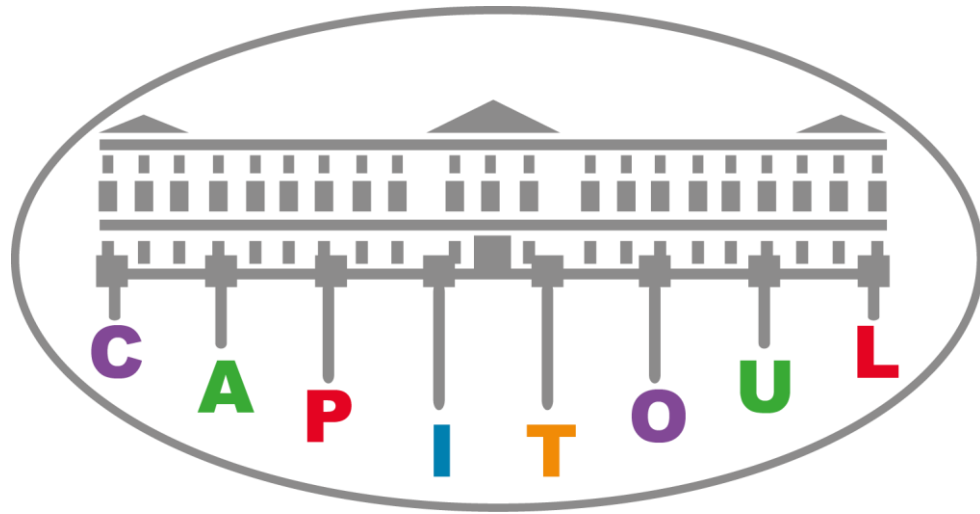
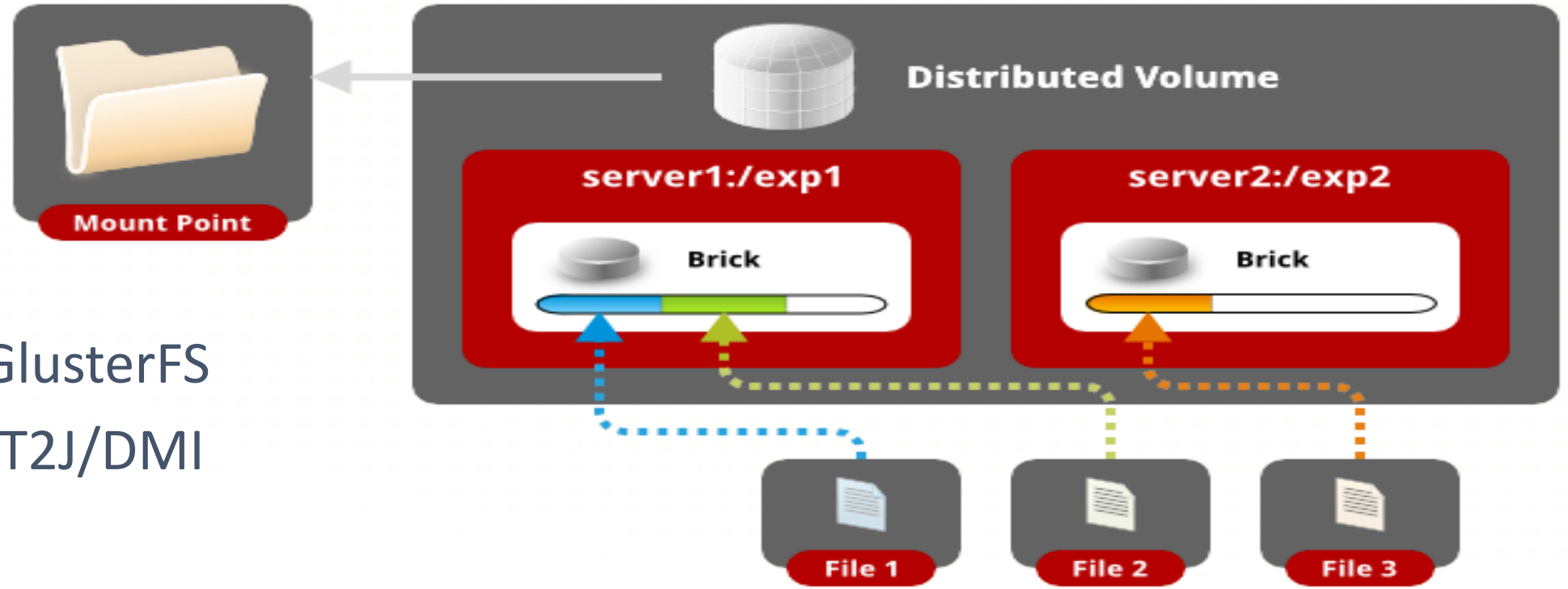


