Ceph - Bug #13988

new OSD re-using old OSD id fails to boot

12/05/2015 04:40 PM - Loic Dachary

Status:	Resolved	**Start date:**	12/05/2015
Priority:	Urgent	**Due date:**	
Assignee:	Loic Dachary	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:	other	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor		

Description

Steps to reproduce

```
teuthology-openstack --verbose --key-filename ~/Downloads/myself --key-name loic --teuthology-git-
url http://github.com/dachary/teuthology --teuthology-branch wip-suite --ceph-qa-suite-git-url htt
p://github.com/dachary/ceph-qa-suite --suite-branch wip-ceph-disk --ceph-git-url http://github.com
/dachary/ceph --ceph master --suite ceph-disk --filter ubuntu_14.04
```

It will sleep forever with two targets provisionned and ready to be used.

- ssh to the target that runs the monitory
- git clone http://github.com/ceph/ceph
- cd ceph/workunits/qa/ceph-disk
- sudo bash
- bash ceph-disk.sh
- Control-c when it starts to run the tests

Although the problem shows when running the tests, it is easier to reproduce as follows:

ceph version 10.0.0-855-g15a81bb (15a81bb7121799ba1b71b88b356998ebc8effec9)

```
[root@target167114226249 ceph-disk]# uuid=$(uuidgen) ; ceph-disk prepare --osd-uuid $uuid /dev/vdd
[root@target167114226249 ceph-disk]# id=$(ceph osd create $uuid)
[root@target167114226249 ceph-disk]# echo $id
4
[root@target167114226249 ceph-disk]# ceph osd tree
ID WEIGHT  TYPE NAME                  UP/DOWN REWEIGHT PRIMARY-AFFINITY
...
 4 0.00969         osd.4                    up  1.00000          1.00000
[root@target167114226249 ceph-disk]# ceph-disk deactivate --deactivate-by-id $id ; ceph-disk destr
oy --zap --destroy-by-id $id
[root@target167114226249 ceph-disk]# ceph osd tree
ID WEIGHT  TYPE NAME                  UP/DOWN REWEIGHT PRIMARY-AFFINITY
-1 0.01938 root default
-3       0     rack localrack
-2       0         host localhost
-4 0.01938     host target167114226249
 2 0.00969         osd.2                  down  1.00000          1.00000
 3 0.00969         osd.3                  down  1.00000          1.00000
[root@target167114226249 ceph-disk]# ceph-disk list /dev/vdd
/dev/vdd other, unknown
[root@target167114226249 ceph-disk]# ceph-disk prepare --osd-uuid $uuid /dev/vdd
```

```
[root@target167114226249 ceph-disk]# sleep 300 ; ceph osd tree
...
 4       0 osd.4                             down  1.00000          1.00000
[root@target167114226249 ceph-disk]#
```

Related issues:

Related to Ceph - Bug #13989: OSD boot fails with os/FileJournal.cc: 1907: FA...	**Duplicate**	**12/05/2015**
Related to Ceph - Bug #19119: pre-jewel "osd rm" incrementals are misinterpreted	**Resolved**	**03/01/2017**
Blocks Ceph - Bug #14080: ceph-disk: use blkid instead of sgdisk -i	**Resolved**	**12/14/2015**
Blocks Ceph - Bug #13970: ceph-disk list fails on /dev/cciss!c0d0	**Resolved**	**12/03/2015**

Associated revisions

Revision c6cdc334 - 12/10/2015 07:46 PM - Loic Dachary

tests: verify it is possible to reuse an OSD id

When an OSD id is removed via ceph osd rm, it will be reused by the next ceph osd create command. Verify that and OSD reusing such an id successfully comes up.

http://tracker.ceph.com/issues/13988 Refs: #13988

Signed-off-by: Loic Dachary <loic@dachary.org>

Revision 4e28f9e6 - 12/11/2015 04:50 PM - Sage Weil

osd/OSDMap: clear osd_info, osd_xinfo on osd deletion

If we destroy an OSD in the map, clear not just the uuid but also all the metadata about it.

Specifically, we care about up_from, which can prevent a new OSD from booting if it starts with a map prior to the deletion when it sends it boot. Specifically, the osd epoch may be 0 and if the latest osd epoch is also small the osd decide it is "close enough" to the latest epoch and sends the boot message. In practice this problem wouldn't surface on any cluster that isn't brand new.

Note that this changes the result of applying an incremental. As such, it will cause lots of old OSDs to request full maps from the mon, spiking load during an upgrade. This is as it should be.

