> Nhanhoa Cloud Team
>
> Tunning Storage
>
> Date: 04/01/2018

MỤC LỤC

1. Tunning sysctl.conf

> No require restart Server, Ceph Service

- Đối tượng áp dụng: CephNodeSSD01->06
- Mục đích: Tối ưu TCP, UDP, memory, tăng số lượng files kernel có thể mở và xử lý
- Revert lại cấu hình mặc định nếu có vấn đề

1.1 Cấu hình điều chỉnh tăng các tham số limit của Kernel

```
cat << EOF >> /etc/sysctl.conf
# Increase Linux autotuning TCP buffer limits
# Set max to 16MB for 1GE and 32M (33554432) or 54M (56623104) for 10GE
# Use 128M buffers
kernel.pid_max=4194303
fs.file-max=26234859
```

```
# VM Settings
vm.swappiness=1
vm.vfs_cache_pressure=1

net.ipv4.tcp_rmem=67108864 134217728 56623104
net.ipv4.tcp_wmem=67108864 134217728 56623104
net.ipv4.tcp_low_latency=1
net.ipv4.tcp_adv_win_scale=1

net.core.rmem_max= 56623104
net.core.wmem_max= 56623104
net.core.rmem_default= 56623104
net.core.wmem_default= 56623104
net.core.netdev_budget=1200
net.core.optmem_max=134217728

# Max packet backlog
net.core.somaxconn=32768

# Increase the length of the processor input queue
net.core.netdev_max_backlog=250000
net.ipv4.tcp_max_syn_backlog=30000
net.ipv4.tcp_max_tw_buckets=2000000
net.ipv4.tcp_tw_reuse=1
net.ipv4.tcp_tw_recycle=1
net.ipv4.tcp_fin_timeout=5

# Disable TCP slow start on idle connections
net.ipv4.tcp_slow_start_after_idle=0
net.ipv4.conf.all.send_redirects=0
net.ipv4.conf.all.accept_redirects=0
net.ipv4.conf.all.accept_source_route=0

# If your servers talk UDP, also up these limits
net.ipv4.udp_rmem_min=8192
net.ipv4.udp_wmem_min=8192

# Disable source routing and redirects
net.ipv4.conf.all.send_redirects=0
net.ipv4.conf.all.accept_redirects=0
net.ipv4.conf.all.accept_source_route=0
kernel.perf_event_max_sample_rate = 80000
EOF

sysctl -p
```

1.2 Revert lại cấu hình

```
Bỏ các cấu hình phía trên và chạy lệnh `sysctl -p`
```

2. Tunning Ulimit

> No require restart Server, Services

- Đối tượng áp dụng: CephNodeSSD01->06
- Mục đích: Tăng số lượng files kernel có thể mở và xử lý
- Cài đặt
- Revert lại cấu hình mặc định nếu có vấn đề

2.1 Setup

```
ulimit -n 10240
```

Add vào khởi động cùng hệ thống `/etc/rc.local` để tự add cấu hình lúc reboot server

```
...
ulimit -n 10240
exit 0
```

Kiểm tra

```
ulimit -a
```

2.2 Revert cấu hình mặc định

```
ulimit -n 1024
```

Bỏ cấu hình trong `/etc/rc.local`

3. Tunning Hugepages (dd write-sequence)

> No require restart Server, Services
>
> Cân nhắc không cần tunning tham số này

- Đối tượng áp dụng: CephNodeSSD01->06
- Mục đích: Cung cấp giải pháp thay thế cho các gói tin có kích thước 4k (16k) nhằm tăng tốc độ write Seq - Kiểm tra với câu lệnh `dd`
- Cài đặt và cấu hình
- Revert lại cấu hình mặc định nếu có vấn đề

3.1 Tăng cấu hình tham số hugepage

```
cat << EOF >> /etc/sysctl.conf
vm.nr_hugepages = 10
EOF
```

Kiểm tra

```
cat /proc/meminfo | grep hugepages
sysctl a | grep nr_hugepages
```