GlusterFS



Introduction

- Présentation GlusterFS
- Utilisation à UT2J/DMI
- Références
- Conclusions



GlusterFS

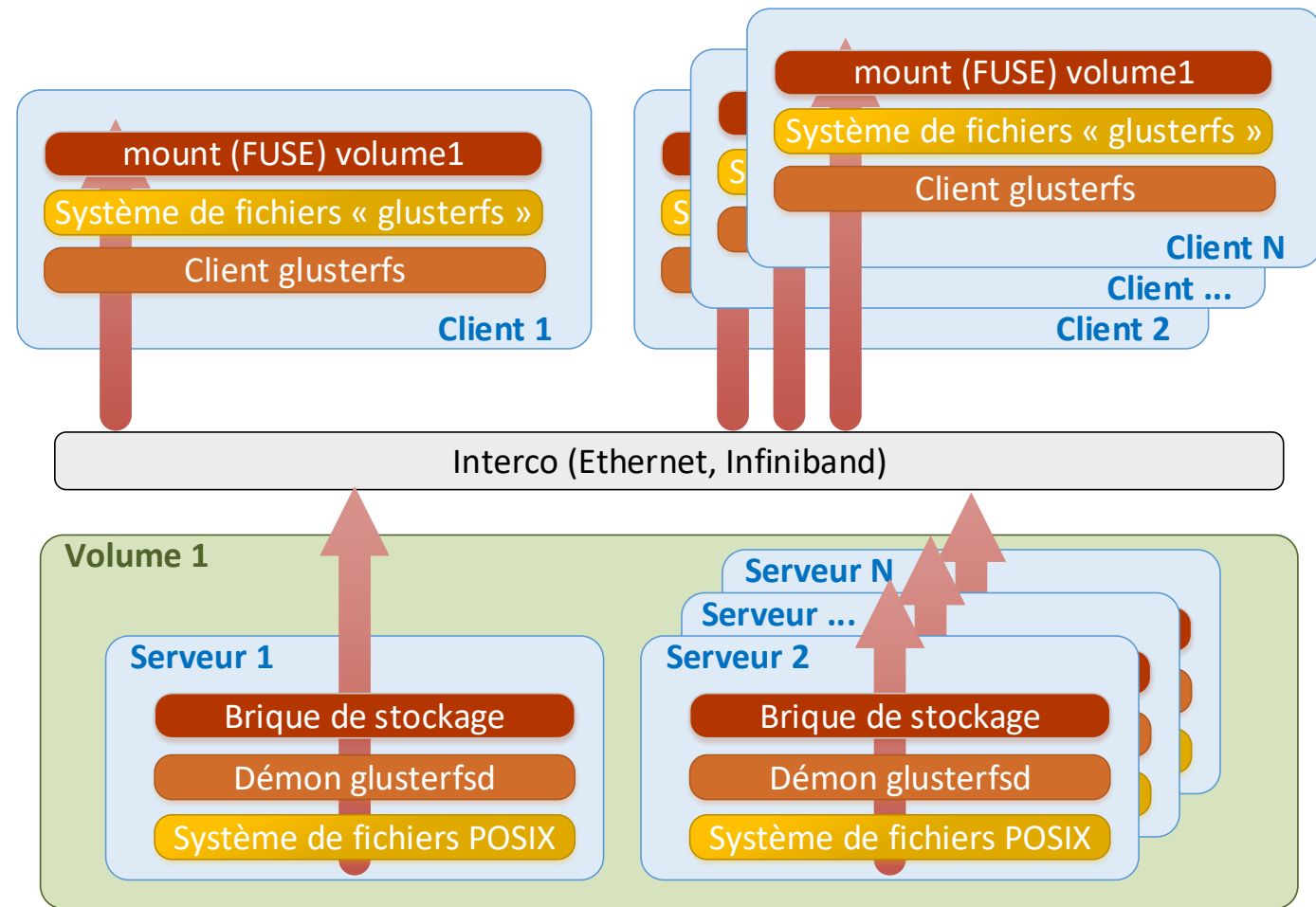
- Présentation de GlusterFS
- Fonctionnement
- Modes de fonctionnement des volumes

GlusterFS - Présentation

- Système de fichier distribué en parallèle (grappe de serveurs)
- Agrégation de plusieurs sources de stockage sous un seul espace de nommage
- Plusieurs pétaoctets (10^{15} octets)
- Capable de supporter des centaines de clients
- Conforme POSIX
- Supporte la réplication, les quotas, la geo-réplication (réplication entres sites distants) et la détection de la dégradation des informations stockées
- Sans serveur(s) centralisé(s) de métadonnées
- Gestion simplifiée via une seule commande sur les serveurs : gluster

GlusterFS - Fonctionnement

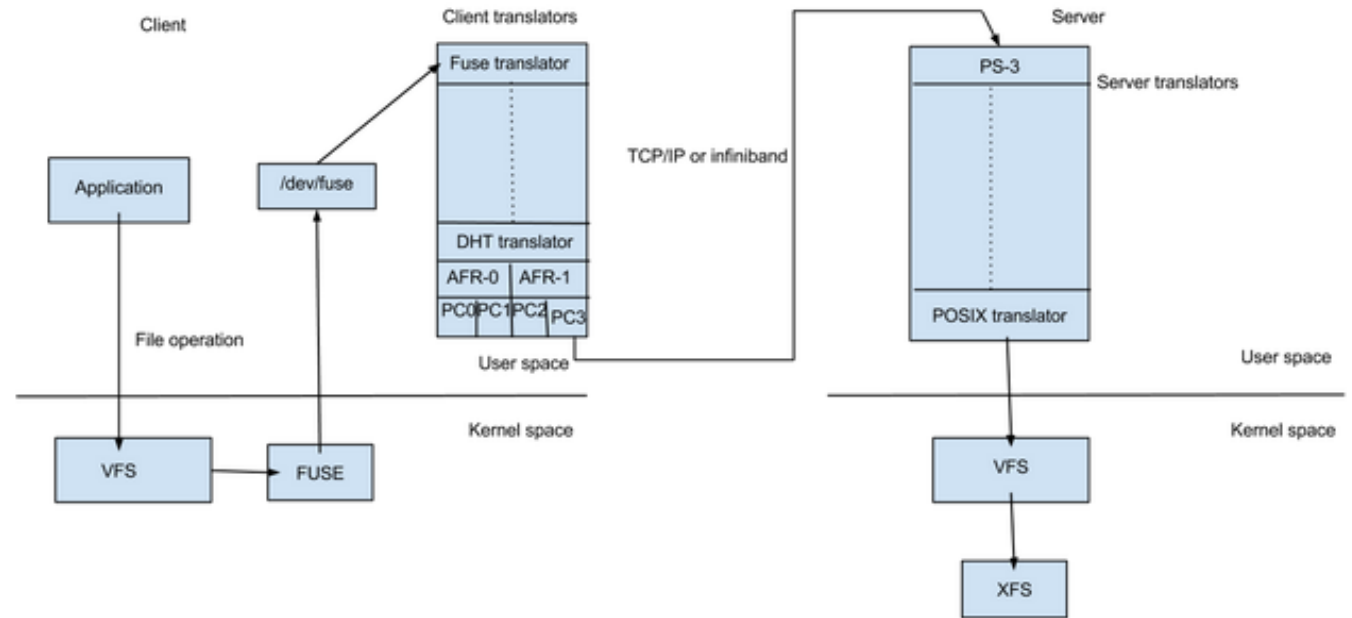
- Un volume est un assemblage de briques exportées par les démons glusterfsd
- Le volume contient les données montrées au client.
- Chaque volume est composé d'un ensemble de sous-volumes (sous volume élémentaire : la brique de stockage)



