

Ceph: Un Voyage vers 1 To/s

Par Mark Nelson

En 2023, Clyso a collaboré avec une entreprise avant-gardiste pour faire la transition de leur cluster Ceph d'un stockage basé sur des disques durs (HDD) vers un déploiement NVMe de 10 Pétaoctets (PB). Cette affiche décrit le parcours pour atteindre un débit record de 1 To/s avec un cluster Ceph, détaillant les défis rencontrés et les solutions mises en œuvre. Le projet a démontré le potentiel de Ceph à offrir des performances extrêmes sur une infrastructure matérielle et réseau bien conçue.

01. Introduction



Le projet a commencé lorsqu'un client a contacté Clyso avec l'exigence unique de migrer son cluster Ceph existant vers une configuration haute performance basée sur NVMe. Le cluster devait fonctionner de manière transparente sur 17 racks, sans interruption de service. Cette affiche donne un aperçu de la configuration du cluster, de l'optimisation des performances et des méthodologies de test qui ont permis d'atteindre un débit sans précédent.

02. Configuration du Cluster



Configuration Matérielle:

- **Nœuds:** 68 x Dell PowerEdge R6615
- **CPU:** 1 x AMD EPYC 9454P (48C/96T)
- **Mémoire:** 192 Go DDR5 RAM
- **Réseau:** 2 x 100 GbE Mellanox ConnectX-6
- **Stockage:** 10 x Dell 15.36 To NVMe Entreprise

Configuration Logicielle:

- **OS:** Ubuntu 20.04.6 (Focal)
- **Version Ceph:** Quincy v17.2.7 (Paquets Deb Upstream)

Considérations de Conception:

- Consommation électrique limitée à 1000-1500W par rack.
- Intégration dans un cluster existant sans interruption de service.

03. Configuration des Tests



Tests de Rodage:

- Clusters Ceph éphémères déployés à l'aide de CBT (Ceph Benchmarking Tool).
- OSDs configurés avec 8 Go `osd_memory_target`.

Évaluation des Performances:

- Utilisation de FIO avec librbd pour les tests de performance.
- Tests effectués avec une réplication x3 et un codage d'effacement (erasure coding) EC 6+2.
- Concentration sur des tests I/O de 4 Mo et 4 Ko pour une évaluation complète des performances.

Considérations Spéciales:

Mode sécurisé activé avec le chiffrement msgr V2.

04. Analyse des Performances



Problèmes Initiaux:

- Variabilité des performances entre les nœuds.
- Goulots d'étranglement au niveau du noyau liés à l'IOMMU et aux transitions d'état C du CPU (CPU c-state transitions).

Correctifs Implémentés:

- Désactivation des c-states du CPU pour une latence améliorée.
- Désactivation de l'IOMMU dans le noyau pour résoudre la contention de verrouillage tournant (spin lock contention).
- Recompile de RocksDB avec des flags optimisés pour améliorer la performance d'écriture aléatoire 4 Ko.

Ceph: Un Voyage vers 1 To/s

Par Mark Nelson

En 2023, Clyso a collaboré avec une entreprise avant-gardiste pour faire la transition de leur cluster Ceph d'un stockage basé sur des disques durs (HDD) vers un déploiement NVMe de 10 Pétaoctets (PB). Cette affiche décrit le parcours pour atteindre un débit record de 1 To/s avec un cluster Ceph, détaillant les défis rencontrés et les solutions mises en œuvre. Le projet a démontré le potentiel de Ceph à offrir des performances extrêmes sur une infrastructure matérielle et réseau bien conçue.

05. Défis et Correctifs



- Correctif 01.**

Problème: Latence due aux transitions d'état C du CPU (c-state).

Solution: Désactivation des c-states dans le BIOS, entraînant un gain de performance substantiel de 10 à 20 %.
- Correctif 02.**

Problème: Contention du noyau sur le verrouillage tournant (spin lock) de l'IOMMU.

Solution: Désactivation de l'IOMMU dans le noyau, entraînant une amélioration significative des performances.
- Correctif 03.**

Problème: Compaction lente dans RocksDB en raison de drapeaux de compilation incorrects.

Solution: Recompilation avec les drapeaux corrects, entraînant une compaction 3 fois plus rapide et des performances d'écriture aléatoire 4 Ko doublées.

06. Résultats



Performances des OSDs (Débit FIO 4 Mo)

OSDs (Nœuds)	Performance de Lecture Séquentielle 4 Mo (Go/s)	Performance d'Écriture Séquentielle 4 Mo (Go/s)
30 OSDs (3 Nœuds)	63	15
100 OSDs (10 Nœuds)	214	46
320 OSDs (32 Nœuds)	635	133
630 OSDs (63 Nœuds)	1025	270

07. Conclusion



Ce projet a prouvé la capacité de Ceph à atteindre un débit de 1 To/s sur un cluster NVMe optimisé. Les principaux enseignements portent sur l'influence cruciale du matériel, des réglages du noyau (kernel) et des ajustements logiciels sur les performances globales. Les prochaines étapes viseront à résoudre le phénomène de latence des PGs (laggy PGs) et à explorer de nouvelles optimisations du multithreading des OSD pour accroître davantage les débits.

08. Remerciements



Un grand merci à notre client pour sa collaboration et son soutien tout au long de ce projet. Notre gratitude s'adresse également à IBM/Red Hat et Samsung pour avoir fourni du matériel pour les tests de comparaison, ainsi qu'à tous les contributeurs de Ceph pour leurs efforts continus visant à améliorer la plateforme. Un grand merci sincère à Anthony D'Atri et Lee-Ann Pullar pour leur assistance en matière de relecture (copyediting).

Correctifs Implémentés : Mise à l'Échelle des OSD - Débit FIO 4 Mo (Meilleurs Résultats)