Cách tính Hugepage

```
HugePages_Total - HugePages_Free = HugePages_Used
```

3.2 Revert lại cấu hinh mặc định

Bỏ cấu hình hugepages trong `/etc/sysctl.conf`

Cập nhật lại cấu hình `sysctl -p`

4. Tunning Block Device (I/O Block Device)

> No require restart Server, Services

- Đối tượng áp dụng: CephNodeSSD01->06
- Mục đích: Tunning Disk I/O cho các OSD của các node CephSSD
- Cấu hình tunning block device cho các CephNodeSSD
- Revert cấu hình mặc định nếu có vấn đề

4.1 Cấu hình

```
# Bỏ qua Block Device cài đặt OS cấu hình cho các Block Device OSD
for device in $(find /sys/block/* | grep -E '\/sd' ) ; do
if [ $device != "/sys/block/sda" ]; then
        echo 0 > $device/queue/add_random
        echo 0 > $device/queue/rotational
        echo 2 > $device/queue/rq_affinity
        echo 2048 > $device/queue/read_ahead_kb
        echo 512 > $device/queue/nr_requests
        echo "noop" > $device/queue/scheduler
fi
done
```

Bổ sung vào `/etc/rc.local` để tự động set nhận lúc reboot server

```
...
for device in $(find /sys/block/* | grep -E '\/sd' ) ; do
if [ $device != "/sys/block/sda" ]; then
        echo 0 > $device/queue/add_random
        echo 0 > $device/queue/rotational
        echo 2 > $device/queue/rq_affinity
        echo 2048 > $device/queue/read_ahead_kb
        echo 512 > $device/queue/nr_requests
        echo "noop" > $device/queue/scheduler
fi
done
exit 0
```

4.2 Revert cấu hình mặc định

Revert lại cấu hình mặc định

```
for device in $(find /sys/block/* | grep -E '\/sd' ) ; do
if [ $device != "/sys/block/sda" ]; then
        echo 0 > $device/queue/add_random
        echo 0 > $device/queue/rotational
        echo 1 > $device/queue/rq_affinity
        echo 128 > $device/queue/read_ahead_kb
        echo 128 > $device/queue/nr_requests
        echo "noop [deadline] cfq" > $device/queue/scheduler
fi
done
```

Bỏ cấu hình trong `/etc/rc.local`

5. Tunning Tunned (latency)

> No require restart Server, Services

- Đối tượng áp dụng: CephNodeSSD01->06, Compute 1->7
- Mục đích: Điều chỉnh server áp dụng các loại setup tunning theo profile
- Cài đặt và cấu hình profile `throughput-performance`
- Revert lại cấu hình mặc định nếu có vấn đề

5.1 Cài đặt

Các bước cấu hình:

```
yum -y install tuned
systemctl start tuned
systemctl enable tuned
```

Check

```
tuned-adm list
```

Set cho các node Ceph và Compute

```
tuned-adm profile throughput-performance
```

5.2 Revert lại cấu hình

Set cho các node Ceph và Compute

```
tuned-adm profile balanced
```

6. Tunning Jumbo Frame (MTU9000) Switch

- Đối tượng áp dụng: Sw Ceph-Com, Sw Ceph-Replicate, CephNodeSSD01->06, CephBKA01-02
- Mục đích: Enable Jumbo Frame MTU 9000 trên Switch Nexus 3000 (Ceph-Com) và Swith IBM G8124 (Ceph-Rep)
- Config MTU 900 trên các Interface Ceph và Compute
- Restart network trên Ceph và Compute
- Check packet trên Sw và Ceph, Compute đảm bảo các gói tin Jumbo Frame được đẩy qua
- Revert lại cấu hình mặc định nếu có vấn đề

6.1 Switch Nexus 3000 (Ceph-Com)

> No require restart Sw

Show toàn bộ qos

`show class-map type network-qos` Displays the type network qos class maps.

`show policy-map type network-qos` Displays the type network qos policy maps.

`show policy-map system type network-qos` Displays the active type network qos class maps.