GlusterFS - Fonctionnement

Au moment du montage, le client glusterfs communique avec un des serveurs et récupère les fichiers de configuration du volume:

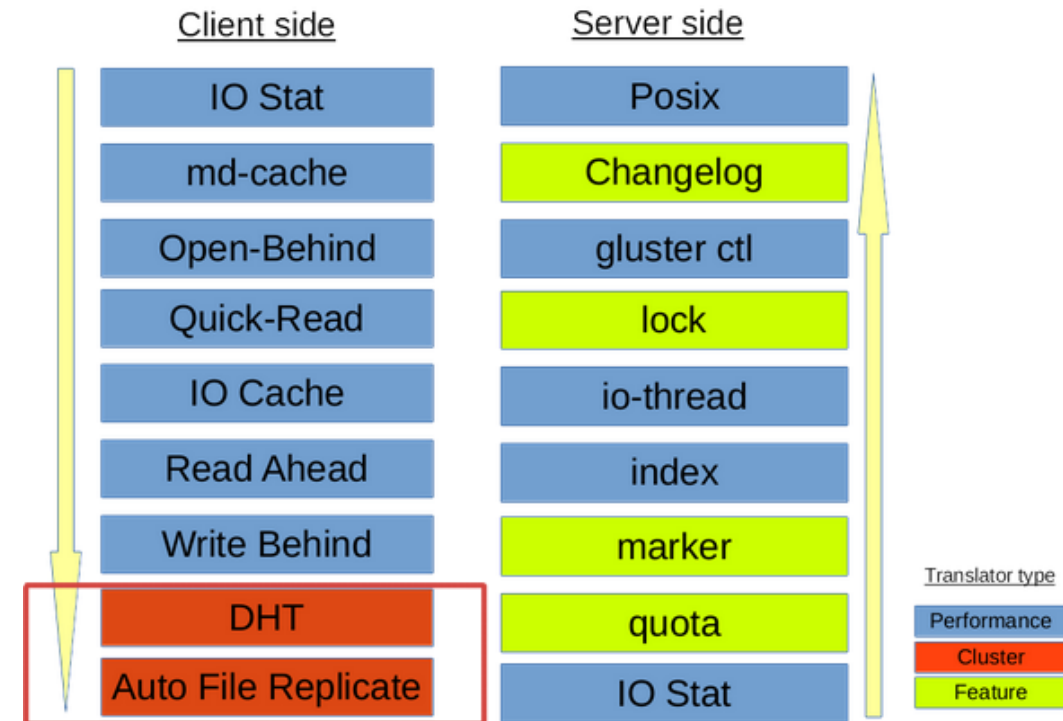
- La liste des traducteurs utilisés par le client (comment)
- Les informations pour joindre chacune des briques du volume (qui/où)



GlusterFS - Fonctionnement

Les traducteurs convertissent les requêtes utilisateurs en requêtes pour le stockage. Ils peuvent:

- Modifier la requête initiale (transformation de type ou modification des chemins, des drapeaux ou même des données ex: chiffrement)
- Intercepter ou bloquer une requête (contrôle d'accès)
- Créer une nouvelle requête



GlusterFS - Fonctionnement

Les traducteurs peuvent se classer en plusieurs catégories mais les principales sont:

- la catégorie Cluster
- La catégorie Performance

Parmi les traducteurs de la catégorie « cluster » on trouve:

- Le traducteur DHT (Distributed hash table)
- Le traducteur AFR (Automatic File Réplication)

CATEGORIE DE TRADUCTEURS	RÔLES
Storage	Traducteurs de plus bas niveau. Permet de stocker et d'accéder aux données du système de fichiers local.
Debug	Fournit une interface et des statistiques pour les erreurs et le débogage.
Cluster	Gère la distribution et la réplication des données pour l'écriture et la lecture des données à partir des briques et de nœuds.
Encryption	Traducteurs supplémentaires pour le chiffrement déchiffrement à la volée des données stockées.
Protocol	Traducteurs d'extension pour les protocoles de communication client / serveur.
Performance	Traducteurs d'optimisation / d'adaptation à la charge de travail et aux profils d'E / S.
Bindings	Traducteur d'interfaçage permettant d'étendre l'API d'interaction avec GlusterFS
System	Traducteurs d'accès au système (interfaçage avec le contrôle d'accès au système de fichiers, etc...)
Scheduler	Planificateurs d'E / S qui déterminent comment répartir les nouvelles opérations d'écriture sur les systèmes en cluster.
Features	Traducteurs permettant des fonctionnalités supplémentaires telles que les quotas, les filtres, les verrous, etc.

GlusterFS - Fonctionnement

DHT(Distributed Hash Table) Translator :

- Au cœur du fonctionnement de GlusterFS
- Responsable du placement de chaque fichier (ou morceau de fichier) sur exactement un de ses sous-volume (n'est pas responsable de la réplication ou du découpage)
- Fonctionne au moyen d'un hachage:
 - A chaque brique est assigné une plage dans un espace de hachage de 32 bits qui couvre toutes les briques (sans trous ni recouvrement).
 - A chaque fichier de la brique est assigné une valeur dans la plage de la brique (hash du nom du fichier). La connaissance du hash du fichier implique la connaissance de la brique sur lequel il se trouve.

