RHCS6: Quorum disk and heuristics

Article Number: 199 | Rating: Unrated | Last Updated: Sun, Jun 3, 2018 9:39 AM

RHCS: Quorum disk and heuristics

```
# Tested on RHEL 6

# A quorum disk is usually used as a tie-breaker to determine which
node should be fenced
# in case of problems.

# It adds a number of votes to the cluster in a way that a "last-man-
standing" scenario
# can be configured.

# The node with the lowest nodeid that is currently alive will become
the "master", who
# is responsible for casting the votes assigned to the quorum disk as
well as handling
# evictions for dead nodes.

# Every node of the cluster will write at regular intervals to its
own block on a
# quorum disk to show itself as available; if a node fails to update
its block it will
# be considered as unavailable and will be evicted. Obviously this is
useful to
# determine whether a node that doesn't respond over the network is
really down or just
# having network problems. 'cman' network timeout for evicting nodes
should be set at
# least twice as high as the timeout for evicting nodes based on
```

```
their quorum disk
# updates.
# From RHEL 6.3 on, a node that can communicate over the network but
has problems to
# write to quorum disk will send a message to other cluster nodes and
will avoid to
# be evicted from the cluster.

# 'cman' network timeout is called "Totem Timeout" can be set by
adding
#   <totem token="timeout_in_ms"/>  to /etc/cluster/cluster.conf

# Quorum disk has to be at least 10MB in size and it has to be
available to all nodes.

# A quorum disk may be specially useful in these configurations:
#
# - Two node clusters with separate network for cluster
communications and fencing.
#   The "master" node will win any fence race. From RHEL 5.6 and RHEL
6.1 delayed
#   fencing should be used instead
#
# - Last-man-standing cluster




# Configuring a quorum disk
# ------------------------------------------------------------------
----------------------

# Take a disk or partition that is available to all nodes and run
following command
# mkqdisk -c <device> -l <label>:

mkqdisk -c /dev/sdb -l quorum_disk
   mkqdisk v3.0.12.1
```

```
   Writing new quorum disk label 'quorum_disk' to /dev/sdb.
   WARNING: About to destroy all data on /dev/sdb; proceed [N/y] ? y
   Initializing status block for node 1...
   Initializing status block for node 2...
   [...]
   Initializing status block for node 15...
   Initializing status block for node 16...


# Check (visible from all nodes)

mkqdisk -L
   mkqdisk v3.0.12.1

   /dev/block/8:16:
   /dev/disk/by-id/ata-VBOX_HARDDISK_VB0ea68140-6869d321:
   /dev/disk/by-id/scsi-SATA_VBOX_HARDDISK_VB0ea68140-6869d321:
   /dev/disk/by-path/pci-0000:00:0d.0-scsi-1:0:0:0:
   /dev/sdb:
           Magic:                eb7a62c2
           Label:                quorum_disk
           Created:              Thu Jul 31 16:57:36 2014
           Host:                 nodeB
           Kernel Sector Size:   512
           Recorded Sector Size: 512




# Scoring & Heuristics
# ------------------------------------------------------------------
-----------------------

# As an option, one or more heuristics can be added to the cluster
configuration.
# Heuristics are tests run prior to accessing the quorum disk. These
are sanity checks for
```

```
# the system. If the heuristic tests fail, then qdisk will, by
default, reboot the node in
# an attempt to restart the machine in a better state.

# We can configure up to 10 purely arbitrary heuristics. It is
generally a good idea to
# have more than one heuristic. By default, only nodes scoring over
1/2 of the total
# maximum score will claim they are available via the quorum disk,
and a node whose score
# drops too low will remove itself (usually, by rebooting).

# The heuristics themselves can be any command executable by "sh -c".

# Typically, the heuristics should be snippets of shell code or
commands which help
# determine a node's usefulness to the cluster or clients. Ideally,
we want to add traces
# for all of our network paths, and methods to detect availability of
shared storage.




# Adding the quorum disk to our cluster
# ------------------------------------------------------------------
-----------------------

# On 'luci' management console (first, check the box under Homebase
--> Preferences to have
# access to "Expert" mode), go to cluster administration -->
Configure --> QDisk, check
# "Use a Quorum Disk" and "By Device Label", and enter the label
given to the quorum disk.
# Define a TKO (Times to Knock Out), the number of votes and the
interval for the quorum
# disk to be updated by every node.
```

```
# The interval (timeout) of the qdisk is by default 1 second. If the
load on the system is
# high, it is very easy for the qdisk cycle to take more than 1
second (I'll set it to 3).

# Totem token is set to 10 seconds by default. This is too short in
most cases. A simple
# rule to configure totem timeout could be "a little bit" more than 2
x qdiskd's timeout.
# I'll set it to 50 seconds (50000 ms)

# After adding the quorum disk on Luci console, we'll have following
entry in our
# /etc/cluster/cluster.conf file:

#       <quorumd interval="3" label="quorum_disk" tko="8"
votes="1"/>