Cấu hình trên Sw Nexus3000 cấu hình này sẽ ăn cho toàn bộ port trên SW trừ `Management port`

```
switch# configure terminal
switch(config)# policy-map type network-qos jumbo
switch(config-pmap-nq)# class type network-qos class-default
switch(config-pmap-c-nq)# mtu 9216
switch(config-pmap-c-nq)# exit
switch(config-pmap-nq)# exit
switch(config)# system qos
switch(config-sys-qos)# service-policy type network-qos jumbo
```

6.2 Switch IBM G8124 (Ceph-Replicate)

> No require restart Sw

```
Tự động nhận cấu hình MTU không cần cấu hình
Jumbo frame support is automatic
```

6.3 Cấu hình MTU cho các Interface Linux Server

> Require restart network on Node Ceph, Node Compute

Ceph 1-6, CephBKA01 (CentOS7)

```
# Interface ceph-com ceph-replicate
for interface in p4p1 p4p2 p6p1 p6p2 bond0 bond2
do
        file="/etc/sysconfig/network-scripts/ifcfg-$interface"
        if [ -f "$file" ]
                echo MTU="9000" >> $file
        fi
done
```

CephBKA02 (Ubuntu16.04) `/etc/network/interfaces` cấu hình cho mỗi interface bổ sung

```
...
MTU 9000
```

COMPUTE (CentOS7)

```
for interface in {interface ceph-com: physical and bond network
192.168.72.0/24}
do
        file="/etc/sysconfig/network-scripts/ifcfg-$interface"
        if [ -f "$file" ]
                echo MTU="9000" >> $file
```

```
        fi
done
```

6.4 Kiểm tra

1. Kiểm tra gói tin ping từ Server

```
ping -l 9000 x.x.x.x
```

2. Nexus 3000 Kiểm tra jumbo packet thay đổi (lưu ý phải đẩy traffic có MTU > 1500)

```
switch# sh queuing int e1/19
show interface e1/44 counters detail
```

Show toàn bộ qos

`show class-map type network-qos` Displays the type network qos class maps.

`show policy-map type network-qos` Displays the type network qos policy maps.

`show policy-map system type network-qos` Displays the active type network qos class maps.

6.5 Revert lại cấu hình MTU 1500 mặc định

1. Switch Nexus 300 (Ceph-Com)

Cấu hình qos jumbo về class default

```
switch# configure terminal
switch(config)# system qos
switch(config-sys-qos)# service-policy type network-qos class-default
```

2. Switch IBM G8124 (Ceph-Replicate)

```
Không cần cấu hình
```

3. Bỏ config MTU trên các Interface

- Bỏ các cấu hình MTU 9000 trong các interface
- Restart network

7. Tunning ceph.conf

> No require restart Server
>
> Require restart Ceph-mon Service, Cinder-* Service trên Compute

- Đối tượng áp dụng: CephNodeSSD01->06 , Controller
- Mục đích:
  - Disable ceph auth
  - Enable warning max, min osd
  - Append max pool, pgnum
  - Append bluestore_rocksdb_options, cached
- Revert lại cấu hình mặc định nếu có vấn đề

7.1 Cấu hình config hiện tại của hệ thống

Chỉnh sửa ceph.conf trong thư mục `/home/cephuser/ceph-deploy` trên node `Cephssd01`