GlusterFS - Fonctionnement

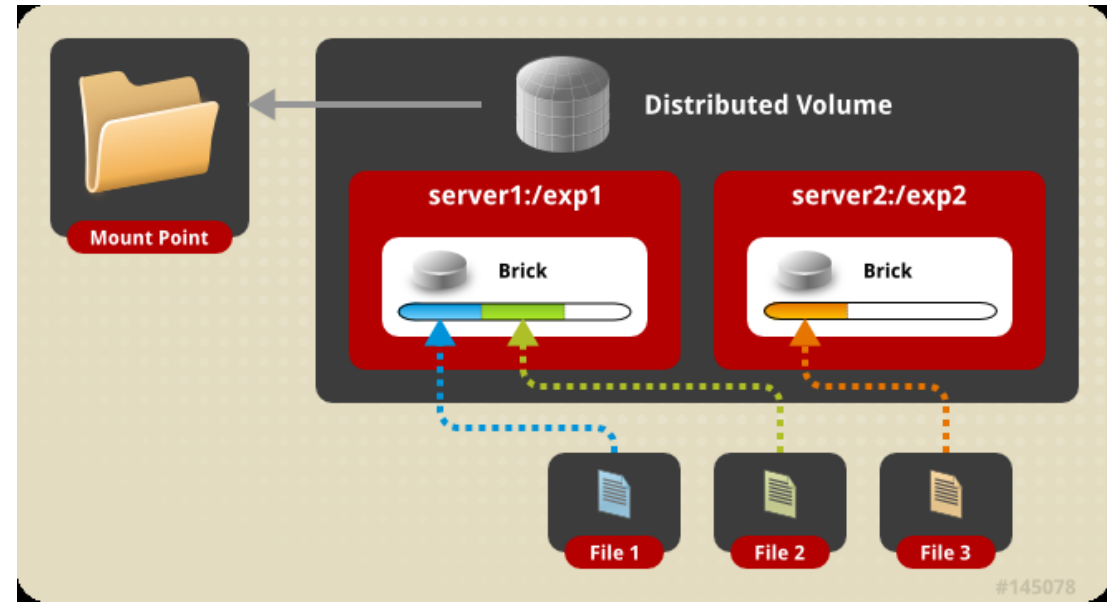
AFR(Automatic File Replication) Translator :

- Garde les traces des opérations sur les fichiers
- Responsable de la réplication des données à travers les briques :
 1. Doit maintenir de la cohérence de la réplication (ex: les données répliquées sur deux briques doivent être identiques même en cas d'opérations en parallèle depuis plusieurs applications / clients / points de montage, tant que toutes les briques du jeu de réplicas sont actives)
 2. Doit fournir un moyen de recouvrer les données en cas de dysfonctionnement tant qu'il reste au moins une brique qui contient les données correctes
 3. Doit fournir la dernière version des données pour les opérations sur les fichiers

GlusterFS – Modes possibles

Volume distribué:

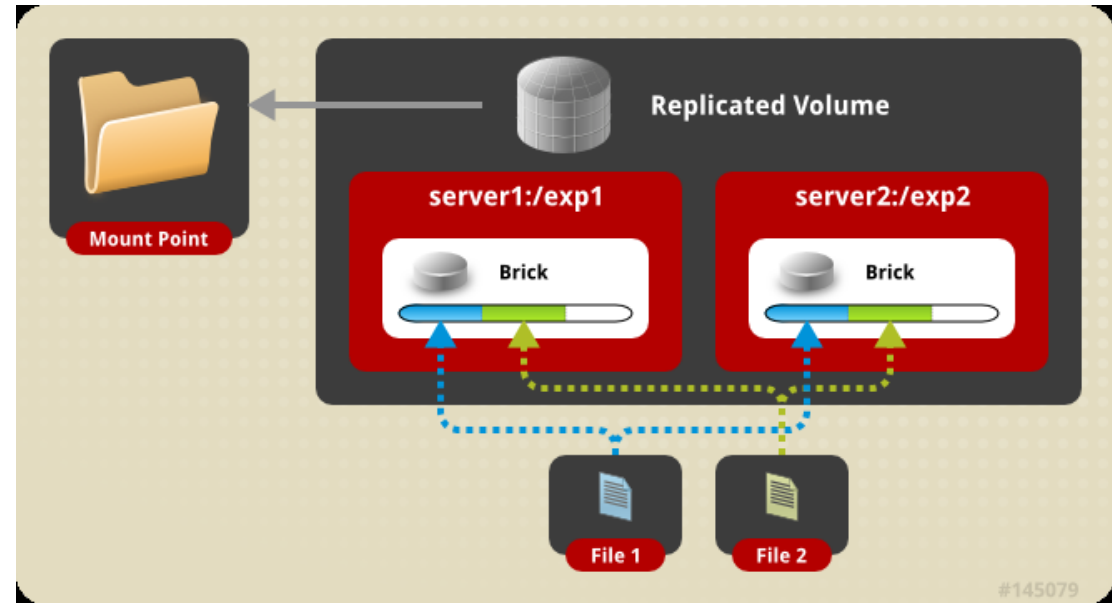
- Les fichiers sont affectés de façon uniforme sur les briques
- Grande performance
- Attention : pas de réplication:
 - Perte des fichiers présent sur une brique si elle devient indisponible.
 - Raid obligatoire pour le système de fichier local



GlusterFS – Modes possibles

Volume répliqué:

