Stockage distribué @ LCPQ

TutoJRES n°18 – 4 mai 2016

David Sanchez
<david.sanchez@irsamc.ups-tlse.fr>

Laboratoire de Chimie et Physique Quantiques

Le LCPQ

- **Laboratoire de Chimie et Physique Quantiques.**
- **UMR Université Toulouse 3 – Paul Sabatier / CNRS.**
- **Recherche théorique.**
- **Calcul scientifique.**
- **http://www.lcpq.ups-tlse.fr/**

A mon arrivée fin 2013

- **4 clusters.**
- **2 salles « maison ».**
- **Projet de nouvelle salle unique dans notre bâtiment.**
- **Rebondissement de dernière minute : le projet de salle est annulé !**

Plan B

- **Déménagement dans le datacenter de la DSI.**
- **Au final c'est bien mieux que le projet de nouvelle salle dans notre bâtiment.**

L'infra en 2014

- **4 clusters sous Rocks Clusters, en v5 et v6.**
- **Dont un cluster qui dispose d'un stockage lustre sur infiniband.**
- **Les autres n'ont que du NFS entre le maître et les nœuds de calcul.**
- **Un cinquième cluster vient s'ajouter (ANR privatisé).**

Les baies du LCPQ

Et si on fusionnait ?

- **Fusion des 4 clusters en un seul.**
- **Facilité d'administration et d'utilisation.**
- **Mettre à disposition de tous les nœuds de calcul un stockage distribué ?**

Jusqu'à maintenant on avait Lustre...

- **Sauf que compliqué à mettre en place à l'époque de l'installation du cluster l'utilisant.**
- **Upgrade très compliqué (kernel lustre nécessaire sur tous les nœuds clients).**
- **Lustre version 2009 : instable.**
- **Module noyau, kernel panic, reboot oss.**
- **Et si on regardait ailleurs ?**

Critères de sélection

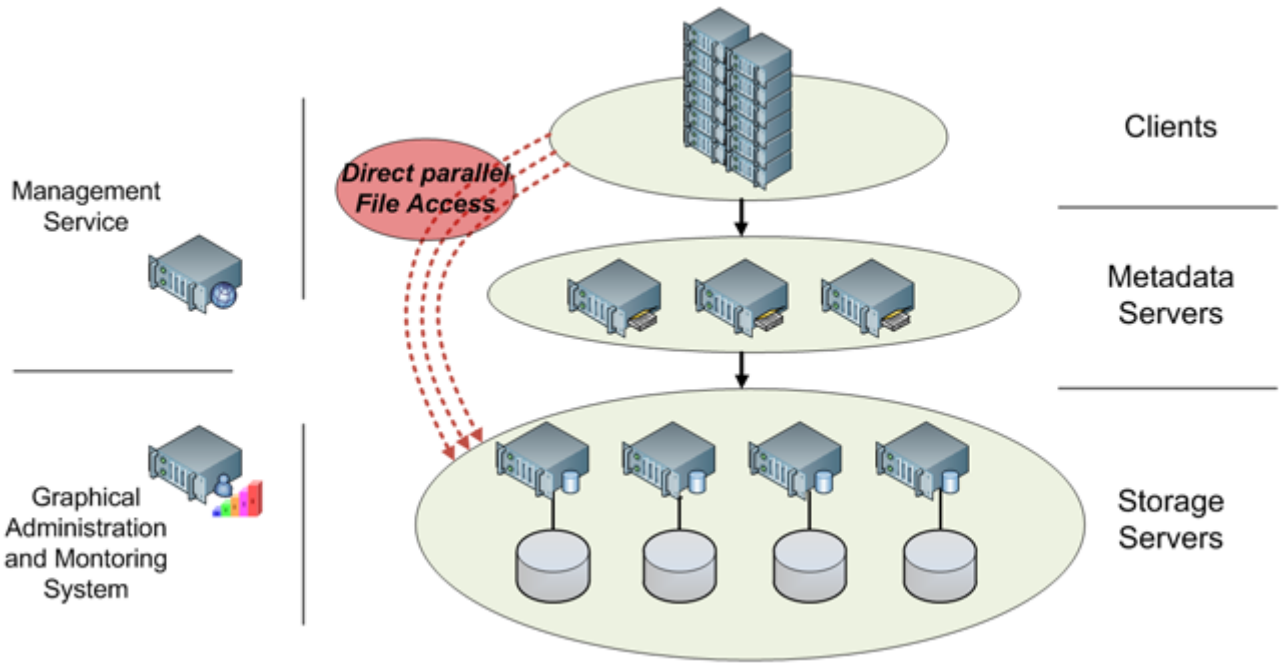
- **Sous GNU/Linux.**
- **Non lié au kernel.**
- **Utiliser divers protocoles réseaux : ethernet, InfiniBand.**
- **Performances.**
- **Stabilité.**
- **Facilité d'administration.**

Le choix : BeeGFS

- **Ex-FhGFS.**
- **Système de fichiers conçu pour le calcul distribué.**
- **L'institut du Climat de Catalogne l'utilise sur une solution de 1PB qui pourra être augmentée jusqu'à 10PB.**
- **Découpage des fichiers pour stockage sur multiples serveurs de stockage.**
- **Utilisation des disques locaux des nœuds de calcul (ex : Strasbourg).**