Fixes: #13988
Signed-off-by: Sage Weil <sage@redhat.com>

Revision e102e5a0 - 01/11/2016 03:58 PM - Loic Dachary

tests: verify it is possible to reuse an OSD id

When an OSD id is removed via ceph osd rm, it will be reused by the next ceph osd create command. Verify that and OSD reusing such an id successfully comes up.

http://tracker.ceph.com/issues/13988 Refs: #13988

Signed-off-by: Loic Dachary <loic@dachary.org>
(cherry picked from commit 7324615bdb829f77928fa10d4e988c6422945937)

History

#1 - 12/05/2015 04:46 PM - Loic Dachary

- Related to Bug #13989: OSD boot fails with os/FileJournal.cc: 1907: FAILED assert(0) added

#2 - 12/08/2015 03:35 PM - Loic Dachary

- File test.out.gz added

#3 - 12/08/2015 03:36 PM - Loic Dachary

- Duplicated by Bug #13986: TEST_scrub_snaps: timeout > 300 added

#4 - 12/08/2015 03:47 PM - Loic Dachary

- Subject changed from new OSD re-using old OSD id fails to boot to new OSD fails to boot

#5 - 12/08/2015 04:56 PM - Loic Dachary

- File deleted (test.out.gz)

#6 - 12/08/2015 04:56 PM - Loic Dachary

- Duplicated by deleted (Bug #13986: TEST_scrub_snaps: timeout > 300)

#7 - 12/08/2015 11:55 PM - Loic Dachary

- Subject changed from new OSD fails to boot to new OSD re-using old OSD id fails to boot

#8 - 12/09/2015 01:52 PM - Loic Dachary

- Description updated

#9 - 12/09/2015 04:04 PM - Loic Dachary

- File ceph-mon.a.log.gz added

- File ceph-osd.2.log.gz added

#10 - 12/09/2015 04:09 PM - Loic Dachary

The attached files were obtained with a slightly different set of commands:

```
root@target167114253183:~/ceph/qa/workunits/ceph-disk# ceph-disk prepare /dev/vdb
root@target167114253183:~/ceph/qa/workunits/ceph-disk# id=2
root@target167114253183:~/ceph/qa/workunits/ceph-disk# ceph-disk deactivate --deactivate-by-id $id ; ceph-disk
 destroy --zap --destroy-by-id $id
root@target167114253183:~/ceph/qa/workunits/ceph-disk# ceph-disk prepare /dev/vdb
```

#11 - 12/09/2015 05:16 PM - Loic Dachary

- File osdmap.15.plain added

- File osdmap.16.plain added

- File osdmap.17.plain added

#12 - 12/09/2015 05:34 PM - Loic Dachary

Removed all OSDs (0, 1, 2) and then ceph-prepare /dev/vdb which created the osd.0 and shows the same symptoms. The osdmap at this point is:

```
epoch 31
fsid 8441dd6a-940e-4e1b-a209-a155dcaf7d9b
created 2015-12-09 14:06:12.811944
modified 2015-12-09 17:28:45.500737
flags sortbitwise

pool 0 'rbd' replicated size 2 min_size 1 crush_ruleset 0 object_hash rjenkins pg_num 8 pgp_num 8 last_change
1 flags hashpspool stripe_width 0

max_osd 3
osd.0 down in  weight 1 up_from 3 up_thru 3 down_at 25 last_clean_interval [0,0) 167.114.253.182:6800/9101 167
.114.253.182:6801/9101 167.114.253.182:6802/9101 167.114.253.182:6803/9101 e\
xists,new d73314b6-2633-4732-921a-9be60bf850f3
```

#13 - 12/09/2015 08:34 PM - Loic Dachary

To repeat the problem :

- on master https://github.com/ceph/ceph/pull/6876
- on infernalis https://github.com/ceph/ceph/pull/6882

#14 - 12/09/2015 08:34 PM - Loic Dachary

- Status changed from Verified to In Progress

#15 - 12/10/2015 12:00 PM - Wei-Chung Cheng

Hi loic,

I try to trace this problem.
I think the problem is on FileJournal::check()

Due to the same uuid, it will pass **header.fsid != fsid**.
It will let us to use the old header information. (Like old committed_up_to)

So It will get the assertion.

I just add another condition to check not only fsid but also journaled_seq.
It looks like work.
I will test on my environment some times. If it runs Ok, I will pull another request to fixed.

