Maintenance d'un cluster Ceph

Gérer les pannes

Etat du cluster

```
[ceph: root@ceph1 /]# ceph -s
  cluster:
    id:     bbb39ea2-ef21-11ec-b071-fa163e6de05c
    health: HEALTH_OK

  services:
    mon: 4 daemons, quorum ceph1,ceph2,ceph4,ceph3 (age 15h)
    mgr: ceph2.xesuzz(active, since 3d), standbys: ceph1.glefiw
    mds: 1/1 daemons up, 1 standby
    osd: 8 osds: 8 up (since 14h), 8 in (since 15h)
    rgw: 2 daemons active (2 hosts, 1 zones)

  data:
    volumes: 1/1 healthy
    pools:   13 pools, 321 pgs
    objects: 1.82k objects, 4.6 GiB
    usage:   15 GiB used, 305 GiB / 320 GiB avail
    pgs:     321 active+clean
```

En cas de problème, utiliser la commande ceph health detail pour plus d'information

```
[ceph: root@ceph1 /]# ceph health detail
HEALTH_WARN 2 osds down; 1 host (2 osds) down; Degraded data redundancy: 1296/5601 objects
degraded (23.139%), 135 pgs degraded
[WRN] OSD_DOWN: 2 osds down
    osd.0 (root=default,host=ceph1) is down
    osd.4 (root=default,host=ceph1) is down
[WRN] OSD_HOST_DOWN: 1 host (2 osds) down
    host ceph1 (root=default) (2 osds) is down
[WRN] PG_DEGRADED: Degraded data redundancy: 1296/5601 objects degraded (23.139%), 135 pgs
degraded
```

Monitorer l'activité de gestion du cluster :

```
ceph -W cephadm
```

Autres commandes pour l'état du cluster

```
ceph pg stat
ceph df
ceph df detail
```

Gestion des services

Les démons ceph sont gérés par systemd et les logs sont gérés par journald.

Un serveur physique peut être dans plusieurs clusters Ceph, et pour différencier les services de chacun des clusters ceph, chaque service est précédé de l'id du cluster

```
systemctl status ceph-bbb39ea2-ef21-11ec-b071-fa163e6de05c@osd.0.service
journalctl -u ceph-bbb39ea2-ef21-11ec-b071-fa163e6de05c@osd.0.service
```

Pour démarrer l'ensemble des services ceph sur un hôte, utiliser ceph.target

```
[root@ceph1 ~]# systemctl status ceph.target
ceph.target - All Ceph clusters and services
   Loaded: loaded (/etc/systemd/system/ceph.target; enabled; vendor preset: disabled)
   Active: active since Sat 2022-06-18 16:14:51 UTC; 2 weeks 5 days ago
```

Configuration des Logs

Traditionnellement, les démons Ceph enregistrent les leurs logs dans /var/log/ceph. Avec cephadm les démons Ceph enregistrent dans stderr et les journaux sont capturés par l'environnement d'exécution du conteneur. Pour la plupart des systèmes, par défaut, ces journaux sont envoyés à journald et accessibles via journalctl.

```
journalctl -u ceph-bbb39ea2-ef21-11ec-b071-fa163e6de05c@osd.0.service
```

Activer les fichiers log

```
ceph config set global log_to_file true
ceph config set global mon_cluster_log_to_file true
```

Les logs sont ensuite disponibles dans le répertoire /var/log/ceph/

```
ceph.log
ceph-mon.hostmane.log
ceph-mgr.hostname.log
ceph-mds.cephfs.hostname.log
ceph-ods.xxx.log
ceph-volume.log
```

Configuration de niveau de logs

Vous pouvez changer le niveau de logs des sous-systèmes Ceph à tout moment afin de faciliter le dépannage de tout problème qui pourrait survenir.

```
ceph tell TYPE.ID injectargs --debug-SUBSYSTEM VALUE [--NAME VALUE]
```