Schéma architecture

- **http://www.beegfs.com/wiki/SystemArchitecture**



Architecture d'une solution BeeGFS

- **beegfs-mgmtd : démon de gestion du cluster de stockage (sur le serveur maître).**
- **beegfs-storage : démon pour le stockage des données (sur les oss).**
- **beegfs-meta : démon pour la gestion des métadonnées (sur le mds).**
- **beegfs-client : démon pour monter le système de fichiers sur les nœuds de calcul.**
- **beegfs-helperd : démon pour la partie espace utilisateur du client (sur nœuds de calcul).**

L'architecture du LCPQ

- **3 baies.**
- **1 switch Dell N4064 par baie : 48 ports 10GbE, 2 uplinks 40Gb.**
  - Anneau pour interconnecter les 3 switchs.
- **Réutilisation du switch InfiniBand sur les nœuds non dotés d'une carte réseau 10GbE.**
- **Nœuds de calcul :**
  - Dell R620, R630, R710, R730, R815, R910, HP DL180G6.
  - Chassis HPe moonshot.
- **NAS maison : Dell R530.**
- **Serveur maître : Dell R520.**

Architecture de stockage

- **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 HDD, 33Tb utilisable sur le cluster.
- **112 nœuds de calcul :**
  - Diverses machines : voir slide précédent.
  - 10GbE ou InfiniBand, sauf Moonshot en GbE.

Installation de la solution

- **Alignement sur rocks clusters : Centos 6.**
- **Il suffit d'ajouter un dépôt yum pour installer les paquets nécessaires.**
- **Les oss et le mds ont des cartes raid LSI/Dell Perc H730/H730p (12Gbps).**
- **La documentation recommande comme scheduler : deadline.**
- **On fait confiance aux cartes raid : noop.**

Sur les clients

- **Installation de beegfs-client.**
- **Ajout de paquets pour faire du verbs sur les machines InfiniBand. (libmlx4 par exemple).**
- **A l'installation, détection présence d'InfiniBand sur le nœud.**
  - Si présence : ajout lien symbolique pour compilation support InfiniBand.
- **Au premier lancement du client, compilation pour s'adapter à la machine.**

Tuning

- **Adaptable par fichiers et dossiers.**
- **Striping**
  - Chunk = 2M
  - Targets = 4
- **Compromis pour avoir plusieurs jobs en simultanés sans gêne entre les jobs.**
- **On peut connaître la configuration avec la commande :**

  $ beegfs-ctl --getentryinfo /mnt/beegfs
- **Pour modifier la configuration**

  $ beegfs-ctl --setpattern –numtargets=4 \
  --chunksize=2m /mnt/beegfs

Tuning – OSS

- **tuneNumWorkers = nombre de disques.**
- **Passage de l'IO scheduler en noop.**
- **Augmentation du nombre de requêtes :**

  $ echo 4096 > /sys/block/sdX/queue/nr_requests
- **Augmentation du nombre de lectures en séquentiel :**

  $ echo 4096 > /sys/block/sdX/queue/read_ahead_kb
- **Modification des caches dans le noyau :**

  $ echo 5 > /proc/sys/vm/dirty_background_ratio

  $ echo 10 > /proc/sys/vm/dirty_ratio

Tuning – MDS

- **ext4 pour la partition/volume qui contiendra les métadonnées.**
- **Activation des attributs étendus s'ils ne le sont pas.**
  - tune2fs -o user_xattr /dev/sdX
- **Changement d'IO scheduler : deadline recommandé sur le wiki, noop utilisé ici (carte raid hardware).**
- **tuneNumWorkers = 64 pour 100-200 nodes, 128 pour plus.**

Performances

(A ± 30Mo/s)

Performances – séquentiel

**$ dd if=/dev/zero **
of=/mnt/beegfs/dsanchez/benchmark/node1/test bs=4k count=10000000

(41 GB) copied, 109.886 s, 373 MB/s

- En dessous de 2Go, aucun intérêt de bench, les caches des cartes faussent les tests.

Performances – parallèle

- Sur un seul nœud de calcul, X dd :

dd	1	2	3	4
Débit	416 Mo/s	415 Mo/s	322 Mo/s	260 Mo/s
Cumulé	416 Mo/s	830 Mo/s	966 Mo/s	1040 Mo/s

Performances – parallèle

- Sur X nœuds, un dd par nœud.

Nombre de nœuds	2	3	4	5	6
Débit	386 Mo/s	354 Mo/s	318 Mo/s	353 Mo/s	297 Mo/s
Cumul	772 Mo/s	1062 Mo/s	1272 Mo/s	1765 Mo/s	1782 Mo/s

- Les machines Intel font 350 Mo/s de moyenne, celle en AMD 230 Mo/s (à partir de 4 nœuds).
- Sur 5 nœuds, un nœud fait tourner 2 dd :
  - 328 Mo/s pour les deux dd sur la même machine.
- Sur 6 nœuds, dont le serveur maître et la machine AMD, un dd par nœud sauf un qui en fait tourner 2.

Performances – parallèle

- **Sur 3 nœuds, 3 dd par nœud :**
  - 259 Mo/s de moyenne par dd.
- **Débit cumulé : 2,2Go/s, saturation des interfaces 10GbE des OSS.**

Performances – IB DDR

Nombre de dd	1	2	3	4	5
Débit	319 Mo/s	283 Mo/s	238 Mo/s	211 Mo/s	190 Mo/s
Cumul	319 Mo/s	566 Mo/s	714 Mo/s	844 Mo/s	950 Mo/s

Performances – latences

- **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 un petit cluster est possible et utilisable.**

Fin

Merci !

Questions ?