> Lưu ý: Phần BlueStoreDB đối với các hệ thống đang chạy có thể gây ra crash khi điều chỉnh size + config, hệ thống

```
[global]
fsid = 700b2fe0-54e0-43b5-93d6-7f58a3f5639d
mon_initial_members = ceph1, ceph2, ceph3
mon_host = 10.10.13.67,10.10.13.68,10.10.13.69
public network = 10.10.13.0/24
cluster network = 10.10.14.0/24
osd objectstore = bluestore
osd pool default size =2
osd pool default min size = 1
auth client required = none
auth cluster required = none
auth service required = none
auth supported = none
cephx require signatures = false
cephx sign messages = false
mon_allow_pool_delete = false
mon_max_pg_per_osd = 800
mon pg warn max per osd = 800
ms_crc_header = false
ms_crc_data = false
ms type = async
perf = true
rocksdb_perf = true
debug_lockdep = 0/0
debug_context = 0/0
debug_crush = 0/0
debug_buffer = 0/0
debug_timer = 0/0
debug_filer = 0/0
debug_objecter = 0/0
debug_rados = 0/0
debug_rbd = 0/0
debug_ms = 0/0
```

```
debug_monc = 0/0
debug_tp = 0/0
debug_auth = 0/0
debug_finisher = 0/0
debug_heartbeatmap = 0/0
debug_perfcounter = 0/0
debug_asok = 0/0
debug_throttle = 0/0
debug_mon = 0/0
debug_paxos = 0/0
debug_rgw = 0/0

[mon]
mon_max_pool_pg_num = 166496
mon_osd_max_split_count = 10000

[client]
rbd_cache = false
rbd_cache_writethrough_until_flush = false

[osd]
bluestore_csum_type = none
bluestore_cache_kv_max = 200G
bluestore_cache_kv_ratio = 0.2
bluestore_cache_meta_ratio = 0.8
bluestore_cache_size_ssd = 18G
bluestore_extent_map_shard_min_size = 50
bluestore_extent_map_shard_max_size = 200
bluestore_extent_map_shard_target_size = 100
osd_min_pg_log_entries = 10
osd_max_pg_log_entries = 10
osd_pg_log_dups_tracked = 10
osd_pg_log_trim_min = 10
bluestore_rocksdb_options =
compression=kNoCompression,max_write_buffer_number=64,min_write_buffer_numb
er_to_merge=32,recycle_log_file_num=64,compaction_style=kCompactionStyleLev
el,write_buffer_size=4MB,target_file_size_base=4MB,max_background_compactio
ns=64,level0_file_num_compaction_trigger=64,level0_slowdown_writes_trigger=
128,level0_stop_writes_trigger=256,max_bytes_for_level_base=6GB,compaction_
threads=32,flusher_threads=8,compaction_readahead_size=2MB
```

Push config qua các node CephSSD01-06

```
su cephuser
cd /home/cephuser/ceph-deploy/
ceph-deploy --overwrite-conf config push nhcephssd01 nhcephssd02
nhcephssd03 nhcephssd04 nhcephssd05 nhcephssd06
```

Push config qua các node Controller, Compute

```
cd /home/cephuser/ceph-deploy/
scp /etc/ceph/ceph.conf root@{host}:/etc/ceph/
```

Restart service Ceph-mon trên cả 3 node CephSSD01-> 03

```
systemctl restart ceph-mon@$(hostname)
```

7.2 Revert

Cấu hình Ceph.conf ban đầu

```
[global]
# Debug config
debug lockdep = 0/0
debug context = 0/0
debug crush = 0/0
debug mds = 0/0
debug mds balancer = 0/0
debug mds locker = 0/0
debug mds log = 0/0
debug mds log expire = 0/0
debug mds migrator = 0/0
debug buffer = 0/0
debug timer = 0/0
debug filer = 0/0
debug objecter = 0/0
debug rados = 0/0
debug rbd = 0/0
debug journaler = 0/0
debug objectcacher = 0/0
debug client = 0/0
debug osd = 0/0
debug optracker = 0/0
debug objclass = 0/0
debug filestore = 0/0
debug journal = 0/0
debug ms = 0/0
debug mon = 0/0
debug monc = 0/0
debug paxos = 0/0
debug tp = 0/0
debug auth = 0/0
debug finisher = 0/0
debug heartbeatmap = 0/0
debug perfcounter = 0/0
debug rgw = 0/0
debug hadoop = 0/0
debug asok = 0/0
```

```
debug throttle = 0/0
rbd default format = 2

# Network
public network = 192.168.72.0/24
cluster network = 192.168.73.0/24

# UUID
fsid = 4d41a358-1ab8-xxxxxx-xxxxxxxx
mon_initial_members = nhcephssd01, nhcephssd02, nhcephssd03
mon_host = 192.168.72.51,192.168.72.52,192.168.72.53

# Authentication mechanism
auth_cluster_required = cephx
auth_service_required = cephx
auth_client_required = cephx

# Bluestore
bluestore block db size = 5737418240
bluestore block wal size = 2737418240

# Replica
osd pool default size = 2
osd pool default min size = 1

# Chooseleaf type
osd crush chooseleaf type = 1

# Disable auto update crush => Modify Crushmap OSD tree
osd crush update on start = false

# Backfilling and recovery
osd max backfills = 1
osd recovery max active = 1
osd recovery max single start = 1
osd recovery op priority = 1

# Osd recovery threads = 1
osd backfill scan max = 16
osd backfill scan min = 4
osd backfill full ratio = 0.95

[mon]
        mon data = /var/lib/ceph/mon/\$cluster-\$id
        #mon allow pool delete = true

[mon.nhcephssd01]
        host = nhcephssd01
        mon data = /var/lib/ceph/mon/ceph-nhcephssd01
        mon addr = 192.168.72.51:6789

[mon.nhcephssd02]
        host = nhcephssd02
        mon data = /var/lib/ceph/mon/ceph-nhcephssd02
```

```
        mon addr = 192.168.72.52:6789

[mon.nhcephssd03]
        host = nhcephssd03
        mon data = /var/lib/ceph/mon/ceph-nhcephssd03
        mon addr = 192.168.72.53:6789
```