TYPE : le type de daemon ceph (osd, mon, ou mds)
ID : l'ID spécifique du deamon ceph. Vous pouvez utiliser * pour tous les daemons d'un type particulier
SUBSYSTEM : le sous-système du daemon
VALUE : un nombre de 0 à 20, où 0 est arrêt, 1 est concis et 20 est verbeux.
Exemple :

```
# ceph tell osd.0 injectargs --debug-osd 0/5
```

Voir la configuration

```
ceph daemon osd.0 config show
```

Problème avec les OSD

```
HEALTH_WARN 1 nearfull osds
osd.2 is near full at 85%
```

ou

```
HEALTH_ERR 1 full osds
osd.3 is full at 95%
```

Ceph renvoie le message nearfull osds lorsque le cluster atteint la capacité définie par le paramètre mon osd nearfull ratio. Par défaut, ce paramètre vaut 0,85, ce qui signifie 85 % de la capacité du cluster. Pour éviter d'arriver à une situation de blocage, il faut prévoir l'ajout de nouveaux serveurs avec du stockage supplémentaire.

Dans le cas d'une mauvaise répartition des données, il se peut qu'un OSD n'ai plus de place de libre.

Dans ce cas reportez-vous aux articles ci-dessous :

Troubleshoot Ceph OSDs Reporting Full

https://support.hpe.com/hpesc/public/docDisplay?docId=a00117940en_us&docLocale=en_US&page=Troubleshoot_Ceph_OSDs_Reporting_Full.html

Data distribution not equal across OSDs

https://www.suse.com/support/kb/doc/?id=000018889

Voir l'utilisation de « ceph osd reweight-by-utilization »

Scrub et deepScrub

En plus de réaliser plusieurs copies d'objets, Ceph assure l'intégrité des données en nettoyant les groupes de placement. Le nettoyage que réalise Ceph est analogue à l'exécution de fsck sur la couche de stockage d'objets. Pour chaque groupe de placement, Ceph génère un catalogue de tous les objets et compare chaque objet principal et ses répliques pour s'assurer qu'aucun objet n'est manquant ou discordant. Le nettoyage léger est réalisé quotidiennement et vérifie la taille et les attributs de l'objet, tandis que le nettoyage approfondi, hebdomadaire, lit les données et utilise les sommes de contrôle pour garantir l'intégrité de celles-ci.

Le nettoyage est essentiel au maintien de l'intégrité des données. Vous pouvez ajuster les paramètres suivants pour augmenter ou réduire la fréquence des opérations de nettoyage :

osd_max_scrubs : nombre maximum d'opérations de nettoyage simultanées pour un Ceph OSD. La valeur par défaut est 1.

osd_scrub_begin_hour, osd_scrub_end_hour : Heures du jour (0 à 24) qui définissent une fenêtre temporelle pendant laquelle le nettoyage peut avoir lieu. Par défaut, elle commence à 0 et se termine à 24.

```
#afficher les valeurs
ceph: root@ceph1 /]# ceph config show-with-defaults osd.0 |grep scrub
```

Exemple d'une détection et d'une correction pour un « scrub »

```
# ceph -s
  cluster:
    id:     xxxx
    health: HEALTH_ERR
            1 scrub errors
            Possible data damage: 1 pg inconsistent
  data:
    pools:   13 pools, 289 pgs
    objects: 39.28k objects, 152 GiB
    usage:   456 GiB used, 8.8 TiB / 9.3 TiB avail
    pgs:     288 active+clean
             1   active+clean+inconsistent

# ceph health detail
HEALTH_ERR 1 scrub errors; Possible data damage: 1 pg inconsistent
OSD_SCRUB_ERRORS 1 scrub errors
PG_DAMAGED Possible data damage: 1 pg inconsistent
    pg 9.2 is active+clean+inconsistent, acting [14,4,1]
# ceph pg repair 9.2
instructing pg 9.2 on osd.14 to repair
# ceph -s
  cluster:
    id:     xxxx
    health: HEALTH_OK
  data:
    pools:   13 pools, 289 pgs
    objects: 39.28k objects, 152 GiB
    usage:   456 GiB used, 8.8 TiB / 9.3 TiB avail
    pgs:     289 active+clean
```

Remarque : Le cluster CEPH est de nouveau dans un état de santé optimal. Si votre cluster indique fréquemment des PGs en erreur, il est impératif de vérifier vos disques de stockage.