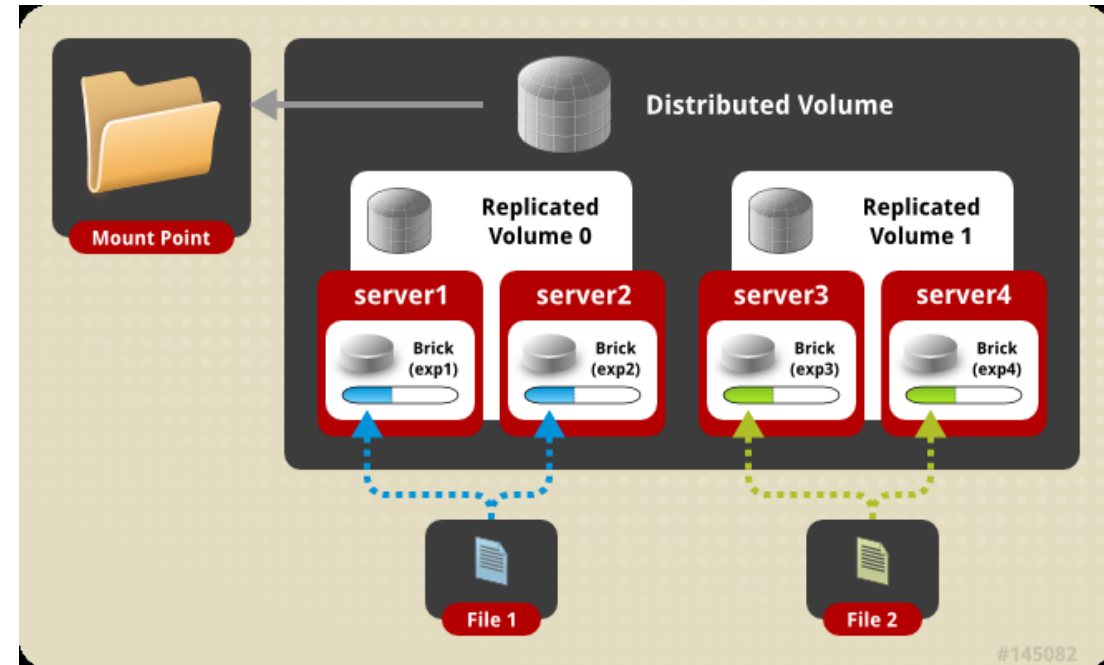
- Les fichiers répliqués de manière synchrone sur N nœuds
- Pénalise les performances en écriture (écriture terminée quand la réplication sur toutes les briques est terminée).
- Peut améliorer les performances en lecture: un fichier peut être lu sur plusieurs briques) mais les appels à « stat » sont pénalisés (stat sur toutes les briques pour obtenir la dernière version)



GlusterFS – Modes possibles

Volume distribué et répliqué:

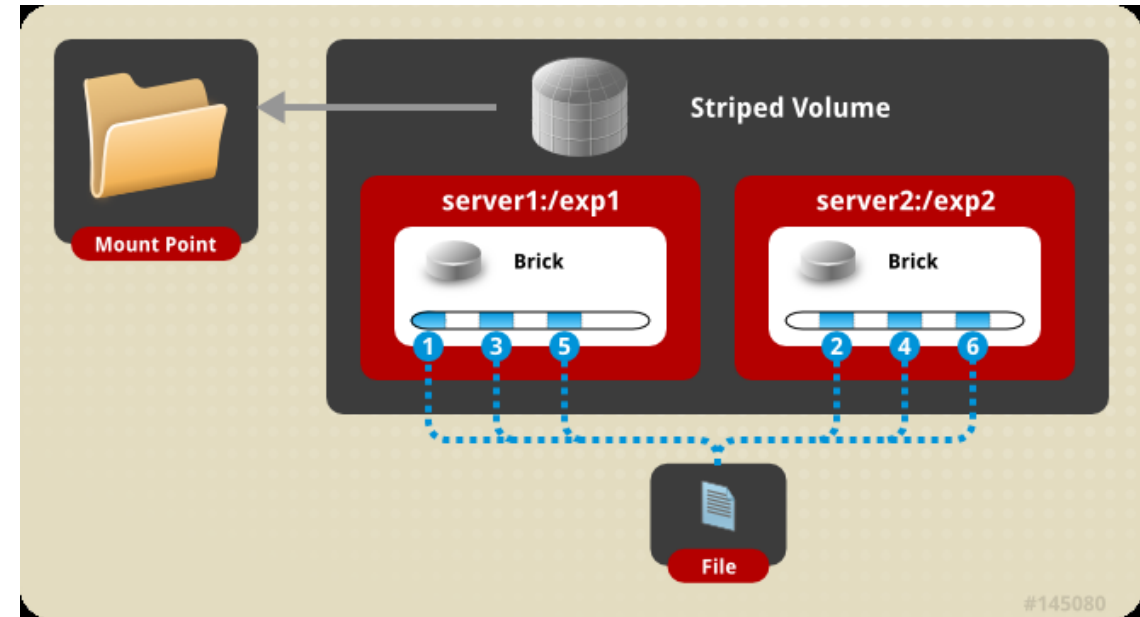
- Fichiers répliqués au sein de sous-volume qui sont assemblés dans un volume distribué.
- Bon compromis entre performance en lecture/écriture et haute disponibilité



