BeeGFS au LCPQ

David Sanchez

Capitoul - 13 décembre 2018

- 4 clusters de calcul, dont un seul avec un stockage distribué.
- Lustre version 2009, 10To :
  - Rapide pour l'époque, mais commençait à s'essouffler.
  - Forte adhérence au noyau, nécessaire de recompiler.
  - Crash de Lustre en cas de crash d'un disque ou du module noyau, reboot nécessaire.

- GNU/Linux.
- Le moins d'adhérences au noyau.
- Pouvoir utiliser diverses interfaces réseaux : ethernet et InfiniBand.
- Performances.
- Stabilité.
- Facilité d'administration.

BeeGFS

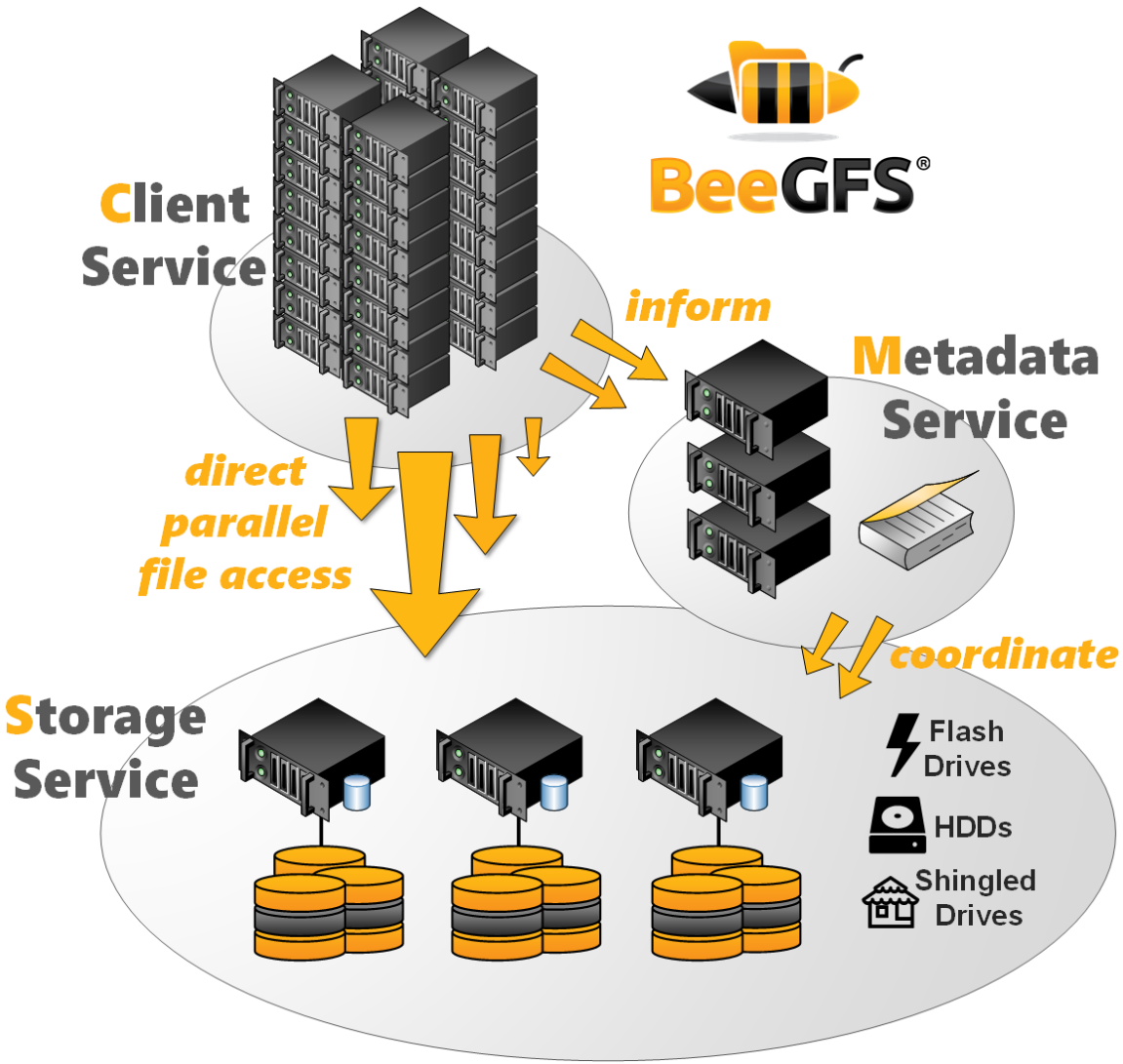
Historique de BeeGFS

- Début du projet en 2005 : Fraunhofer Gesellschaft File System (FhGFS).
- Création de ThinkParQ en 2014 pour commercialiser la solution.
- Il existe des installations de 1Po et plus.
- L'université de Strasbourg l'utilise pour /scratch et /home (après soucis avec RozoFS).
- Version actuelle (24/09/2018) : 7.1 (7.0 au LCPQ).
- BeeOND / "BeeGFS On Demand" : utiliser les disques des nœuds.

Architecture BeeGFS

- mgmtd : démon de gestion/administration du cluster.
- storage : démon pour le stockage des données.
- meta : démon pour le stockage des métadonnées.
- client : démon pour monter le système de fichier distribué.
- helperd : userspace pour les clients.

Architecture BeeGFS



- Faire du "tiering" SSDs / HDDs : "Storage Pools"
- Plutôt que de faire des caches avec SSDs puis stockage sur HDDs, on fait des silos.
- Ex : le dossier des calculs actifs utilise le silo SSD tandis que les calculs terminés sont déplacés dans le silo HDD.
- On peut aussi utiliser différents silos dans le même dossier !
- Ex : on crée un fichier dans le silo SSD au début du calcul, puis à la fin de celui-ci on le migre dans le silo HDD.

Architecture BeeGFS LCPQ

- 3 baies pour serveurs.
- 1 switch Dell N4064 par baie, 48x10Gb + 2x40Gb.
- 40Gb en anneau pour transferts/communications.
- Switch InfiniBand pour les serveurs restants sur cette technologie.
