ZFS on PowerEdge R730xd

Martin Souchal, 3 octobre 2017

Le besoin

Cahier des charges

- Héberger des fichiers partagés en NFS sur des serveurs linux
- Pas de performance en I/O requise
- Pas de budget ni d'utilité (redondance…) pour applicance dédiée (NAS Dell ou NetApp…)
- Disponible sur MATINFO
- 20 TB de stockage utile

Etat de l'art

Solutions envisagées en septembre 2016

- DELL Storage Center (SVC 2000 2 controleurs, 8g RAM, 10 x 4 To SAS nearline) : 26 110 €
- NetApp FAS2650HA (2 controleurs, 24x900 SAS) : 31 316 € (45 % de réduction)
- Dell R730xd (2 * Intel E5-2609 v4 8c, 64 Go de ram, 10 x 4 To SAS nearline 7,2k) : 6000 €

DELL PowerEdge R730xd

- Serveur unique 64b
- châssis 2,5" pour maximum 24 + 2 disques durs Hot plugs en 2,5"
- châssis 3,5" pour maximum 12 + 4 Interne Hot plug + 2 disques durs Hot plugs en 2,5"
- Compatible disques SSD, SAS, SAS 10k
- Acheté sans carte controleur RAID
- Cartes réseau 1G/10G/40G/IB... disponibles
- Double alimentation
- 10 disques 4 To données, 2 disques 600 Gb systeme
- 2 * Intel E5-2609 v4, 64 Go de ram
- Sans OS

Historique

- ZFS est le filesystem développé par Sun pour Solaris en 1993
- Code source de solaris ouvert en 2005 => naissance d'OpenSolaris
- ZFS on Linux est issu d'OpenZFS, lui même issu de OpenSolaris
- La premiere release stable de ZFS on linux date de 2013
- Officiellement intégré et supporté à partir de Ubuntu Xenial 16.04
- Ne fait pas partie du noyau Linux pour des problèmes de licenses

Caractéristiques

- "ZFS" signifie "Zettabyte File System". Aujourd'hui il peut supporter jusqu'a 256 10^24 ZiB (zebibytes), soit 3x10^34 Gb.
- A la fois système de fichier et gestionnaire de volume
- Déduplication
- Snapshots
- Compression
- Chiffrement $
- Supporte SMB et NFS
- Quotas utilisateurs et groupes
- Gestion du cache disque

ZFS On linux - RAID Logiciel

- Il est extrement déconseillé d'utiliser du RAID matériel
- raid0 : les données sont réparties sur les disques sans redondance. très performants mais pas sécurisé (une erreur disque, tout est perdu)
- raid1 : miroir. Sécurisé mais très limité (la taille disponible est égale à la taille du disque le plus petit, les performances sont celle du disque).
- raid5 or raidz : simple parité, il est possible de perdre un disque. De nombreuses études ont montrées que le RAID 5 est trop risqué, car la charge placée sur les disques restants au moment de la reconstruction d'un disque endommagé augmente de 8% les chances de perdre un deuxieme disques et donc toutes les données. Source : http://storagemojo.com/2007/02/26/netapp-weighs-in-on-disks/
- raid6 or raidz2 : double parité, il est possible de perdre 2 disques.

ZFS On linux - RAID Logiciel

- raid7 or raidz3 : triple parité, il est possible de perdre 3 disques.
- raid10 or raid1+0 : miroir + distribution des données. 4 disques et 2 miroirs : le 1 et 2 sont des miroirs et les 3 et 4 sont un deuxieme miroir. Les données sont ensuites réparties sur les 2 miroirs. Il est possible de perdre 2 disques sur les deux miroirs. Vous ne pouvez pas perdre les 2 disques d'un même miroir. Rapide en lecture, le RAID 10 est très lent en écriture et limité en place.
- raid60 or raid6+0 : 2+ volumes RAID 6. Avantage du RAID 6 : on peut perdre deux disques par volume et la vitesse de lecture est améliorée .
- raid70 or raid7+0 : 2+ volumes raid7. Mêmes avantages et inconvénients que le raid60.

=> Plus l'indice du raid est élévé, plus les performances sont dégradées. RAID0 > RAID 1 > RAID 2, etc...

Configuration

- 10 disques 4 Tb répartis en 9 disques raidz2 et un spare pour volume final de 21 To utiles
- 2 disques de 600 Go en mirroir pour l'OS
- OS : CentOS 7 avec module DKMS ZFSoL => contraignant pour les updates kernel
- scrub 8 jours
- Script journalier de vérification d'erreur disque
- autoreplace on
- compression on
- autoexpand off
- secteurs de 512b

Avantages

- Un NAS toutes options pour 6000€
- Facilité de gestion / administration
- Controle total de la machine et de la configuration
- double alimentation

Inconvénients

- Moins robuste : pas de redondance controleur, pas de redondance proc, mémoire, etc
- Il faut un admin