#       <totem token="50000"/>

# On the command line, we can run following commands to obtain same
result:

ccs -f /etc/cluster/cluster.conf --settotem token=70000
ccs -f /etc/cluster/cluster.conf --setquorumd interval=3
label=quorum_disk tko=8 votes=1

# If we had defined a heuristic in "Heuristics" section, by entering
heuristic program
# (I'll use three pings to different servers as heuristics), interval
score and tko, we'd
# have the following:

#   <quorumd interval="3" label="quorum_disk" tko="8" votes="1">
#      <heuristic program="/sbin/ping nodeC -c1 -w1" tko="8"/>
#      <heuristic program="/sbin/ping nodeD -c1 -w1" tko="8"/>
#      <heuristic program="/sbin/ping nodeE -c1 -w1" tko="8"/>
#   </quorumd>
```

```
# Once our quorum disk is configured, the option "expected_votes" must be adapted
and
# the option "two_node" is not necessary anymore so have to change following line
in
# cluster.conf file:
#
      <cman expected_votes="1" two_node="1">


# by
       <cman expected_votes="3">




# Do not forget to propagate changes to the rest of nodes in the
cluster

ccs -h nodeA -p myriccipasswd --sync --activate




# Quorum disk timings
# ------------------------------------------------------------------
-----------------------

# Qdiskd should not be used in environments requiring failure
detection times of less than
# approximately 10 seconds.
#
# Qdiskd will attempt to automatically configure timings based on the
totem timeout and
# the TKO. If configuring manually, Totem's token timeout must be set
to a value at least
# 1 interval greater than the the following function:
```

```
#
# interval * (tko + master_wait + upgrade_wait)
#
# So, if you have an interval of 2, a tko of 7, master_wait of 2 and
upgrade_wait of 2,
# the token timeout should be at least 24 seconds (24000 msec).
#
# It is recommended to have at least 3 intervals to reduce the risk
of quorum loss during
# heavy I/O load. As a rule of thumb, using a totem timeout more than
2x of qdiskd's
# timeout will result in good behavior.
#
# An improper timing configuration will cause CMAN to give up on
qdiskd, causing a
# temporary loss of quorum during master transition.




# Show cluster basic information
# ---------------------------------------------------------------------
----------------------

# Before adding a quorum disk:

cat /etc/cluster/cluster.conf
   <?xml version="1.0"?>
   <cluster config_version="25" name="mycluster">
      <clusternodes>
         <clusternode name="nodeA" nodeid="1"/>
         <clusternode name="nodeB" nodeid="2"/>
      </clusternodes>
      <cman expected_votes="1" two_node="1">
         <multicast addr="239.192.XXX.XXX"/>
      </cman>
      <rm log_level="7"/>
</cluster>
```

```
clustat   # (RGManager cluster)
   Cluster Status for mycluster @ Thu Jul 31 17:13:41 2014
   Member Status: Quorate

    Member Name
ID   Status
    ------ ----
---- ------

nodeA                                                    1
Online, Local

nodeB                                                    2
Online


cman_tool status
   Version: 6.2.0
   Config Version: 25
   Cluster Name: mycluster
   Cluster Id: 4946
   Cluster Member: Yes
   Cluster Generation: 168
   Membership state: Cluster-Member
   Nodes: 2
   Expected votes: 1
   Total votes: 2
   Node votes: 1
   Quorum: 1
   Active subsystems: 8
   Flags: 2node
   Ports Bound: 0 11
   Node name: nodeA
   Node ID: 1
   Multicast addresses: 239.192.XXX.XXX
   Node addresses: XXX.XXX.XXX.XXX
```

```
# After adding a quorum disk:

cat /etc/cluster/cluster.conf
   <?xml version="1.0"?>
   <cluster config_version="26" name="mycluster">
      <clusternodes>
         <clusternode name="nodeA" nodeid="1"/>
         <clusternode name="nodeB" nodeid="2"/>
      </clusternodes>
      <cman expected_votes="3">
         <multicast addr="239.192.XXX.XXX"/>
      </cman>
      <rm log_level="7"/>
      <quorumd interval="3" label="quorum_disk" tko="8" votes="1">
         <heuristic program="/sbin/ping nodeC -c1 -w1" tko="8"/>
         <heuristic program="/sbin/ping nodeD -c1 -w1" tko="8"/>
         <heuristic program="/sbin/ping nodeE -c1 -w1" tko="8"/>
      </quorumd>
      <totem token="70000"/>
</cluster>


clustat   # (RGManager cluster)
   Cluster Status for mycluster @ Thu Jul 31 17:20:07 2014
   Member Status: Quorate

    Member Name
ID   Status
    ------ ----
---- ------

nodeA                                                    1
Online, Local

nodeB                                                    2
```

```
Online

/dev/sdb                                                        0
Online, Quorum Disk


cman_tool status
   Version: 6.2.0
   Config Version: 28
   Cluster Name: mycluster
   Cluster Id: 4946
   Cluster Member: Yes
   Cluster Generation: 168
   Membership state: Cluster-Member
   Nodes: 2
   Expected votes: 3
   Quorum device votes: 1
   Total votes: 3
   Node votes: 1
   Quorum: 2
   Active subsystems: 11
   Flags:
   Ports Bound: 0 11 177 178
   Node name: nodeA
   Node ID: 1
   Multicast addresses: 239.192.XXX.XXX
   Node addresses: XXX.XXX.XXX.XXX
```

Posted - Sun, Jun 3, 2018 9:39 AM. This article has been viewed 4307 times.

Online URL: http://kb.ictbanking.net/article.php?id=199