Recherche des OSD lents

Un disque peut avoir des lenteurs. Ce qui va impacter les performances lors de l'écriture des données, car Ceph vérifie que les opérations d'écriture soient complètes sur tous les disques associés à un PG.

Il est possible d'examiner chaque OSD afin de trouver ceux qui ralentissent le processus d'écriture avec la commande :

```
[ceph: root@ceph1 /]# ceph tell osd.* bench
osd.0: {
    "bytes_written": 1073741824,
    "blocksize": 4194304,
    "elapsed_sec": 1.833926951,
    "bytes_per_sec": 585487782.60470641,
    "iops": 139.59116520993862
}
osd.1: {
```

Vous devez ensuite comparer la valeur de IOPS pour identifier les OSD les plus lents.

Redémarrer un nœud

Lors des opérations de maintenance, il est parfois nécessaire de redémarrer un nœud du cluster. Cela est supporté par défaut, à condition d'avoir un cluster dans un état sain.

Après avoir vérifié l'état du cluster, on commence par désactiver temporairement le balancing du cluster. L'objectif est d'empêcher le cluster de répliquer les objets non disponibles sur d'autres serveurs. Sur un des MON, faire :

```
# ceph osd set noout
# ceph osd set norebalance
```
Sur le nœud à éteindre :

```
# reboot
```
Quand la machine est de nouveau disponible, on vérifie l'état du cluster. Il faut que les PGs soient active+clean. Sur un des nœuds :

```
# ceph -s
```
Puis réactiver le balancing du cluster.

```
# ceph osd unset noout
# ceph osd unset norebalance
```
Pour finir, on s'assure que tout va bien :

```
# ceph status
```

Remplacement d'un OSD

Maintenant qu'un OSD est HS que faire ? D'abord vérifier le matériel avec smartools. Si le matériel est en cause, il faut le changer et créer un nouvel OSD ;). Il est conseillé de ne pas chercher à réparer un OSD, mais plutôt laisser Ceph répliquer les données sur un nouveau disque, ce qui se fait assez facilement.

Méthode avec cephadm et ceph orch :

```
# identification du disque
[ceph: root@ceph1 /]# ceph device ls |grep osd.5
7811529f-a85f-44c0-8  ceph3:vdc  osd.5
# remarque : vdc sur ceph3
# suppression avec draining
[ceph: root@ceph1 /]# ceph orch osd rm 5
Scheduled OSD(s) for removal
[ceph: root@ceph1 /]# ceph orch osd rm status
OSD  HOST   STATE     PGS  REPLACE  FORCE  ZAP    DRAIN STARTED AT
5    ceph3  draining   27  False    False  False  2022-08-25 16:36:26.900875
[ceph: root@ceph1 /]#  ceph health
HEALTH_WARN Reduced data availability: 6 pgs inactive; Degraded data redundancy: 601/262269
objects degraded (0.229%), 74 pgs degraded, 36 pgs undersized
[ceph: root@ceph2 /]# ceph orch osd rm status
OSD  HOST   STATE                    PGS  REPLACE  FORCE  ZAP    DRAIN STARTED AT
5    ceph3  done, waiting for purge    0  False    False  False  2022-08-25 17:36:04.821739
[ceph: root@ceph2 /]# ceph orch daemon stop osd.5
Scheduled to stop osd.5 on host 'ceph3'
[ceph: root@ceph2 /]# ceph osd tree
#plus de trace de l'osd.5


[ceph: root@ceph2 /]# ceph orch device ls ceph3
# il reste le lvm
[ceph: root@ceph2 /]# ceph orch device zap ceph3 /dev/vdc --force
zap successful for /dev/vdc on ceph3
[ceph: root@ceph2 /]# ceph orch device ls ceph3
# au bout d'un temps un nouveau lvm apparait
[ceph: root@ceph2 /]# ceph osd tree
#retour de ceph osd.5 creation automatique par ceph orch
[ceph: root@ceph2 /]# ceph health
HEALTH_WARN Reduced data availability: 6 pgs inactive; Degraded data redundancy: 608/260175
objects degraded (0.234%), 67 pgs degraded, 34 pgs undersized

# remarque Removing an OSD using ceph orch requires some additional cleanup
# https://www.suse.com/support/kb/doc/?id=000020642
# cephadm ceph-volume lvm zap --destroy --osd-id <id>
```