GlusterFS – Modes possibles

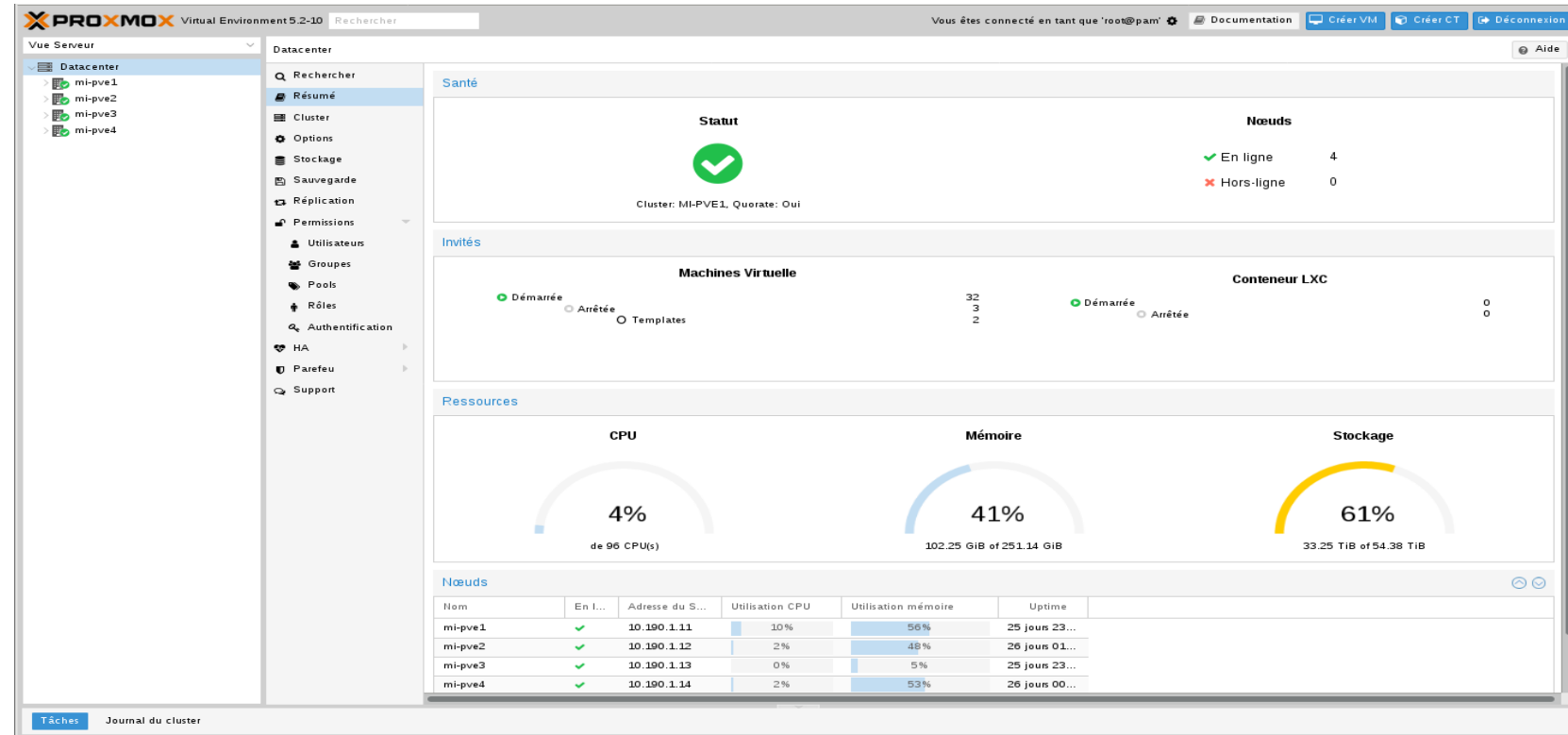
Volume découpé (éclaté ou emmental):

- Les fichiers sont découpés et les morceaux répartis entre les nœuds
- Améliore les performances (hébergement de disques de machines virtuelles)
- Perte de données catastrophique si une brique devient indisponible



Retour d'expérience

- Dans quel cadre ?
 - Cluster PROXMOX avec quatre nœuds sous utilisés (de la ram et du temps cpu dispo)
 - 3 nœuds disposants d'espace disque non utilisé (6To raid 5 matériel, disques sata non performants)
 - Interco 10Gbs
- Pourquoi ?
 - Facile à mettre en place
 - Robuste
 - Adaptable à l'usage
 - Performant si besoin



Retour d'expérience

Sélection du mode de fonctionnement du volume

☐ Besoin (psychologique) de robustesse

☐ Pas besoin de hautes performances

☐ ...et surtout, pas la possibilité / les moyens de mettre plus de serveurs

✓ Type de volume choisi : Réplica à trois membres

😞 Tant pis pour les performances...

Retour d'expérience

Installation des paquets nécessaires

- Configuration du dépôt d'installation (tous les serveurs et les clients)
 - PROXMOX 5.2 -> Debian Stretch
 - Utilisation du dépôt officiel GlusterFS pour Debian afin d'utiliser la dernière version

```
wget -O - https://download.gluster.org/pub/gluster/glusterfs/5/rsa.pub | apt-key add -  
DEBID=$(grep 'VERSION_ID=' /etc/os-release | cut -d '=' -f 2 | tr -d '"')  
DEBVER=$(grep 'VERSION=' /etc/os-release | grep -Eo '[a-z]+')  
DEBARCH=$(dpkg --print-architecture)  
echo deb https://download.gluster.org/pub/gluster/glusterfs/LATEST/Debian/${DEBID}/${DEBARCH}/apt \  
${DEBVER} main > /etc/apt/sources.list.d/gluster.list
```
- Clients

```
apt update && apt install glusterfs-client glusterfs-common
```
- Serveurs

```
apt update && apt install glusterfs-server
```

Retour d'expérience – Mise en place

Création des système de fichiers locaux sur les serveurs (volumes logiques formatés en xfs)

#/dev/sdX est un disque vierge

devicePath=« /dev/sdX"

deviceSize=5T

#On labélise le disque en GPT

(echo g; echo w) | fdisk \${devicePath}

#création de la partition

(echo n; echo p; echo 1; echo ; echo +\${deviceSize}; echo t; echo 8e; echo w) | fdisk \${devicePath}

devicePart=\${devicePath}1

#creation du volume physique

pvcreate \${devicePart}

#creation du groupe de volume

vgcreate vg-glbr1 \${devicePart}

yes | lvcreate -n lv-glbr1 -l 100%FREE vg-glbr1

mkfs.xfs -f -i size=512 /dev/vg-glbr1/lv-glbr1

Retour d'expérience – Mise en place

Création du point de montage et montage de la partition (tous les serveurs)

```
deviceMountPath=/exportfs/glusterfs/brick1
```

```
mkdir -p ${deviceMountPath}
```

```
mount /dev/vg-glbr1/lv-glbr1 ${deviceMountPath}
```

Création du répertoire exporté par GlusterFS (tous les serveurs)