Questions

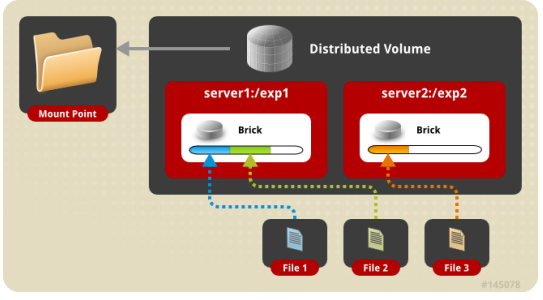
GlusterFS

GlusterFS

- Utilisé sur cluster de calcul
- Distribue sur chaque noeud de cluster /home, /workdir et espaces projets depuis un attachement SAN iSCSI 10Gb/s Equallogic
- Volumétrie 80T
- Permet de palier à une limitation equalogic (volume max de 15T)

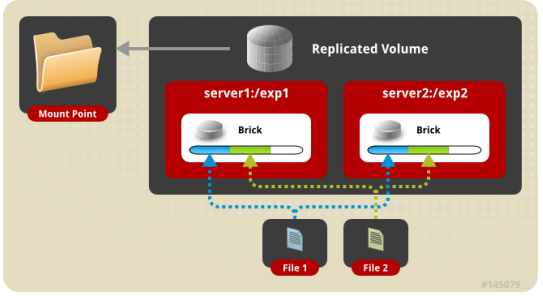
GlusterFS

Architecture



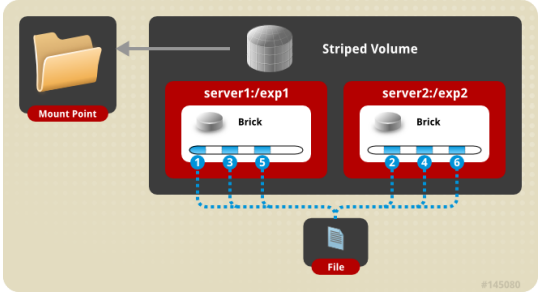
GlusterFS

Architecture



GlusterFS

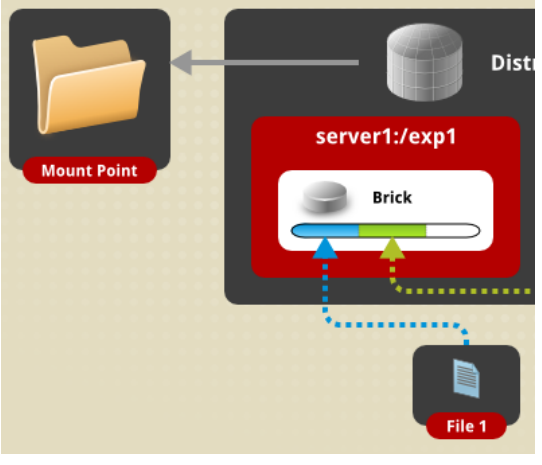
Architecture



GlusterFS

Architecture

- En pratique chez nous :



GlusterFS

Avantages

- Simple à installer et a mettre en place

Serveur

```
yum install glusterfs-server
service glusterd start
gluster volume create gv0 replica 2 serv
gluster volume start gv0
```

Client

```
yum install glusterfs glusterfs-fuse
mount -t glusterfs server1:/test-volume /
```

GlusterFS

Avantages

- Simple a mettre en oeuvre
- libre et gratuit
- Maintenu par Red Hat
- Diversité des approches
- Compatible RDMA

GlusterFS

Inconvénients

- Fuse
- Pas très performant
- Sensibles aux dossiers avec bcp de fichiers
- Stabilité / bugs

Questions

BeeGFS

BeeGFS

- Stockage distribué performant hardware-independant
- Gratuit*, open source*
- Developé par Fraunhofer / ThinkParQ
- 3 briques : metadata, storage, management
- rapprocher les données du processeur
- gestion de filesystems "on the fly"
- pas de patch du kernel

Architecture

- Client : monte un FS
- Storage Service : contient les données
- Metadata : Informations sur les fichiers (postion,taille,permissions), pas d'accés entre ouverture/fermeture
- Management Service : Registre et surveillance
- Un seul gros filesystem pour tout le cluster

Tous les services peuvent être situés sur une même machine ou sur X machines. Plus on augmente le nombre de noeuds, plus on augmente les perfs.

Bonnes pratiques

- Metadata en ext4 sur SSD / RAID 1 ou 10
- RAID 6 pour assurer l'intégrité des données
- Plusieurs serveurs metadata
- Plusieurs serveur de données pour améliorer le débit

Fonctionalités

- ACLs
- Quota
- Data et metadata mirroring
- Stats par utilisateur

Tests à l'APC

- Partage d'un filesystem iSCSI sur baie Equallogic en 10 Gb/s
- Metadata / Storage / Management sur le même serveur => Pas les conditions idéales

Bilan

- Plus performant que GlusterFS
- Aussi simple
- Documentation succinte, peu de communauté
- Production sans payer ?