Méthode sans ceph orch :

```
[ftceph@deploy ceph]$ ceph device ls
DEVICE               HOST:DEV            DAEMONS LIFE EXPECTANCY
1743e4d3-ae99-45f0-9 ceph2.novalocal:vdc osd.4
3015d4c8-0d84-4826-b ceph3.novalocal:vdc osd.5
709417f2-6faa-46c3-a ceph2.novalocal:vdb osd.1
7b4325d7-5040-4426-b ceph1.novalocal:vdb osd.0
ae6894f6-aa21-419b-9 ceph3.novalocal:vdb osd.2
d693d370-c1dc-476b-a ceph1.novalocal:vdc osd.3
# on retrouve le disque vbd sur ceph3

[ftceph@deploy ceph]$ ceph osd purge 2 --yes-i-really-mean-it
purged osd.2
# l'osd 2 est supprimé du cluster

[ftceph@deploy ceph]$ ceph osd tree
ID  CLASS WEIGHT  TYPE NAME      STATUS REWEIGHT PRI-AFF
-13       0.19547 root default
-16       0.07819     host ceph1
  0   hdd 0.03909         osd.0      up  1.00000 1.00000
  3   ssd 0.03909         osd.3      up  1.00000 1.00000
-15       0.07819     host ceph2
  1   hdd 0.03909         osd.1      up  1.00000 1.00000
  4   ssd 0.03909         osd.4      up  1.00000 1.00000
-14       0.03909     host ceph3
  5   ssd 0.03909         osd.5      up  1.00000 1.00000
#plus de référence osd.2;)

#faire le ménage dans LVM
[ftceph@deploy ceph]$ ssh ceph3 sudo ceph-volume lvm zap --destroy --osd-id 2
[...]
 stdout: Volume group "ceph-4ac3dfbf-3a32-4d21-912e-c760eea0b781" successfully removed
--> Zapping successful for OSD: 2

# changer physiquement le disque
# préparer le nouveau disque
[ftceph@deploy ceph]$ ceph-deploy disk zap ceph3 /dev/vdb
[...]
[ceph3][WARNIN] --> Zapping successful for: <Raw Device: /dev/vdb>

# créer un nouveau OSD sur le disque
[ftceph@deploy ceph]$ ceph-deploy osd create --data /dev/vdb ceph3
[...]
[ceph3][WARNIN] --> ceph-volume lvm activate successful for osd ID: 2
[ceph3][WARNIN] --> ceph-volume lvm create successful for: /dev/vdb

# vérifier que l'OSD est présent
[ftceph@deploy ceph]$ ceph osd tree
ID  CLASS WEIGHT  TYPE NAME      STATUS REWEIGHT PRI-AFF
-13       0.23456 root default
-16       0.07819     host ceph1
  0   hdd 0.03909         osd.0      up  1.00000 1.00000
  3   ssd 0.03909         osd.3      up  1.00000 1.00000
-15       0.07819     host ceph2
  1   hdd 0.03909         osd.1      up  1.00000 1.00000
  4   ssd 0.03909         osd.4      up  1.00000 1.00000
-14       0.07819     host ceph3
  2   hdd 0.03909         osd.2      up  1.00000 1.00000
  5   ssd 0.03909         osd.5      up  1.00000 1.00000

# suivre les opérations de reconstruction
[ftceph@deploy ceph]$ ceph -w
```

```
# ceph indique un problème de crash avec l'osd.2
[ftceph@deploy ceph]$ ceph health detail
HEALTH_WARN 1 daemons have recently crashed
RECENT_CRASH 1 daemons have recently crashed
    osd.2 crashed on host ceph3.novalocal at 2020-06-19 07:09:15.227175Z

# [ftceph@deploy ceph]$ ceph crash archive-all
[ftceph@deploy ceph]$ ceph health detail
HEALTH_OK
#retour à l'état sain sans perte de données, Youpi !
```