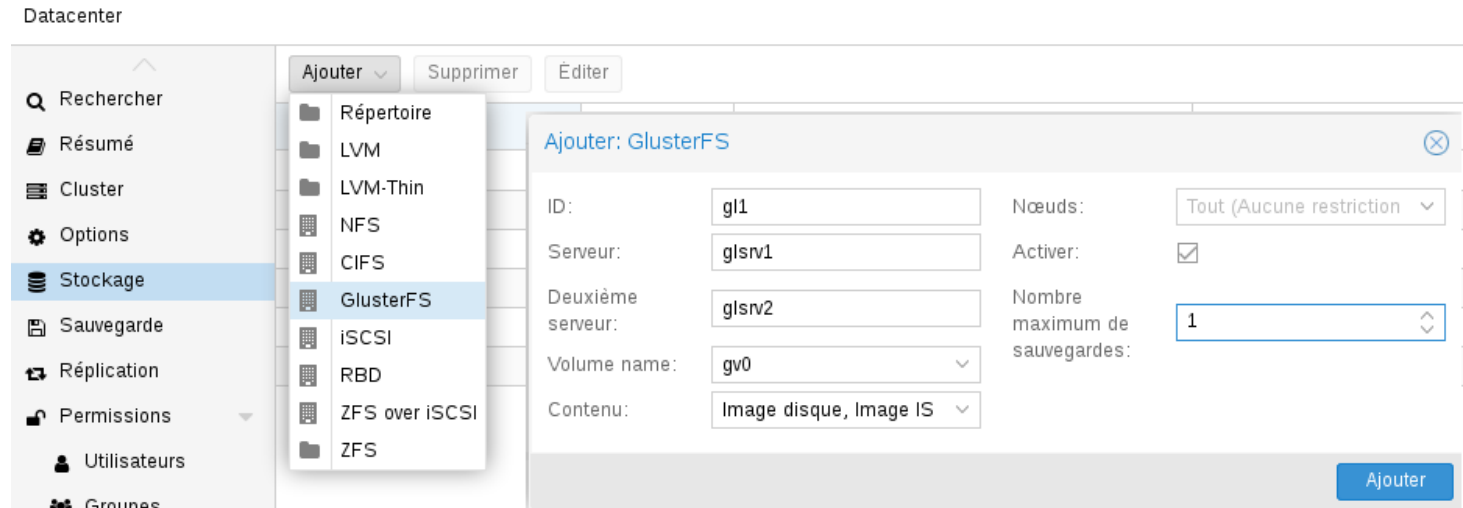
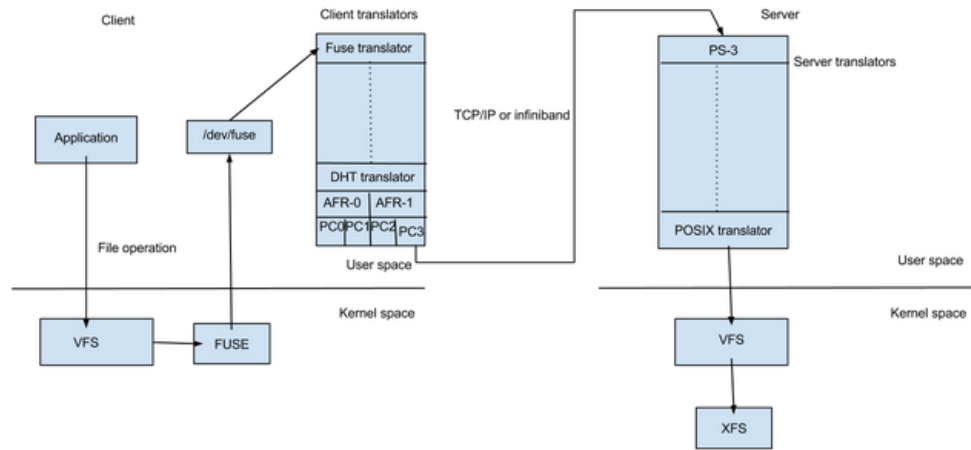
```
mkdir -p ${deviceMountPath}/brick
```

Création et démarrage du volume (sur un serveur)

```
gluster volume create nomVolume replica 3 nœud{1..3} =/exportfs/glusterfs/brick1/brick
```

```
gluster volume start
```

Retour d'expérience – Mise en place



Montage via l'interface PROXMOX

Ou manuellement...

mount -t glusterfs HOSTNAME-OR-IPADDRESS:/VOLNAME MOUNTDIR

...ou dans le fichier fstab pour un montage automatique au boot

HOSTNAME-OR-IPADDRESS:/VOLNAME MOUNTDIR glusterfs defaults,_netdev 0 0

Retour d'expérience – Surveillance

Obtenir des informations des nœuds / d'un volume (sur un des serveurs)

- Informations détaillées sur glusterfsd

gluster get-state detail (informations détaillées sur glusterfsd et sur le volume)

- Informations détaillées sur les options d'un volume

gluster get-state volumeoptions

- Etat d'un volume

gluster volume status

- Santé des données d'un volume

gluster volume heal <NOM_VOLUME> info

Retour d'expérience – En cas de (petit) pépin

Que faire si une brique est partie proprement en vacances mais a fini par revenir (reboot d'un nœud, par exemple pour update) ?

Pour qu'une brique soit à nouveau active (qu'elle réplique) dans le volume, il faut que les fichiers qui la composent soient synchronisés avec les fichiers des autres briques.

Au retour de la brique en défaut, il faut resynchroniser les données (fait automatiquement toutes les 10min dans les dernières version de GlusterFS):

- Juste les fichiers nécessitant d'être resynchronisés:

gluster volume heal <VOLNAME>

- Tous les fichiers du volume

gluster volume heal <VOLNAME> full

Retour d'expérience – En cas de (petit ?) pépin

- Obtenir la liste des fichiers resynchronisés

gluster volume heal <VOLNAME> info healed

- Obtenir la liste des fichiers non resynchronisés

gluster volume heal <VOLNAME> info failed

- Obtenir la liste des fichiers en état de split-brain

gluster volume heal <VOLNAME> info split-brain

Retour d'expérience – En cas de gros pépin

Malgré la lecture de la doc sur « comment supprimer et remplacer une brique défaillante », votre volume ne veut plus fonctionner ?

Tant qu'un nœud de stockage est fonctionnel et qu'il n'y a pas de perte du système de fichier local (celui exporté par glusterfs), vous avez accès aux données.

