NFS => CEPHFS

PIERRE GAMBAROTTO

Created: 2024-02-29 jeu. 09:43

QUI SUIS-JE ?

INSTITUT DE MATHÉMATIQUES
de TOULOUSE

Responsable informatique

- dans un laboratoire
- avec une petite équipe

- réseau métier des ASR des laboratoires de Math
- membre de la PLMTeam : gère des services au niveau national pour les mathématiciens

PLAN

- contexte et besoin
- cephfs ?
- plateforme de test
- la suite

POINT DE DÉPART

labo : ~400 utilisateurs actifs

baie netapp, partage en NFSv4/Kerberos

UTILISATION

Partage	Taille	%	Utilisation 1	Utilisation 2
homes	7To	70%	postes fixes NFSv4/ Kerberos	Accès serveur ssh
pages web perso	700Go	20%	frontal apache	Accès ssh/sftp pour l'édition
web divers	300Gb	40%	frontal apache	Accès ssh/sftp pour l'édition

=> faibles volumes de données

=> pas de besoin de performance

HOMEDIR : GESTION DES DONNÉES UTILISATEURS

passage au tout portable sur la fin :

- forte baisse de l'utilisation de la baie
- données perso gérées par seafile

SITES WEB

Migration des pages perso existante vers :

- pages manuelles: générateur de site statique en intégration continue par gitlab
- wordpress pour le reste

=> migration à faire

LES BESOINS

Baie netapp en fin de vie, évolution des besoins

- plateforme de calcul:
  - home et scratch partagés sur les nœuds du cluster
  - store guix/nix en montage commun
- accès ssh/sftp pour gestion à conserver

Labo de mathématiques : volume très faible des données

mais nette évolution côté calcul

CHOIX TECHNIQUE

- cluster proxmox de 5 nœuds disponible, dell R440
  - utilisé pour virtualisation des serveurs
  - ceph intégré, en mode block pour les vms
  - réseau d'accès vm 10Gb
  - réseau ceph 10Gb séparé
  - SATA ssd, 3*900Go
  - cephfs disponible, peu utilisé (templates vm/ct)

DÉBUTS 2023

Test sur l'espace de scratch des serveurs de calcul

=> go technique pour cephfs

Alternatives envisagées :

- nfs sur rbd
- pnfs/ganesha

ÉVOLUTION DE LA CAPACITÉ DE STOCKAGE

3 baies de disques de libre par serveur

fin 2023 : => 65To

- Ajout de 3 disques SATA SSD de 3.8To
- tiroirs Workdone

Prix dell : 2200€ par disque + tiroir …

MMI : intel D3-S4520 3.84To + tiroir ~ 380 €HT

Merci à la liste Capitoul et à L. Guerby pour le conseil :-)

CEPHFS

Les notions à retenir :

Ceph :

- osd : gestion d'un disque dur physique
- unité de stockage : objet RADOS
- pool ceph : partition logique, assure la réplication d'objets RADOS sur un ensemble d'OSD
- MONiteur : serveur ceph, utilisé par les clients : un des membres du cluster

RÉPLICATION

Choix classique : 3 exemplaires de chaque objet

=> diminution de la capacité réelle

=> augmentation de la qualité de vie

UN PARTAGE CEPHFS

- un pool ceph pour les données
- un pool ceph pour les métadonnées

Accès fichier similaire à NFS

Par rapport au mode block de ceph :

- mds : serveur de métadonnées, 1 actif, plusieurs secondaires
- point faible : un seul serveur actif en même temps, risque si beaucoup de fichiers

=> point à surveiller pendant les backups

CEPH BY PROXMOX

- base debian
- distribution spécifique de ceph : dépôt debian spécifique, un par version de ceph
- GUI web
- cli ceph classique
- très bonne expérience, simplifie largement la mise en place d'un cluster ceph.
- les montées de version se passent bien

PLATEFORME DE TEST

Scénario 1 : serveur pour accès ssh/sftp des homedir, ou nœud de calcul

NŒUD CEPH/PROXMOX

On réutilise le cluster ceph fourni par proxmox

partage cephfs : 2 pools ceph un pour les données, un pour les méta données

```
ceph fs ls
```

```
name: cephfs, metadata pool: cephfs_metadata, data pools: [cephfs_
```

Réutilisation du partage cephfs *cephfs* existant.

Même principe que NFSv4 : on va gérer des répertoires dans le partage

GESTION DES DONNÉES CÔTÉ SERVEUR

- Sur n'importe quel nœud proxomx
- le partage *cephfs* est monté sur `/mnt/pve/ceph-fs`
- gestion par sous-répertoire

Par exemple, pour la gestion des `home` :

- répertoire `/homes/*` du partage *cephfs*
- monté sur `/mnt/pve/ceph-fs/homes/`

CRÉATION D'UN UTILISATEUR CEPHFS

client cephfs :

- nom
- clef
- liste de droits.

```
# CEPHfs share : cephfs
# client name : homeserv
ceph fs authorize cephfs client.homeserv /homes rw
```

=> génère une clef identifiant le client

=> lecture/écriture sur le répertoire `/homes` du partage *cephfs*

RÉCUPÉRER LES INFORMATIONS

```
ceph auth get client.homeserv
```

```
[client.homeserv]
    key = ABCDXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX==
    caps mds = "allow rw fsname=cephfs path=/homes"
    caps mon = "allow r fsname=cephfs"
    caps osd = "allow rw tag cephfs data=cephfs"
```