Gestion centralisée de la configuration

Depuis la version Mimic de Ceph, il est possible de gérer de manière centralisée la configuration du cluster qui était jusqu'à présent dans les fichiers ceph.conf

Historiquement, les admins devaient modifier manuellement les fichiers ceph.conf et les distribuer sur les nœuds, puis s'assurer que les services étaient redémarrés à tour de rôle pour la prise en compte des changements.

Cette nouvelle fonctionnalité a été conçue afin d'éviter l'utilisation d'outils externes pour gérer les fichiers de configurations ceph.conf

Fonctionnement

Les moniteurs gèrent conjointement une base de données de configuration. La base de données a la même structure sémantique qu'un fichier ceph.conf :

- Il y a des noms d'options (par exemple, osd_scrub_load_threshold) et des valeurs.
- Un paramètre peut être associé à un groupe "global", et à un groupe de service de même nature (par exemple, "osd" ou "mds"), ou à un service spécifique (par exemple, "osd.123").

Vous pouvez toujours mettre des paramètres dans ceph.conf. L'ordre de priorité utilisé par Ceph pour définir les options est le suivant:

1. Valeurs par défaut à la compilation
2. Base de données de configuration de cluster (la nouveauté!)
3. Fichier ceph.conf local
4. Remplacement à chaud (via «ceph daemon <daemon> config set…» ou «ceph tell <daemon> injectargs…»)

Commande en ligne

ceph config -h affiche les options disponibles dont voici les plus importantes :

```
config dump                             Show all configuration option(s)
config get <who> {<key>}                Show configuration option(s) for an entity
config help <key>                       Describe a configuration option
config set <who> <name> <value>         Set a configuration option for one or more entities
config show <who> {<key>}               Show running configuration
config show-with-defaults <who>         Show running configuration (including compiled-in
defaults)
```

Voir la configuration du cluster :

```
[ftceph@deploy ceph]$ ceph config dump
WHO   MASK LEVEL    OPTION                          VALUE                   RO
  mgr      advanced mgr/dashboard/RGW_API_ACCESS_KEY RGWV4GK9PO3H6C5V0YKZ   *
  mgr      advanced mgr/dashboard/RGW_API_SECRET_KEY Z0UvbIp...             *
  mgr      advanced mgr/dashboard/password           $2b$12$...             *
  mgr      advanced mgr/dashboard/ssl                false                  *
```

```
  mgr      advanced mgr/dashboard/username           xstradm                *
```

Afficher l'aide concernant un paramètre :

```
[ftceph@deploy ceph]$ ceph config help debug_osd
debug_osd - Debug level for osd
  (str, advanced)
  Default: 1/5
  Can update at runtime: true

The value takes the form 'N' or 'N/M' where N and M are values between 0 and 99.  N is the
debug level to log (all values below this are included), and M is the level to gather and
buffer in memory.  In the event of a crash, the most recent items <= M are dumped to the
log file.
```

Modifier et consulter les modifications pour un service :

```
[ftceph@deploy ceph]$ ceph config set osd.0 debug_osd 10
[ftceph@deploy ceph]$ ceph config get osd.0
WHO   MASK LEVEL    OPTION    VALUE RO
osd.0      advanced debug_osd 10/10
```

Supprimer une configuration :

```
[ftceph@deploy ceph]$ ceph config rm osd.0 debug_osd
[ftceph@deploy ceph]$ ceph config get osd.0
WHO MASK LEVEL OPTION VALUE RO
```

