Ceph - Bug #490

Cluster stays in a degraded state

10/14/2010 12:19 PM - Wido den Hollander

Status:	Can't reproduce	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.50 hour
Assignee:			
Category:			
Target version:			
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

My cluster is staying in a degraded state for the last few days.

```
2010-10-14 21:17:38.322482    pg v34543: 3176 pgs: 72 active, 3104 active+clean; 831 GB data, 2474
 GB used, 3852 GB / 6326 GB avail; 300/3108741 degraded (0.010%)
2010-10-14 21:17:38.330531   mds e60: 1/1/1 up {0=up:replay(laggy or crashed)}
2010-10-14 21:17:38.330562   osd e466: 12 osds: 12 up, 12 in
2010-10-14 21:17:38.330681   log 2010-10-14 14:27:06.289919 mon0 [2001:16f8:10:2::c3c3:af78]:6789/
0 12 : [INF] mon.logger@0 won leader election with quorum 0,1,2
2010-10-14 21:17:38.330935   class rbd (v1.2 [x86-64])
2010-10-14 21:17:38.330953   mon e1: 3 mons at {logger=[2001:16f8:10:2::c3c3:af78]:6789/0,node13=[
2001:16f8:10:2::c3c3:3f9b]:6789/0,node14=[2001:16f8:10:2::c3c3:2e5c]:6789/0}
```

I checked all the pools (if one of them is blocking), but with **rados -p <pool> ls** I can list all my pools.

Restarting the OSD's doesn't make any difference, those 300 objects stay degraded.

Any way to list or find out which objects are degraded? Is there a way to dump that?

My OSD status:

```
2010-10-14 21:19:53.686620 mon <- [osd,dump]
2010-10-14 21:19:53.692862 mon1 -> 'dumped osdmap epoch 466' (0)
epoch 466
fsid 795255b3-7f59-193f-153b-929336fdf29c
created 2010-10-09 11:27:28.419690
modifed 2010-10-13 11:21:04.983593
flags

pg_pool 0 'data' pg_pool(rep pg_size 3 crush_ruleset 0 object_hash rjenkins pg_num 768 pgp_num 768
 lpg_num 2 lpgp_num 2 last_change 19 owner 0)
pg_pool 1 'metadata' pg_pool(rep pg_size 3 crush_ruleset 1 object_hash rjenkins pg_num 768 pgp_num
 768 lpg_num 2 lpgp_num 2 last_change 15 owner 0)
pg_pool 2 'casdata' pg_pool(rep pg_size 2 crush_ruleset 2 object_hash rjenkins pg_num 768 pgp_num
768 lpg_num 2 lpgp_num 2 last_change 1 owner 0)
pg_pool 3 'rbd' pg_pool(rep pg_size 3 crush_ruleset 3 object_hash rjenkins pg_num 768 pgp_num 768
lpg_num 2 lpgp_num 2 last_change 23 owner 0)
pg_pool 4 'iscsi' pg_pool(rep pg_size 2 crush_ruleset 0 object_hash rjenkins pg_num 8 pgp_num 8 lp
g_num 0 lpgp_num 0 last_change 241 owner 0)

max_osd 12
osd0 in weight 1 up   (up_from 392 up_thru 453 down_at 391 last_clean 248-390) [2001:16f8:10:2::c3
c3:8f6b]:6800/2836 [2001:16f8:10:2::c3c3:8f6b]:6801/2836
```

```
osd1 in weight 1 up   (up_from 398 up_thru 454 down_at 396 last_clean 258-395) [2001:16f8:10:2::c3
c3:a24f]:6800/28331 [2001:16f8:10:2::c3c3:a24f]:6801/28331
osd2 in weight 1 up   (up_from 403 up_thru 453 down_at 401 last_clean 274-400) [2001:16f8:10:2::c3
c3:4a8c]:6800/20449 [2001:16f8:10:2::c3c3:4a8c]:6801/20449
osd3 in weight 1 up   (up_from 417 up_thru 454 down_at 416 last_clean 409-416) [2001:16f8:10:2::c3
c3:2e3a]:6800/1600 [2001:16f8:10:2::c3c3:2e3a]:6801/1600
osd4 in weight 1 up   (up_from 413 up_thru 453 down_at 412 last_clean 307-411) [2001:16f8:10:2::c3
c3:fa1b]:6800/11251 [2001:16f8:10:2::c3c3:fa1b]:6801/11251
osd5 in weight 1 up   (up_from 417 up_thru 453 down_at 416 last_clean 324-415) [2001:16f8:10:2::c3
c3:3b6a]:6800/8984 [2001:16f8:10:2::c3c3:3b6a]:6801/8984
osd6 in weight 1 up   (up_from 423 up_thru 454 down_at 421 last_clean 329-420) [2001:16f8:10:2::c3
c3:3f6c]:6800/9704 [2001:16f8:10:2::c3c3:3f6c]:6801/9704
osd7 in weight 1 up   (up_from 432 up_thru 453 down_at 430 last_clean 346-429) [2001:16f8:10:2::c3
c3:2f6c]:6800/24892 [2001:16f8:10:2::c3c3:2f6c]:6801/24892
osd8 in weight 1 up   (up_from 437 up_thru 453 down_at 436 last_clean 359-435) [2001:16f8:10:2::c3
c3:1b6c]:6800/8675 [2001:16f8:10:2::c3c3:1b6c]:6801/8675
osd9 in weight 1 up   (up_from 442 up_thru 453 down_at 441 last_clean 366-440) [2001:16f8:10:2::c3
c3:2e56]:6800/4813 [2001:16f8:10:2::c3c3:2e56]:6801/4813
osd10 in weight 1 up   (up_from 448 up_thru 453 down_at 447 last_clean 371-446) [2001:16f8:10:2::c
3c3:2bfe]:6800/9206 [2001:16f8:10:2::c3c3:2bfe]:6801/9206
osd11 in weight 1 up   (up_from 453 up_thru 453 down_at 452 last_clean 376-451) [2001:16f8:10:2::c
3c3:ab76]:6800/30731 [2001:16f8:10:2::c3c3:ab76]:6801/30731

2010-10-14 21:19:53.692938 wrote 2738 byte payload to -
```

History

#1 - 10/14/2010 01:02 PM - Greg Farnum

ceph pg dump -o -
should let you know which PGs are degraded. If you're still running Cephx and having issues between two of your OSDs, I bet it's because there's a PG placed on those OSDs (with one as primary).

#2 - 10/15/2010 06:01 AM - Wido den Hollander

- File pg.txt added

Tnx, this shows me a lot of information, but it's not clear what tells me which PG is degraded.

Just checked my OSD logs, nothing new after the logrotate of this morning.

Do you think this is related to #462 ?

I've attached the pg dump, anything special there?

#3 - 11/12/2010 01:16 PM - Sage Weil

- Status changed from New to Can't reproduce

Files

pg.txt	278 KB	10/15/2010	Wido den Hollander