- 55 Dell R[620/630/640/710/730/740].
- HP Moonshot.
- Rocks Clusters (Centos 7).

Machines BeeGFS

1 Dell R520 pour le management

C'est aussi le contrôleur SLURM, login, etc
Ethernet et InfiniBand.

1 Dell R630 pour le MDS

2 SSD / MLC / 200GB / raid 1.
Ethernet et InfiniBand.

2 Dell R630 pour les OSS

10 HDD / 10k tpm / 1,8TB / indépendants.
Ethernet et InfiniBand.

Total : 20 disques durs, 33TB utilisables sur le cluster.

Installation de BeeGFS

- Facile : ajout du dépôt BeeGFS sur les machines et installer les paquets nécessaires au type de machine.
- Pour les machines IB : ajout des paquets InfiniBand (ex : mlx4), ainsi au lancement il détectera la carte, et il ajoutera le support InfiniBand au démon concerné.

- Changer l'ordonnanceur noyau par défaut : *deadline*.
- Si on utilise des cartes raid avec cache, et qu'on veut leur faire confiance, utiliser : *noop*.
- Système de fichiers : *xfs* sur les OSS, *ext4* sur les MDS.

- Chunk : 2M
- Targets : 4
- Permettre avec 5 fichiers d'avoir le max de performance (20HDDs / 4 cibles = 5)
- Pour connaître la configuration sur un système donné :

```
beegfs-ctl --getentryinfo /beegfs
```

Tester avec des fichiers de moins de 2Go est inutile dans mon cas, les cartes raid ont 2Go de cache.

- mono-nœud : de 416Mo/s (1 job) à 275Mo/s (4 jobs). Saturation du lien 10G à 4 jobs.
- multi-nœuds : voir tableau (en Mo/s).

Nœuds	2	3	4	5	6
Débit	386	354	318	353	297
Cumulé	772	1062	1272	1765	1782

Mesures avec fio.

- 50 µs pour un SSD.
- 100 µs pour BeeGFS sur infiniband.
- 250 µs pour BeeGFS sur 10GbE.
- 500 µs pour BeeGFS sur GbE.
- 4-10 ms pour un disque 10k tpm.

Faire du stockage distribué en 10GbE sur une petite grappe de calcul est possible.

Fin

Des questions ?

Références

https ://www.beegfs.io/
https ://www.beegfs.io/wiki/BeeOND
https ://www.beegfs.io/wiki/StoragePools
https ://www.beegfs.io/wiki/Upgrade6xOr70To71
https ://www.beegfs.io/wiki/AboutMirroring
https ://www.beegfs.io/wiki/EnableQuota