Pour le choix du fs (xfs...ou autre), ne négligez pas celui qui vous est le plus familier, au cas où il faille réparer...

Retour d'expérience – Test de performances

Finalité : avoir une idée des perfs de GlusterFS face à NFS (et à un système de fichier local) sur des bases matérielles identiques (raid 5, disques capacitifs peu véloce)

Outil utilisé : FIO

Mode opératoire :

- 4 tests exécutés en simultanés
 - écriture
 - lecture
 - écriture aléatoire
 - lecture aléatoire
- 1 fichier de 128G
- taille de block de 4k

Retour d'expérience – Test de performances

ECRITURE							LECTURE					
	BW (KB/S)	IOPS	lat min (μ s)	lat moy (μ s)	lat max (μ s)	95 % clat (μ s)	BW (KB/S)	IOPS	lat min (μ s)	lat moy (μ s)	lat max (μ s)	95 % clat (μ s)
xfs	18467	4615	1000	80000	6930	15168	2067100	529395	3	60	13426	96
nfs	51418	12853	305	2487	27890	9792	53038	13258	288	2410	853843	3312
glusterfs	21941	5484	24	5753	309816	828	74029	18506	129	1727	51127	51127

ECRITURE ALEATOIRE							LECTURE ALEATOIRE					
	BW (KB/S)	IOPS	lat min (μ s)	lat moy (μ s)	lat max (μ s)	95 % clat (μ s)	BW (KB/S)	IOPS	lat min (μ s)	lat moy (μ s)	lat max (μ s)	95 % clat (μ s)
xfs	1695	422	10000	75050	395000	161000	1124200	287787	1	107	1205400	23
nfs	5377	1343	336	25663	3333900	201728	93	21	38000	16039400	2306000	7767000
glusterfs	340	81	9000	380650	1124000	1106000	79010	19751	150	83012	1617	1784

Retour d'expérience – Test de performances

A la réflexion, j'aurais peut être dû rester sur la version de GlusterFS livrée avec PROXMOX...

Ajouter ▼DétacherEditerRe-dimensionner le disqueDéplacer le disqueRevenir en arrière

Disposition du clavier	Défaut
Mémoire	2.00 GiB [balloon=0]
Processeurs	4 (1 sockets, 4 cores)
Affichage	Défaut
Lecteur CD/DVD (ide2)	none,media=cdrom
Disque Dur (sata0)	rks2:18120501/vm-18120501-disk-0.qcow2,size=320G
Carte réseau (net0)	e1000=C2:6F:CF:95:DC:AB,bridge=vmbri190
Disque inutilisé 0	rks4:18120501/vm-18120501-disk-0.qcow2

Task viewer: VM 18120501 - Déplacer le disque

SortieStatut

Stopper

create full clone of drive sata0 (rks2:18120501/vm-18120501-disk-0.qcow2)
[2018-12-12 15:59:13.687524] E [MSGID: 101191] [event-epoll.c:671:event_dispatch_epoll_worker] 0-epoll: Failed to dispatch handler
[2018-12-12 15:59:13.796545] E [MSGID: 108006] [afr-common.c:5302: __afr_handle_child_down_event] 0-gv0-replicate-0: All subvolumes are down
The message "E [MSGID: 101191] [event-epoll.c:671:event_dispatch_epoll_worker] 0-epoll: Failed to dispatch handler" repeated 6 times between [2018-12-12 15:59:14.681988] E [MSGID: 101191] [event-epoll.c:671:event_dispatch_epoll_worker] 0-epoll: Failed to dispatch handler
[2018-12-12 15:59:34.981159] E [MSGID: 108006] [afr-common.c:5302: __afr_handle_child_down_event] 0-gv0-replicate-0: All subvolumes are down
Formatting 'gluster://glsvr1/gv0/images/18120501/vm-18120501-disk-0.qcow2', fmt=qcow2 size=343597383680 cluster_size=65536 preallocation=me
The message "E [MSGID: 101191] [event-epoll.c:671:event_dispatch_epoll_worker] 0-epoll: Failed to dispatch handler" repeated 19 times between [2018-12-12 15:59:35.736972] E [MSGID: 101191] [event-epoll.c:671:event_dispatch_epoll_worker] 0-epoll: Failed to dispatch handler
transferred: 0 bytes remaining: 343597383680 bytes total: 343597383680 bytes progression: 0.00 %
transferred: 6871947673 bytes remaining: 336725436007 bytes total: 343597383680 bytes progression: 2.00 %
transferred: 10307921510 bytes remaining: 333289462170 bytes total: 343597383680 bytes progression: 3.00 %
The message "E [MSGID: 101191] [event-epoll.c:671:event_dispatch_epoll_worker] 0-epoll: Failed to dispatch handler" repeated 20 times between [2018-12-12 15:59:36.736972] E [MSGID: 101191] [event-epoll.c:671:event_dispatch_epoll_worker] 0-epoll: Failed to dispatch handler
transferred: 13743895347 bytes remaining: 329853488333 bytes total: 343597383680 bytes progression: 4.00 %
transferred: 17179869184 bytes remaining: 326417514496 bytes total: 343597383680 bytes progression: 5.00 %
transferred: 20615812820 bytes remaining: 326261672060 bytes total: 343597383680 bytes progression: 6.00 %

Déc 12 16:59:13	Déc 12 17:48:32	mi-pve3	root@pam	VM 18120501 - Déplacer le disque
Déc 12 11:31:21	Déc 12 12:21:04	mi-pve3	root@pam	VM 18120501 - Déplacer le disque

Conclusion

GlusterFS, c'est

- Facile
- Fiable
- Robuste
- Pour les perfs, il faudra attendre la v3:
 - Passage à un mode répliqué - réparti
 - Traducteur client md-cache (utilisation de cache SSD)
 - Utilisation de disques SAS 15k et de contrôleurs SAS plus performant

Références

GlusterFS

- <https://www.gluster.org/>
- <https://docs.gluster.org/en/latest/>

FIO

- <https://github.com/axboe/fio>
- <https://fio.readthedocs.io/en/latest/>