Push lại config qua các node CephSSD01-06

```
su cephuser
cd /home/cephuser/ceph-deploy/
ceph-deploy --overwrite-conf config push nhcephssd01 nhcephssd02
nhcephssd03 nhcephssd04 nhcephssd05 nhcephssd06
```

Push config qua các node Controller, Compute

```
cd /home/cephuser/ceph-deploy/
scp /etc/ceph/ceph.conf root@{host}:/etc/ceph/
```

Restart service Ceph-mon trên cả 3 node CephSSD01-> 03

```
systemctl restart ceph-mon@$(hostname)
```

Restart service cinder trên controller

```
systemctl restart cinder-*
```

8. Tunning TCmalloc 256MB (IOPS)

> Require restart Ceph-mon service, ceph-osd server

- Đối tượng áp dụng: CephNodeSSD01->06, Compute 1->7
- Mục đích: Tăng số lượng files kernel có thể mở và xử lý
- Cài đặt và cấu hình `tuned-adm profile throughput-performance`
- Revert lại cấu hình mặc định nếu có vấn đề

Ubuntu

```
/etc/default/ceph
```

CentOS

```
/etc/sysconfig/ceph
```

Tham số 128MB

```
TCMALLOC_MAX_TOTAL_THREAD_CACHE_BYTES==134217728
```

9. Tài liệu tham khảo

- Config MTU Nexus 3000: https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus3000/sw/qos/503_u1_1/b_Cisco_nexus3000_qos_config_gd_503_u1_1/b_Cisco_nexus3000_qos_config_gd_503_u1_1_chapter_010.html
- https://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus3000/sw/qos/7x/b_3k_QoS_Config_7x/b_3k_QoS_Config_7x_chapter_010.html
- https://www.cisco.com/c/en/us/support/docs/switches/nexus-9000-series-switches/118994-config-nexus-00.html
- https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/2/html/ceph_object_gateway_for_production/deploying_a_cluster
- https://cromwell-intl.com/open-source/performance-tuning/disks.html
- https://www.kernel.org/doc/Documentation/block/queue-sysfs.txt
- https://wiki.mikejung.biz/Ubuntu_Performance_Tuning
- https://access.redhat.com/sites/default/files/attachments/20150325_network_performance_tuning.pdf
- TCP tunning sysctl: https://fatmin.com/2015/08/19/ceph-tcp-performance-tuning/
- Advance Tunning BlockStorage: https://www.openstack.org/assets/presentation-media/Advanced-Tuning-and-Operation-guide-for-Block-Storage-using-Ceph-Boston-2017-final.pdf
- TCMalloc tunning: https://access.redhat.com/documentation/en-us/red_hat_ceph_storage/2/html/ceph_object_gateway_for_production/deploying_a_cluster
- ceph osd crush tunables optimal # optimal: the best (ie optimal) values of the current version of Ceph
- http://docs.ceph.com/docs/mimic/rados/operations/crush-map-edits/
- http://docs.ceph.com/docs/jewel/rados/operations/crush-map/
- http://www.mellanox.com/blog/2016/02/making-ceph-faster-lessons-from-performance-testing/