Vérification du schéma

L'avantage de l'utilisation d'un schéma, est que les valeurs des paramètres sont vérifiées avant d'être validées. C'est utile pour savoir si l'option est toujours supportée. On peut vérifier les options. Par exemple :

```
[ftceph@deploy ceph]$ ceph config set osd.0 bluestore_compression_mode 1
Error EINVAL: error parsing value: '1' is not one of the permitted values: none, passive,
aggressive, force
[ftceph@deploy ceph]$ ceph config help bluestore_compression_mode
bluestore_compression_mode - Default policy for using compression when pool does not
specify
  (str, advanced)
  Default: none
  Possible values:  none passive aggressive force
  Can update at runtime: true

'none' means never use compression.  'passive' means use compression when clients hint that
data is compressible.  'aggressive' means use compression unless clients hint that data is
not compressible.  This option is used when the per-pool property for the compression mode
is not present.
```

Historique des changements de configuration

L'utilisation de fichier de configuration au format txt, permet, via des outils externes, d'avoir un historique des versions de configuration. Ceph config fournit un comportement similaire. Chaque changement de configuration est enregistré et facilement visualisable:

```
ceph: root@ceph1 /]# ceph config log
--- 8 --- 2020-06-19 13:03:38.339954 ---
- osd.0/debug_osd = 10/10
--- 7 --- 2020-06-19 13:00:06.917395 ---
+ osd.0/debug_osd = 10/10
--- 6 --- 2020-06-17 06:26:39.506031 ---
+ mgr/mgr/dashboard/RGW_API_SECRET_KEY = Z0UvbIpb9gxwztKDKRgUg8BMiyUguMBBGLsPBV55
--- 5 --- 2020-06-17 06:26:15.189409 ---
+ mgr/mgr/dashboard/RGW_API_ACCESS_KEY = RGWV4GK9PO3H6C5V0YKZ
--- 4 --- 2020-06-17 05:42:20.186943 ---
+ mgr/mgr/dashboard/password = $2b$12$2fM496...
--- 3 --- 2020-06-17 05:42:19.857389 ---
+ mgr/mgr/dashboard/username = xstradm
```

```
--- 2 --- 2020-06-17 05:37:12.284529 ---
+ mgr/mgr/dashboard/ssl = false
--- 1 --- 2020-06-15 19:59:08.439943 ---
```

Revenir à une configuration

Il est possible de revenir à une configuration antérieure, en annulant les dernières modifications. Pour cela, identifiez le numéro de configuration qui précède chaque changement, et utilisez la commande « ceph config reset » suivi du numéro pour appliquer l'ancienne configuration.

```
ceph: root@ceph1 /]# ceph config reset 7
ceph: root@ceph1 /]# ceph config log
--- 9 --- 2020-06-19 13:27:42.755428 --- reset to 7 ---
+ osd.0/debug_osd = 10/10
--- 8 --- 2020-06-19 13:03:38.339954 ---
- osd.0/debug_osd = 10/10
--- 7 --- 2020-06-19 13:00:06.917395 ---
+ osd.0/debug_osd = 10/10
--- 6 --- 2020-06-17 06:26:39.506031 ---
+ mgr/mgr/dashboard/RGW_API_SECRET_KEY = Z0UvbIpb9gxwztKDKRgUg8BMiyUguMBBGLsPBV55
--- 5 --- 2020-06-17 06:26:15.189409 ---
[...]
```

Remarque

La centralisation de la configuration permet de modifier la configuration soit avec le Dashboard soit avec une commande en ligne en fonction de la préférence de chacun, ce qui devrait simplifier la vie des admins Ceph.