If you have any problems, feel free to let me know.

thanks!
vicente

#16 - 12/11/2015 07:42 AM - Loic Dachary

```
bash-4.2$ git bisect run $(pwd)/try.sh
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0.log
Bisecting: 713 revisions left to test after this (roughly 10 steps)
[118fc222074b2730a9d653d3a96349ddaf326833] Merge pull request #6583 from ceph/wip-krbd-map-args
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-715-g118fc22.log
Bisecting: 356 revisions left to test after this (roughly 9 steps)
[37dd1f00e2f367926935b9dc05be80c96d796600] Merge pull request #6454 from H3C/wip-mds
running /home/loic/ceph-centos-7-loic/try.sh
v10.0.0-617-g37dd1f0.log
Bisecting: 177 revisions left to test after this (roughly 8 steps)
[b584388ce9ce998c99e219ec144725beaf09ab28] Merge pull request #6489 from xiexingguo/xxg-wip-13715
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-895-gb584388.log
Bisecting: 85 revisions left to test after this (roughly 7 steps)
[5135292d9557269bab5cefc98d39606174aa6ebe] Merge branch 'wip-bigbang'
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-987-g5135292.log
Bisecting: 45 revisions left to test after this (roughly 6 steps)
[f3e88ace74c896c72f6e8485c44c7432f298d887] Merge remote-tracking branch 'gh/jewel'
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-907-gf3e88ac.log
Bisecting: 39 revisions left to test after this (roughly 5 steps)
[083c2e42c663229ce505f74c40d8261ca530a79b] Merge pull request #6565 from chenji-kael/patch-1
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-920-g083c2e4.log
There are only 'skip'ped commits left to test.
The first bad commit could be any of:
536c70281a8952358e8d88a6ff8d7cd9b8db5a76
fb9dfada02e61928b3b63e0c2794c1885021180f
d781f48438c896c9a4b636a9772420bd718db90d
e5fc790329e5209acd5218bf78ab3aa704f8a1e0
6f30002485009bca046b70817190924f778e6ba5
73bdf0fc044dbdc2780dccb6d9b201256817c0d6
a12dd1b61274ef7c6f63ab601f1318d5bb8e13d6
ae9d5ee65c1a0cd2444671bddfb83fd3a2667552
26496b907758f0f4ac5b33bbc74795ab65227693
7489ec484908188964a66220425df575296b989a
39c1495406cce3f24c427a3d7a53fb7b26a509c9
6557b76f845d8f235259a8f23c4d3dbcd228343f
05aaa60eb53557d157f587da97580e74edecde61
e31b69514a8aafddd0f4e50482d185ddcaa11d5e
39e06ef8f070e136e54452bdea3f6105cd79bb73
8b5b6c85cc6225d961debfb98621927fc50e81b5
242bf504f1c1e1924ef3a4ac74407c0aa120418e
7bc4763ed720705c549bcc24d6d6c24878562a5f
21ca0b591aa495586f771b11a2ad2f5d9b920dfb
b3b0a95e43a0cca255c8cb6ed0ca1125b05d7c60
2754007c4b4036c5a42aee625f4abe6d8947fe34
d4f813b37576992803c950da0faf0c98d64e9561
2af422a5af8d9295167d48e6451b4a87162d4488
dd91837a8e945a6088a9aab899971ea2b90303e5
605e188003699c7cb4923acf5634a4b45af2dd5a
d3eba9b0afe6cf2085ee5704ce7d1eb03c580ad9
21e95c2dac9bceb70619ccfc6e466f4626211112
c1f6eec94b5788d88fb12fca3b824f8c90768758
6cbdd6750cf330047d52817b9ee9af31a7d318ae
894eb2af5b3729dfe9f7492f9ab6d8c2042f57ae
0938bf055deec0183e1159f9bfcfa98a6757bc4b
c9534dfb15a595387f2daf2795fa03e6330efb7b
865ddcac41069f7857a3066b3b04a5c62cc1db8c
c85b15234afb76eb66943f981232ecc78f3b1a4a
7fcffe3d9f748c8f3addf57fdfa6a6bf2fdeff66
```

```
b3ca828ae8ebc9068073494c46faf3e8e1443ada
160a0205c1a74ab594c5914cda54c85af6832900
5e10de4cc02e497a92c7359315c30c30bbe2ebc7
0389763a5b6777ca16b4dec9b13ac4efc0b0973f
12c7e54fd9255a0a05796f0a9163e3d6177c83f5
19b714f519123a014dc81634e13b89694b507a9d
59123824a9be6d27a451fe6995377b18641a9408
23d4df3e012374c1d9520161b1e9829d2c464add
d5a2f9a6c7e1eec2bae7facc670860729bc06408
2d2e6b2ea89233f97d92fdf757297861ba621b44
3ad0c9215f667732eea45034a7e81de2cd33eb40
1f4b7141c5a381a0da759bd5773501f1fbaaa078
c131c81511ca766f96238b1942e8fad566bd7413
093478afa96bf0c99e2ce4ee98e4ab04bdccc54c
9864a79abcc977b7254cfb40814f19ec69d6bc3a
ca75e37a302b6ce48db9f776453a881089f0c82e
e2756f9ab3c394a98c078c42cb6aa9b085b90181
57121dbe2cbbbd188c9aff38a3a534df038982c9
ae1cae027df1ecf3c66dc13def87c95f7abd8207
53f2c7f291d94774dda7182d00fd26af4ee65f6f
75e28c425422bb2bcb98272eff1a23766ebbf881
0269a0c17723fd3e22738f7495fe017225b924a4
17d24292b8121d5d13ddd27179342ef99b9de895
d201c6d93f40affe72d940605c8786247451d3e5
facd36fc14bb855e950fd55a764444f8887e0828
25888bb7f5be5825c3a0d8a6fec3a459e1d678ef
987f68a8df292668ad241f4769d82792644454dd
56dbf7a63cb39999c2eda288e97fcf2b2c778052
72edab282343e8509b387f92d05fc4d6ae96b25b
28138c65a6b02ba0dd0e65105303870bb8d2b86e
9aabc8a9b8d7775337716c4e0fa3cc53938acb45
5135292d9557269bab5cefc98d39606174aa6ebe
We cannot bisect more!
bisect run cannot continue any more
```

