Stockage de données LCG : choix-configuration-optimisation

Eléments de discussion.
Merci à tous ceux qui ont apporté leur contribution

Jean-Michel BARBET, Laboratoire SUBATECH, Nantes

Checklist

- Configuration matérielle
- Configuration RAID
- Partitions
- Systèmes de fichiers
- Optimisation I/O au niveau kernel
- Mesurer la performance

Configuration matérielle

- Type de disque (SATA,SAS,SSD)/ capacité / nombre / bus
- Contrôleur RAID (fonctionalités / performances / taille des caches)
- CPU / mémoire
- Interfaces réseau
  - Équilibre entre bande passante et performances de lecture/écriture des données de chaque serveur
  - Équilibre de l'ensemble du cluster, réseau compris

Configuration RAID

- Niveau de RAID : en général RAID6
- Stripe-element size
- Caches, read-ahead, write cache policy
- Autres réglages ?

Partitions

- Partitionner ou non
- Alignement de partitions [3]

Systèmes de fichiers

- Choix : EXT4, XFS, autres (ZFS en mode JBOD)
- Configuration :
  - Block size (4KB Linux sans HugePages)
  - No atime ?
  - Tenir compte du stripe-width ?
  - Write barriers
  - Taille des métadonnées en accord avec le nombre de fichiers
    - Inode64 (XFS)

Optimisation I/O niveau Kernel

Travaux de J.Pansanel & C.Leroy [1], de C.Diarra [2]

- Algorithme de gestion (I/O scheduler)
  - Choix entre : cfq(default), noop ou deadline (préconisé ?)
- Read ahead
  - getra : 16384
- Taille de la queue d'I/O
  - queue_depth : 256
- nr_requests
  - Supérieur à la taille de la queue d'IO [4] : 512

Notion de « workload »

- Répartition statistique de la taille de fichiers
  - histogramme
- Fréquence des accès en lecture et en écriture
- Mode d'accès : séquentiel ou par morceaux
  - ROOT « baskets » ?
- Lien avec le modèle de calcul
  - types de données (ESD,AOD,conditions,logs)
  - Activités (simulation, analyse,...)

Mesurer la performances

- Quels tests facilement utilisables par tous ?
- Premiers candidats : dd, fio, iozone
- Disposer d'un mode opératoire pour un jeu de tests standard

Sécurité des données

- Le risque de perte simultanée de plusieurs disques
  - temps de reconstruction
  - Spare disk(s)
- Le risque de « punctured RAID », Patrol, surveillance, predictive failure
- Caches et intégrité des données
-

Conclusion

- Disposer d'une méthodologie de configuration
- Pouvoir faire des comparaisons

Références

[1] Storage benchmarking and tuning
http://lcg.in2p3.fr/wiki/index.php?title=Storage-Benches

[2] Linux Kernel Tuning: TCP & disk I/O
http://lcg.in2p3.fr/wiki/index.php?title=TCP-Tuning#Tuning_disk_I.2FO

[3] How to align partitions for best performance using parted
http://rainbow.chard.org/2013/01/30/how-to-align-partitions-for-best-performance-using-parted/

[4] O+P Insights : Linux Hardware RAID Howto
http://insights.oetiker.ch/linux/raidoptimization/

[5] Aligning IO on a hard disk RAID – the Theory
https://www.percona.com/blog/2011/06/09/aligning-io-on-a-hard-disk-raid-the-theory/