Configuration réseau

NTP

Il est important que tous les nodes du cluster Ceph soient bien à l'heure. Ceph affiche un warning s'il constate un décalage de temps supérieur à 0.5 seconde

```
[root@mon ~]# systemctl status chronyd
[root@mon ~]# chronyc sources
[root@mon ~]# chronyc sourcestats
[root@mon ~]# chronyc tracking
```

S'il y a un décalage au niveau de l'heure, il est possible de forcer la mise à l'heure

```
chronyc makestep
```

Ethtool

Vérifier le nombre d'erreurs de transmission

```
ethtool -S ens3 | grep errors
```

Vérifier la vitesse de connexion

```
ethtool INTERFACE
```

Iperf3

Permet de tester les connexions réseau entre les nodes du cluster

Lancement du serveur

```
[root@host01 ~]# iperf3 -s
-----------------------------------------------------------
Server listening on 5201
```

```
-----------------------------------------------------------
```

Lancement du client

```
[root@host02 ~]# iperf3 -c host01
Connecting to host mon, port 5201
[  4] local xx.x.xxx.xx port 52270 connected to xx.x.xxx.xx port 5201
[ ID] Interval           Transfer     Bandwidth       Retr  Cwnd
[  4]   0.00-1.00   sec   114 MBytes   954 Mbits/sec    0    409 KBytes
[  4]   1.00-2.00   sec   113 MBytes   945 Mbits/sec    0    409 KBytes
[  4]   2.00-3.00   sec   112 MBytes   943 Mbits/sec    0    454 KBytes
[  4]   3.00-4.00   sec   112 MBytes   941 Mbits/sec    0    471 KBytes
[  4]   4.00-5.00   sec   112 MBytes   940 Mbits/sec    0    471 KBytes
[  4]   5.00-6.00   sec   113 MBytes   945 Mbits/sec    0    471 KBytes
[  4]   6.00-7.00   sec   112 MBytes   937 Mbits/sec    0    488 KBytes
[  4]   7.00-8.00   sec   113 MBytes   947 Mbits/sec    0    520 KBytes
[  4]   8.00-9.00   sec   112 MBytes   939 Mbits/sec    0    520 KBytes
[  4]   9.00-10.00  sec   112 MBytes   939 Mbits/sec    0    520 KBytes
- - - - - - - - - - - - - - - - - - - - - - - - -
[ ID] Interval           Transfer     Bandwidth       Retr
[  4]   0.00-10.00  sec  1.10 GBytes   943 Mbits/sec    0             sender
[  4]   0.00-10.00  sec  1.10 GBytes   941 Mbits/sec                  receiver

iperf Done.
```

Cette sortie montre une bande passante réseau de 1,1 Gbits/seconde entre les nœuds host01 et host02, ainsi qu'aucune retransmission (Retr) pendant le test indiquant qu'il n'y a pas de perte de paquet.

Les performances

Evaluer les performances avec Rados bench

Ceph inclut la commande « rados bench » pour effectuer des tests de performance. La commande exécute un test d'écriture et deux types de tests de lecture. Il est important d'utiliser l'option --no-cleanup pour tester les performances en lecture et en écriture. Par défaut, la commande rados bench supprime les objets qu'elle a écrits dans le pool de stockage. Le fait de laisser ces objets permet aux deux tests de lecture de mesurer les performances de lecture séquentielle et aléatoire.

Depuis un client :

```
rados bench -p prbd 10 write --no-cleanup
rados bench -p prbd 10 seq
rados bench -p prbd 10 rand
rados -p prbd cleanup
```

Documentation : https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/1.3/html/administration_guide/benchmarking_performance

Nombre de noeuds

Dans une architecture traditionnelle de stockage, le contrôleur ou la carte RAID permet d'accéder à l'ensemble des disques. Lorsque le contrôleur disque atteint une utilisation de 100 %, il crée un goulot d'étranglement. Dans ce cas, l'ajout de disques ne permet pas d'améliorer les performances, mais uniquement d'augmenter la capacité.

Une architecture scale-out élimine ce problème en fournissant de la capacité et des performances en ajoutant de nouveaux nœuds à un cluster de stockage. Lorsque des nœuds sont ajoutés, les performances et la capacité augmentent ensemble. Il est recommandé de ne pas dépasser 12 disques par serveur, mais d'augmenter le nombre de nœuds et le nombre de disques pour avoir plus d'axes de rotation et distribuer les requêtes sur plusieurs serveurs.

