Ceph - Bug #2355

pgs stuck creating (with thrashing)

04/27/2012 01:51 PM - Josh Durgin

Status:	Resolved	**% Done:**	0%
Priority:	High	**Spent time:**	0.00 hour
Assignee:			
Category:	OSD		
Target version:	v0.47		
Source:	Development	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

I was running the following teuthology config, and since test_librbd_fsx creates a pool on each run, new pgs were created after thrashing started.
This was with ceph @ cbe795a7fe6d3da89ccd598f8e6338f33a495088.

```
interactive-on-error: true
roles:
- [mon.a, osd.0, osd.1, osd.2, osd.3, osd.4]
- [mon.b, osd.5, osd.6, osd.7, osd.8, osd.9]
- [mds.a, osd.10, osd.11, osd.12, osd.13, osd.14]
- [mon.c, client.0, client.1, client.2, client.3]
tasks:
- ceph:
    log-whitelist:
    - wrongly marked me down or wrong addr
    - objects unfound and apparently lost
    - clocks not synchronized
    conf:
      osd:
        debug ms: 1
        debug osd: 20
      client:
        debug ms: 1
        debug objecter: 20
        debug rbd: 20
        log_to_stderr: true
        rbd_cache: true
        client_oc_size: 8388608
- thrashosds:
    chance_down: 50
    min_live: 10
    min_in: 10
- rbd_fsx:
    clients: [client.0, client.1, client.2, client.3]
    size: 1073741824
    seed: 4097
    ops: 200000
```

The stuck creating pgs weren't mapped to anything, even though the osdmaptool said they should be at the same epoch:

```
$ grep creating pg_dump
```

```
6.5     0       0       0       0       0       0       0       creating        0.000000        0'
0     0'0     []      []      0'0     0.000000
5.6     0       0       0       0       0       0       0       creating        0.000000        0'
0     0'0     []      []      0'0     0.000000
4.7     0       0       0       0       0       0       0       creating        0.000000        0'
0     0'0     []      []      0'0     0.000000

$ ./osdmaptool --test-map-pg 6.5 osdmap
./osdmaptool: osdmap file 'osdmap'
 parsed '6.5' -> 6.5
6.5 raw [14,0] up [14,0] acting [14,0]
```

The osdmap at this point is:

```
epoch 27
fsid 40e11265-a74f-478f-9432-07887ff468a4
created 2012-04-26 19:00:04.733576
modifed 2012-04-26 19:35:19.330511
flags

pool 0 'data' rep size 2 crush_ruleset 0 object_hash rjenkins pg_num 60 pgp_num 60 lpg_num 0 lpgp_
num 0 last_change 1 owner 0 crash_replay_interval 45
pool 1 'metadata' rep size 2 crush_ruleset 1 object_hash rjenkins pg_num 60 pgp_num 60 lpg_num 0 l
pgp_num 0 last_change 1 owner 0
pool 2 'rbd' rep size 2 crush_ruleset 2 object_hash rjenkins pg_num 60 pgp_num 60 lpg_num 0 lpgp_n
um 0 last_change 1 owner 0
pool 3 'pool_client.0' rep size 2 crush_ruleset 0 object_hash rjenkins pg_num 8 pgp_num 8 lpg_num
0 lpgp_num 0 last_change 9 owner 0
pool 4 'pool_client.1' rep size 2 crush_ruleset 0 object_hash rjenkins pg_num 8 pgp_num 8 lpg_num
0 lpgp_num 0 last_change 9 owner 0
pool 5 'pool_client.3' rep size 2 crush_ruleset 0 object_hash rjenkins pg_num 8 pgp_num 8 lpg_num
0 lpgp_num 0 last_change 9 owner 0
pool 6 'pool_client.2' rep size 2 crush_ruleset 0 object_hash rjenkins pg_num 8 pgp_num 8 lpg_num
0 lpgp_num 0 last_change 9 owner 0

max_osd 15
osd.0 up   in  weight 1 up_from 5 up_thru 20 down_at 0 last_clean_interval [0,0) 10.214.131.33:680
0/1613 10.214.131.33:6801/1613 10.214.131.33:6802/1613 exists,up
osd.1 up   in  weight 1 up_from 6 up_thru 6 down_at 0 last_clean_interval [0,0) 10.214.131.33:6803
/1615 10.214.131.33:6804/1615 10.214.131.33:6805/1615 exists,up
osd.2 up   in  weight 1 up_from 6 up_thru 25 down_at 0 last_clean_interval [0,0) 10.214.131.33:680
6/1616 10.214.131.33:6807/1616 10.214.131.33:6808/1616 exists,up
osd.3 up   in  weight 1 up_from 2 up_thru 20 down_at 0 last_clean_interval [0,0) 10.214.131.33:680
9/1618 10.214.131.33:6810/1618 10.214.131.33:6811/1618 exists,up
osd.4 up   in  weight 1 up_from 5 up_thru 6 down_at 0 last_clean_interval [0,0) 10.214.131.33:6812
/1634 10.214.131.33:6813/1634 10.214.131.33:6814/1634 exists,up
osd.5 up   in  weight 1 up_from 3 up_thru 25 down_at 0 last_clean_interval [0,0) 10.214.132.27:680
0/1661 10.214.132.27:6801/1661 10.214.132.27:6802/1661 exists,up
osd.6 up   in  weight 1 up_from 6 up_thru 20 down_at 0 last_clean_interval [0,0) 10.214.132.27:680
3/1662 10.214.132.27:6804/1662 10.214.132.27:6805/1662 exists,up
osd.7 up   in  weight 1 up_from 6 up_thru 25 down_at 0 last_clean_interval [0,0) 10.214.132.27:680
6/1663 10.214.132.27:6807/1663 10.214.132.27:6808/1663 exists,up
osd.8 up   in  weight 1 up_from 20 up_thru 20 down_at 19 last_clean_interval [3,11) 10.214.132.27:
6815/2260 10.214.132.27:6816/2260 10.214.132.27:6817/2260 exists,up
osd.9 up   in  weight 1 up_from 6 up_thru 9 down_at 0 last_clean_interval [0,0) 10.214.132.27:6812
/1678 10.214.132.27:6813/1678 10.214.132.27:6814/1678 exists,up
osd.10 up   in  weight 1 up_from 5 up_thru 19 down_at 0 last_clean_interval [0,0) 10.214.132.37:68
00/1589 10.214.132.37:6801/1589 10.214.132.37:6802/1589 exists,up
osd.11 up   in  weight 1 up_from 3 up_thru 14 down_at 0 last_clean_interval [0,0) 10.214.132.37:68
03/1591 10.214.132.37:6804/1591 10.214.132.37:6805/1591 exists,up
osd.12 up   in  weight 1 up_from 3 up_thru 25 down_at 0 last_clean_interval [0,0) 10.214.132.37:68
06/1597 10.214.132.37:6807/1597 10.214.132.37:6808/1597 exists,up
osd.13 down out weight 0 up_from 18 up_thru 6 down_at 22 last_clean_interval [5,8) 10.214.132.37:6
```

```
809/2248 10.214.132.37:6816/2248 10.214.132.37:6817/2248 autoout,exists
osd.14 up   in  weight 1 up_from 5 up_thru 25 down_at 0 last_clean_interval [0,0) 10.214.132.37:68
13/1607 10.214.132.37:6814/1607 10.214.132.37:6815/1607 exists,up
```

