliquer

Suppléments