Exemple : gain de performance entre un cluster à 3 ou 5 nœuds

Workload	IOPS	Average Latency	Tail Latency
Random Read	55% Higher	29% Lower	30% Lower
Random Read Write	95% Higher	46% Lower	44% Lower
Random Write	77% Higher	40% Lower	38% Lower

https://ceph.io/en/news/blog/2019/part-3-rhcs-bluestore-performance-scalability-3-vs-5-nodes/

Nombre de CPU par OSD

Avec un stockage NVME, augmenter le nombre de CPUs par OSD permet d'augmenter les performances. Augmenter au-delà de 6 CPUs apporte plus de performance mais pourrait ne pas justifier le coût impliqué du prix de processeurs multi-cœurs

Avec l'utilisation de disques capacitifs l'augmentation de CPU de 2 à 4 par OSD apporte également un gain de performance.

Nombre de CPUs par OSD NVME

	IOPS Increase from 4 to 6 Cores	IOPS Increase from 6 to 8 Cores
Randread	+ 31.96%	+ 0.51%
Randrw	+ 48.71%	+ 17.15%
Randwrite	+ 51.66%	+ 19.60%

https://ceph.io/en/news/blog/2019/part-4-rhcs-3-2-bluestore-advanced-performance-investigation/

Bluestore 8GB cache vs Bluestore 4GB cache

La taille du cache bluestore peut avoir un impact important sur les performances. Si la donnée n'est pas en cache, elle sera lue à partir du disque, puis mise dans un cache Clé/Valeurs via la RockDB.

Avec une augmentation de la taille du cache BlueStore à 8G, la charge de travail de lecture-écriture aléatoire (70/30) pouvait être augmentée jusqu'à 30% d'IOPS en plus et une latence de queue réduite de 50%.

Workload	IOPS	Avg Lat	P95% Lat	P99% Lat
Rand Read	▲ 14.43%	▼ -24.57%	▼ -25.43%	▼ -61.76%
Rand RW(70R/30W)	▲ 30.52%	▼ -32.62%	▼ -52.12%	▼ -11.60%
Random Write	▲ 15.40%	▼ -19.10%	▼ -24.31%	▼ -28.68%

https://ceph.io/en/news/blog/2019/part-4-rhcs-3-2-bluestore-advanced-performance-investigation/

```
bluestore_cache_autotune : True
osd_memory_target_autotune : True
autotune_memory_target_ratio : 0.7  #70% de la ram système
bluestore_cache_size: 0
bluestore_cache_size_hdd: 1073741824
bluestore_cache_size_ssd: 3221225472
bluestore_cache_kv_ratio : .4
bluestore_cache_kv_max : 512 * 1024*1024 (512 MB)
```

Cache SSD pour le WAL/RocksDB

Mesure de la différence entre une configuration BlueStore (rocksdb et WAL) co-localisé sur le même disque que les datas par rapport l'utilisation d'un disque SSD pour la ROCKDB et le WAL

L'utilisation d'un disque SSD permet d'améliorer légèrement les IOPS mais surtout la latence lors de l'écriture aléatoire avec une forte charge ce qui peut être utile pour les applications nécessitant une faible latence comme les bases de données.

Workload	IOPS	Avg Lat	P95% Lat	P99% Lat
Rand Read	▲ 4.06%	▼ -1.05%	▼ -2.46%	▼ -2.29%
Rand RW(70R/30W)	▲ 9.55%	▼ -3.83%	▼ -4.57%	▼ -4.01%
Random Write	▲ 7.23%	▼ -4.40%	▼ -13.08%	▼ -13.82%

https://ceph.io/en/news/blog/2019/part-4-rhcs-3-2-bluestore-advanced-performance-investigation/

Documentation

Product Documentation for Red Hat Ceph Storage 5

https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/5

La documentation de Ceph fournit plusieurs procédures pour détecter les problèmes.

https://docs.ceph.com/docs/master/rados/operations/