OpenNebula - Bug #3335
Use CEPH "MAX AVAIL"

11/13/2014 12:34 PM - Tino Vázquez

Status:	Closed	**Start date:**	11/13/2014
Priority:	High	**Due date:**	
Assignee:		**% Done:**	0%
Category:	Drivers - Storage	**Estimated time:**	0.00 hour
Target version:			
Resolution:	worksforme	**Pull request:**	
Affected Versions:	OpenNebula 4.12		

Description

It accounts for the number of replicas.

Proposed modification:

```
MONITOR_SCRIPT=$(cat <<EOF
to_megabytes() {
value=\$(echo "\$1" | sed 's/[^0123456789].*\$//g')
units=\$(echo "\$1" | sed 's/^[0123456789]*//g' | tr '[:upper:]' '[:lower:]')
case "\$units" in
t|tb) value=\$(expr \$value * 1024 * 1024) ;;
g|gb) value=\$(expr \$value * 1024) ;;
m|mb) value=\$value ;;
k|kb) value=\$(expr \$value / 1024) ;;
b|'') value=0 ;;
*)
value=
echo "Unsupported units '\$units'" >&2
;;
esac
echo "\$value"
}
used=\$(ceph df | awk '{
if (\$1 == "$POOL_NAME") {
print \$3
}
}')
total=\$(ceph df | awk '{
if (\$1 == "$POOL_NAME") {
print \$5
}
}')
used_mb=\$(to_megabytes \$used)
total_mb=\$(to_megabytes \$total)
free_mb=\$(expr \$total_mb - \$used_mb)
echo "USED_MB=\$used_mb"
echo "FREE_MB=\$free_mb"
echo "TOTAL_MB=\$total_mb"
EOF
)
```

Associated revisions

Revision 8174fe08 - 02/17/2015 02:23 PM - Ruben S. Montero

feature #3335: Use MAX AVAIL to compute Ceph datastore usage

Revision c1a798ae - 02/25/2015 04:30 PM - Ruben S. Montero

feature #3335: Better compute of TOTAL. Cluster total cannot be used as it does not consider placement rules in the crushmap

History

#1 - 11/13/2014 11:15 PM - Ruben S. Montero

- Tracker changed from Request to Feature

- Status changed from Pending to New

- Target version set to Release 4.12

#2 - 11/21/2014 02:15 PM - Bill Campbell

Great idea, however I see one problem with your script. It is using the MAX_AVAIL metric in place of the TOTAL in the datastore, which ends up shrinking as you allocate images. I recommend the following modifications to the above:

- Replica size is pulled from the cluster (utilizing 'ceph osd pool get $POOL_NAME size')
- TOTAL is calculated by pulling the total size from 'rados df'->convert to megabytes->divide by replica size
- MAX_AVAIL is collected to show free space in the cluster
- USED stays the same

This way, you have a static total, the MAX_AVAIL is shown as the free space, and we don't need to worry about changing this per pool (in the case you use multiple pools in your cluster with different replication sizes), all while still taking into consideration the replication size of the pool.

Example modification to above, I'm sure OpenNebula gurus can clean this up a bit:

```
MONITOR_SCRIPT=$(cat <<EOF
REPSIZE=\$(/usr/bin/ceph osd pool get "$POOL_NAME" size | awk '{print \$2}')
to_megabytes() {
value=\$(echo "\$1" | sed 's/[^0123456789].*\$//g')
units=\$(echo "\$1" | sed 's/^[0123456789]*//g' | tr '[:upper:]' '[:lower:]')
case "\$units" in
t|tb) value=\$(expr \$value * 1024 * 1024) ;;
g|gb) value=\$(expr \$value * 1024) ;;
m|mb) value=\$value ;;
k|kb) value=\$(expr \$value / 1024) ;;
b|'') value=0 ;;
*)
value=
echo "Unsupported units '\$units'" >&2
;;
esac
echo "\$value"
}
used=\$(ceph df | awk '{
if (\$1 == "$POOL_NAME") {
```

```
print \$3
}
}')
total=\$($RADOS df | grep 'total space' | awk '{print \$3}')
free=\$(ceph df | awk '{
if (\$1 == "$POOL_NAME") {
print \$5
}
}')
used_mb=\$(to_megabytes \$used)
total_mb=\$(expr \$total / 1024)
free_mb=\$(to_megabytes \$free)
total_pool=\$(expr \$total_mb / \$REPSIZE)
echo "USED_MB=\$used_mb"
echo "FREE_MB=\$free_mb"
echo "TOTAL_MB=\$total_pool"
EOF
)
```

#3 - 02/17/2015 03:38 PM - Ruben S. Montero

- Status changed from New to Closed

- Resolution set to fixed

This is now in master. I've used the suggestions by Bill. A few notes for reference

MAX AVAIL, is very conservative. It takes the less "projected" available space (per OSD) in the PG MAP. (The relative weight of the OSD in the crush placement rule is used to make the projection see [1], i.e. if an OSD is 0.5 - because its weight is half of total weight sum in the ruleset - the estimated free space in the pool is 2x the free space of the OSD). When a cluster is beeing re-weighted this should be considered.

Also I'm getting "max_avail": 4611686018427387904 for our develop cluster (4096P), It seems the code is not considering overall weights of 0. This should not happen in production, but I added a fallback when MAX AVAIL is 0

TOTAL is computed as suggested, dividing it by the replica size. The very same is done in the Ceph code for max avail. This will give a better overview than the raw total. Note also that this will not work when the osd type is set to erasure code.