À stocker sur le client dans `/etc/ceph/ceph.client.homeserv.keyring`

SUR LE CLIENT CEPH

- Installation de `ceph-client`
- `/etc/ceph/ceph.client.homeserv.keyring` : identifiant du client
- `/etc/ceph/ceph.conf` : configuration générale pour trouver les serveurs ceph et identifier le cluster

```
ceph config generate-minimal-conf
```

```
# minimal ceph.conf for a1234567-1234-5678-1234-bkc123456789
[global]
    fsid = a1234567-1234-5678-1234-bkc123456789
    mon_host = [v2:10.0.0.1:3300/0,v1:10.0.0.1:6789/0] [v2:10.0.0.
```

MONTAGE MANUEL

```
mount -t ceph :/homes /homesceph -o name=homeserv
# -o secretfile=/etc/ceph/ceph.client.homeserv.keyring
# 10.0.0.1,~,10.0.0.5:/homes : default mon servers from /etc/ceph/
```

Pour une utilisation basique : c’est tout !

MIGRATION NFS => CEPHFS

- monter les anciens homes et les nouveaux
- rsync

trèèèèèèèèès long

MONTAGE AUTOMATIQUE DES HOME

par le classique autofs

```
# auto.master
/home file:/path/to/auto.home -strict

# auto.home
* -fstype=ceph,name=homeserv 10.0.0.1,…,10.0.0.5:/homes/&
```

QUOTAS UTILISATEUR

On peut positionner un quota :

- par répertoire
- nombre max de fichiers
- taille max
- paquetage `attr` : `setfattr/getfattr`

=> gestion classique des attributs étendus, à la XFS

```
setfattr -n ceph.quota.max_bytes -v 50Gi /mnt/pve/ceph-fs/home/gam
setfattr -n ceph.quota.max_files -v 500000 /mnt/pve/ceph-fs/home/g
```

Attention : le respect du quota est géré au niveau client, pas au niveau du serveur.

=> attaque possible d'un client malicieux :-(

ENROBAGE

À l'arrivée d'un utilisateur :

- création du homedir
- positionnement du quota

Bug pour le moment : à faire sur un nœud du cluster ceph pour la partie `setattr`

RETOUR PREMIÈRES EXPÉRIENCES

Testé à ce jour :

- montage noyau manuel et autofs
- sssd pour la liaison au ldap, sans kerberos
- debian 12
- nixos 23.11

Dans tous les cas : utiliser une version de ceph-client supérieure à celle du serveur

SERVEUR DEBIAN 12

- utiliser le même dépôt apt que sur un nœud proxmox

```
# /etc/apt/sources.list.d/ceph.list
deb http://download.proxmox.com/debian/ceph-quincy bookworm no-sub
```

NIXOS

La cli ceph est en python : très sensible à la version

pin au commit qui marche pour `ceph-client`

COMPARAISONS NFSV4/ CEPHFS

On perd :

- la visualisation des quotas par la commande `quota -v` : pas très lisible de toutes façons
- la sécurité offerte par Kerberos : en multiuser, on retombe sur les droits unix

On gagne :

- tolérance aux pannes, grâce à la réplication 3
- capacité évolutive : rajout de disques/nœud

PAS TESTÉ/À FAIRE

- pas de soucis à signaler sur la plateforme de test
- IMT : pas de besoin de performance spécifique

SCRIPT POUR GESTION UTILISATEUR

Intégrer dans la gestion de l'arrivée d'un utilisateur :

- après la création des attributs du schéma POSIX dans l'annuaire LDAP
- création du homedir
- positionnement du quota
- optionnel : création d'un client ceph spécifique à l'utilisateur

TESTS DE PERFORMANCE

- mesurer les capacités brutes d'un disque ssd avec hdparm

```
DISK=/dev/sdX
hdparm -tv $DISK
# cached reads, offset à 20 GB du debut de disque :
hdparm --offset 20 -T $DISK
# bypass, verbeux, du buffer cache memoire du disque :
hdparm -tv --direct $DISK
```

- test sur le système de fichiers monté : iozone
- fio : tests non destructif
- io500 : test pour cluster de fichiers

MONTAGE FUSE

un montage par FUSE permettrait à un utilisateur l'accès à ses données sur son portable.

- déclarer un *client* ceph par utilisateur, restreint au home de l'utilisateur

  => `/etc/ceph/ceph.client.LOGIN.keyring` rm sur `/homes/LOGIN`

```
ceph-fuse -n client.LOGIN /mnt/my_ceph_homedir -r /homes/LOGIN
```

BACKUP

Archivage sur 1 an avec borgbackup

Pour cela :

- snapshot cephfs ? -> génère un `.snap` dans chaque répertoire
  - pas forcément une bonne idée sur le `/homes` entier : le serveur de métadonnées est facilement occupé, générer des tas de fichiers est dangeeurx
  - snapdiff : différentiel entre 2 snapshots, arrive sur ceph *squid*
- machine de backup :
  - client du serveur borg
  - client cephfs en lecture seule
- tester les backups …