Ceph - Bug #1186

Cluster won't recover, OSD's go up and down again (and stay down)

06/14/2011 05:09 AM - Wido den Hollander

Status:	Closed	**% Done:**	0%
Priority:	Normal	**Spent time:**	13.00 hours
Assignee:			
Category:	OSD		
Target version:			
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

Ok, the title might be somewhat confusing, but so is the issue :)

I'm still trying to get my 40 OSD cluster back into a healthy state, but this won't finish.

After a lot of bugs my OSD's don't assert anymore (for now), but they won't come up either.

I started my cluster (d2b7e291f21928f9f0a3e23fb32c94c9cbbc8984) this morning and slowly the OSD's started to come up:

Slowly I saw the OSD's coming up one by one until 26 up/in, after that it started to go down, until I reached:

```
2011-06-14 13:58:31.692401    pg v660657: 10608 pgs: 333 inactive, 153 active+clean, 107 active+de
graded, 134 active+clean+degraded, 7269 crashed+down+peering, 2612 crashed+down+degraded+peering;
2108 GB data, 0 KB used, 0 KB / 0 KB avail; 245445/1626390 degraded (15.091%)
2011-06-14 13:58:31.692564   osd e43540: 40 osds: 0 up, 0 in
```

What I did notice, the whole time the state of the cluster stayed at:

```
2011-06-14 13:58:31.692401    pg v660657: 10608 pgs: 333 inactive, 153 active+clean, 107 active+de
graded, 134 active+clean+degraded, 7269 crashed+down+peering, 2612 crashed+down+degraded+peering;
2108 GB data, 0 KB used, 0 KB / 0 KB avail; 245445/1626390 degraded (15.091%)
2011-06-14 13:58:31.692564   osd e43540: 40 osds: 0 up, 0 in
```

The first thing I did was verifying if all cosd processes are running and yes, they are.

```
root@monitor:~# dsh -g osd-mdb "pidof cosd|wc -w"
4
4
4
4
4
4
4
4
3
3
root@monitor:~#
```

In the last two boxes I have two crashed disks, so I have 38 working OSD's.

At first I thought it is/was the Atom CPU, but the load on the machines isn't that high:

```
root@atom0:~# ps aux|grep cosd
root      3240 22.1 24.1 1493176 981080 ?      Ssl  11:54  29:02 /usr/bin/cosd -i 0 -c /etc/ceph/c
eph.conf
root      3354 70.3  6.9 1336408 281720 ?      Ssl  11:54  91:58 /usr/bin/cosd -i 1 -c /etc/ceph/c
eph.conf
root      3627 20.1 24.6 1523364 1002060 ?     Ssl  11:54  26:17 /usr/bin/cosd -i 2 -c /etc/ceph/c
eph.conf
root      3900 20.0 27.0 1472972 1097488 ?     Ssl  11:54  26:07 /usr/bin/cosd -i 3 -c /etc/ceph/c
eph.conf
root     10566  0.0  0.0   7676   828 pts/0    S+   14:04   0:00 grep --color=auto cosd
root@atom0:~# uptime
 14:04:53 up  2:46,  1 user,  load average: 1.00, 1.04, 1.16
root@atom0:~#
```

'debug osd = 20' is set on all the OSD's, so I have enough log information. Checking out the logs the OSD's all seem to be different stuff, but they actually are active and alive!

My goal is still to recover this cluster as it seems (imho) to be a pretty good test case for bringing a downed cluster back to life, isn't it?

The logs are groing pretty fast, about 15G per hour, so uploading isn't really an option.

History

#1 - 06/14/2011 10:05 AM - Samuel Just

A bit more information:

for i in {0..9}; do ssh root@atom$i 'uptime'; done
19:01:32 up 7:42, 0 users, load average: 4.06, 4.75, 4.89
19:01:33 up 7:42, 0 users, load average: 130.68, 130.63, 128.30
19:01:34 up 7:39, 0 users, load average: 0.21, 0.19, 0.26
19:01:34 up 7:42, 0 users, load average: 143.05, 143.08, 142.22
19:01:35 up 7:42, 0 users, load average: 123.71, 124.29, 123.77
19:01:35 up 7:42, 0 users, load average: 140.73, 140.65, 139.93
19:01:36 up 7:42, 0 users, load average: 4.49, 3.83, 3.51
19:01:37 up 7:42, 0 users, load average: 1.59, 1.21, 1.58
19:01:39 up 7:39, 0 users, load average: 1.56, 1.35, 1.33
19:01:40 up 7:38, 0 users, load average: 0.43, 0.67, 0.62

It looks like some of the osds do have very high load, still looking...

#2 - 06/14/2011 01:33 PM - Samuel Just

atom2 with cosd daemons killed:
procs -----------memory---------- ---swap-- -----io---- ~~system~~ ----cpu----
r b   swpd   free   buff cache   si   so   bi   bo   in   cs us sy id wa
1 0 1710908 3697360   3624 138640 139 104   204   478   27   23 16 15 64 5
1 0 1710908 3697352   3624 138664   0   0   0   0 235   14 0 44 56 0
1 0 1710908 3697352   3624 138664   0   0   0   0 206   16 0 20 80 0
1 0 1710896 3696996   3624 138916 204   0   448   0 251   74 0 24 76 0
1 0 1710896 3697004   3624 138908   0   0   0   0 205   22 0 27 73 0
1 0 1710896 3697004   3624 138908   0   0   0   0 212   18 0 32 68 0
1 0 1710896 3697004   3624 138908   0   0   0   0 228   16 0 20 80 0

```
1  0 1710896 3697004   3624 138908    0    0     0     0  222   18  0 27 73  0
1  0 1710896 3697004   3624 138908    0    0     0     0  239   18  0 23 77  0
1  0 1710896 3697004   3624 138908    0    0     0     0  254   18  0 28 72  0
1  0 1710896 3697004   3624 138908    0    0     0     0  199   16  0 23 77  0
1  0 1710896 3697004   3624 138908    0    0     0     0  202   18  0 32 68  0
1  0 1710896 3697004   3624 138908    0    0     0     0  220   16  0 21 79  0
1  0 1710896 3697004   3624 138908    0    0     0     0  228   16  0 32 68  0
```

Seems like something is using up a lot of system time?

#3 - 06/14/2011 03:16 PM - Samuel Just

The monitor debugging also seems to have been a problem. Turning that down and restarting the machines has allowed 30 or so to come up, watching to see what happens with peering.

#4 - 06/15/2011 01:09 PM - Samuel Just

Failed an assert in ReplicaActive receiving a query of type other than Query::Missing. (assert(query.query.type == Query::MISSING);)

#5 - 06/23/2011 12:51 PM - Wido den Hollander

- Status changed from New to Closed

I'm going to close this one as well. I formatted my cluster today and started with a fresh v0.29.1 cluster, had to due to my monitor being corrupted.