[1] https://github.com/ceph/ceph/blob/master/src/mon/PGMonitor.cc#L1303

#4 - 03/23/2015 04:38 PM - Mo McRoberts

Hi,

For info, this (I'm pretty sure it's this, though I could be wrong) breaks on Ceph 0.80.7-0ubuntu0.14.04.1 — it reports FREE_MB as zero, causing OpenNebula to refuse to schedule any VMs dependent upon images in the Ceph pool. I just upgraded to 4.12.0 and discovered this the hard way and had to do a little digging to track down (fortunately, I've got copies of the old Ceph datastore scripts to compare with).

The culprit appears to be that MAX_AVAIL=\$($CEPH df | grep "$POOL_NAME" | awk '{print \$5}') in remotes/datastore/ceph/monitor always results in MAX_AVAIL being zero.

My ceph df output looks like this:

```
GLOBAL:
    SIZE     AVAIL     RAW USED     %RAW USED
    327T     319T      8530G        2.54
POOLS:
    NAME                 ID     USED       %USED     OBJECTS
    data                 0      73039M     0.02      28120
    metadata             1      247M       0         1784
    rbd                  2      0          0         0
    one                  3      2747G      0.82      736798
                         13     0          0         0
    .rgw                 15     3273       0         21
    .rgw.root            16     822        0         3
    .rgw.control         17     0          0         8
    .rgw.gc              18     0          0         32
    .rgw.buckets         19     8414M      0         322262
    .rgw.buckets.index   20     0          0         11
    .log                 21     0          0         0
    .intent-log          22     0          0         0
    .usage               23     0          0         0
    .users               24     24         0         3
    .users.email         25     24         0         3
    .users.swift         26     0          0         0
    .users.uid           27     1033       0         6
```

I've worked around it by reverting the monitor script to:

```
MONITOR_SCRIPT=$(cat <<EOF
$RADOS df | $AWK '{
    if (\$1 == "total") {

        space = int(\$3/1024)

        if      (\$2 == "used")  {var = "USED_MB"}
        else if (\$2 == "avail") {var = "FREE_MB"}
        else if (\$2 == "space") {var = "TOTAL_MB"}

        print var "=" space
    }
}'
EOF
)
```

(And ensuring that RADOS is set to pass --id if CEPH_USER is set).

I haven't seen any indication of minimum version requirement for Ceph+OpenNebula, nor that 0.80.7 is too old, although it might be because I'm stupid. It's also possible that this broke in a previous release and was worked around by somebody else who did the platform upgrade in our setup. Apologies if this is unhelpful!

#5 - 03/23/2015 05:10 PM - Jaime Melis

- Tracker changed from Feature to Bug

- Status changed from Closed to Pending

- Target version deleted (Release 4.12)

- Resolution deleted (fixed)

- Affected Versions OpenNebula 4.12 added

Review the FREE_MB value in Ceph 0.80.7

#6 - 03/24/2015 09:20 AM - Ruben S. Montero

Just double checking this... In the testing environment we have:

```
ii  ceph                         0.80.7-0ubuntu0.14.10.1             amd64        distributed storage and file system
ii  ceph-common                  0.80.7-0ubuntu0.14.10.1             amd64        common utilities to mount and interact with a ceph storage cluster
```

And the output:

```
oneadmin@kvm1:~$ ceph df
GLOBAL:
    SIZE      AVAIL     RAW USED     %RAW USED
    1951M     1445M         372M         19.07
POOLS:
    NAME         ID     USED     %USED     MAX AVAIL     OBJECTS
    data         0         0         0         4096P           0
    metadata     1         0         0         4096P           0
    rbd          2         0         0         4096P           0
    one          3      164M      8.40         4096P          45
```

It seems that we are using the same version. I do not understand the diferences...

EDIT: BTW we don't have 4096 Petabytes for testing it seems a bug when all de OSDs have 0 weight, but the column its there

#7 - 03/27/2015 10:07 PM - Mo McRoberts

Hello!

Thanks for digging into this… and I've done some of my own.

It turns out the reason seems to be that *some* of the packages on one of the nodes in the cluster were out of date, and it was holding the whole cluster back to an older version.

My output looks a lot more sensible now, so I should be able to restore the monitor script back to its shipped version:

```
GLOBAL:
    SIZE     AVAIL     RAW USED     %RAW USED
    327T      317T       10102G         3.01
POOLS:
```

NAME	ID	USED	%USED	MAX AVAIL	OBJECTS
data	0	360G	0.11	105T	104339
metadata	1	248M	0	105T	1716
rbd	2	0	0	105T	0
one	3	2990G	0.89	105T	799143
	13	0	0	105T	0
.rgw	15	3273	0	105T	21
.rgw.root	16	822	0	105T	3
.rgw.control	17	0	0	105T	8
.rgw.gc	18	0	0	105T	32
.rgw.buckets	19	8504M	0	105T	325058
.rgw.buckets.index	20	0	0	105T	11
.log	21	0	0	105T	0
.intent-log	22	0	0	105T	0
.usage	23	0	0	105T	0
.users	24	24	0	105T	3
.users.email	25	24	0	105T	3
.users.swift	26	0	0	105T	0
.users.uid	27	1033	0	105T	6

Sorry about the noise!

M.

#8 - 03/28/2015 05:26 PM - Ruben S. Montero

- Status changed from Pending to Closed

- Resolution set to worksforme

Thanks for leetting us know, closing this one

Mo McRoberts wrote:

> *Hello!*
>
> *Thanks for digging into this… and I've done some of my own.*
>
> *It turns out the reason seems to be that some* of the packages on one of the nodes in the cluster were out of date, and it was holding the whole cluster back to an older version.
>
> My output looks a lot more sensible now, so I should be able to restore the monitor script back to its shipped version:
>
> [...]
>
> Sorry about the noise!
>
> M.