The pg dump, osd dump, osd map, and crush map are with the logs in metropolis:/home/joshd/teuthology/archive/thrash_stuck_creating

Associated revisions

Revision 81f51d28 - 05/01/2012 05:39 PM - Sage Weil

osd: pg creation calc_priors_during() should count primary as up

If only want to include down osds if **all** of the prior acting osds are
down. If osd->whoami is one of them, then we're okay.

For example, if osd.13 is down, then the below should be satisfied that
osd.14 (osd->whoami) is alive:

```
2012-04-27 10:46:38.746681 7f5258a63700 15 osd.14 27 calc_priors_during 6.5 [9,25)
2012-04-27 10:46:38.746688 7f5258a63700 20 osd.14 27   6.5 in epoch 9 was [13,14]
2012-04-27 10:46:38.746695 7f5258a63700 20 osd.14 27   6.5 in epoch 10 was [13,14]
2012-04-27 10:46:38.746701 7f5258a63700 20 osd.14 27   6.5 in epoch 11 was [13,14]
2012-04-27 10:46:38.746709 7f5258a63700 20 osd.14 27   6.5 in epoch 12 was [13,14]
2012-04-27 10:46:38.746715 7f5258a63700 20 osd.14 27   6.5 in epoch 13 was [13,14]
2012-04-27 10:46:38.746722 7f5258a63700 20 osd.14 27   6.5 in epoch 14 was [13,14]
2012-04-27 10:46:38.746729 7f5258a63700 20 osd.14 27   6.5 in epoch 15 was [14]
2012-04-27 10:46:38.746735 7f5258a63700 20 osd.14 27   6.5 in epoch 16 was [14]
2012-04-27 10:46:38.746742 7f5258a63700 20 osd.14 27   6.5 in epoch 17 was [14]
2012-04-27 10:46:38.746748 7f5258a63700 20 osd.14 27   6.5 in epoch 18 was [13,14]
2012-04-27 10:46:38.746755 7f5258a63700 20 osd.14 27   6.5 in epoch 19 was [13,14]
2012-04-27 10:46:38.746762 7f5258a63700 20 osd.14 27   6.5 in epoch 20 was [13,14]
2012-04-27 10:46:38.746768 7f5258a63700 20 osd.14 27   6.5 in epoch 21 was [13,14]
2012-04-27 10:46:38.746775 7f5258a63700 20 osd.14 27   6.5 in epoch 22 was [14]
2012-04-27 10:46:38.746781 7f5258a63700 20 osd.14 27   6.5 in epoch 23 was [14]
2012-04-27 10:46:38.746788 7f5258a63700 20 osd.14 27   6.5 in epoch 24 was [14]
2012-04-27 10:46:38.746790 7f5258a63700 10 osd.14 27 calc_priors_during 6.5 [9,25) = 13
```

In that case, it wasn't, and the pg creation was blocked.

Fixes: #2355
Signed-off-by: Sage Weil <sage@newdream.net>

History

#1 - 05/01/2012 10:42 AM - Sage Weil

- Status changed from 12 to Fix Under Review

see wip-2355

#2 - 05/01/2012 10:50 AM - Sage Weil

- Status changed from Fix Under Review to Resolved

81f51d28d67c2a58ab621405c3da65aac726d719