#17 - 12/11/2015 07:57 AM - Loic Dachary

@Vicente : I do not see an assertion in http://jenkins.ceph.dachary.org/job/ceph/LABELS=centos-7&&x86_64/10060/consoleText, only the fact that the OSD cannot join. Also, the test osd-reuse-id.sh removes the directory in which the original OSD files are, so I don't think this part of the code is involved. Am I missing something ?

#18 - 12/11/2015 08:10 AM - Loic Dachary

- File l.out added

#19 - 12/11/2015 08:31 AM - Loic Dachary

The bad commit is one of those:

```
# possible first bad commit: [5135292d9557269bab5cefc98d39606174aa6ebe] Merge branch 'wip-bigbang'
# possible first bad commit: [9aabc8a9b8d7775337716c4e0fa3cc53938acb45] test/mon/osd-crush.sh: escape ceph tel
l mon.*
# possible first bad commit: [72edab282343e8509b387f92d05fc4d6ae96b25b] osd: make some of the pg_temp methods/
fields private
# possible first bad commit: [987f68a8df292668ad241f4769d82792644454dd] osdc/Objecter: call notify completion
only once
# possible first bad commit: [d201c6d93f40affe72d940605c8786247451d3e5] mon: change mon_osd_min_down_reporters
 from 1 -> 2
# possible first bad commit: [0269a0c17723fd3e22738f7495fe017225b924a4] mon/OSDMonitor: simplify failure repor
ters vs reports logic
# possible first bad commit: [53f2c7f291d94774dda7182d00fd26af4ee65f6f] osd: simplify pg creation
# possible first bad commit: [57121dbe2cbbbd188c9aff38a3a534df038982c9] mon/MonClient: make _sub_got behave if
 we "got" old stuff
# possible first bad commit: [ca75e37a302b6ce48db9f776453a881089f0c82e] mon/OSDMonitor: fix oldest_map in send
_incremental
# possible first bad commit: [9864a79abcc977b7254cfb40814f19ec69d6bc3a] mon/PGMonitor: avoid useless pg gets w
hen pool is deleted
# possible first bad commit: [1f4b7141c5a381a0da759bd5773501f1fbaaa078] mon/PGMonitor: revamp how pg creates a
re tracked
# possible first bad commit: [3ad0c9215f667732eea45034a7e81de2cd33eb40] mon/PGMonitor: only send pg create mes
sages to up osds
# possible first bad commit: [23d4df3e012374c1d9520161b1e9829d2c464add] mon/PGMonitor: only churn mapping_epoc
h if the primary changes
# possible first bad commit: [59123824a9be6d27a451fe6995377b18641a9408] mon/PGMonitor: a bunch of cosmetic cle
anup
# possible first bad commit: [0389763a5b6777ca16b4dec9b13ac4efc0b0973f] mon/PGMonitor: drop old creating_pgs_b
y_osd
# possible first bad commit: [160a0205c1a74ab594c5914cda54c85af6832900] osd: reduce mon_subscribe messages
# possible first bad commit: [7fcffe3d9f748c8f3addf57fdfa6a6bf2fdeff66] mon/MonClient: only send new subscript
ions
# possible first bad commit: [c85b15234afb76eb66943f981232ecc78f3b1a4a] mon/PGMonitor: send pg creates via per
sistent subscriptions, not spam
# possible first bad commit: [0938bf055deec0183e1159f9bfcfa98a6757bc4b] mon/PGMonitor: only map and send pg cr
eates post paxos update
# possible first bad commit: [6cbdd6750cf330047d52817b9ee9af31a7d318ae] mon/PGMonitor: remove map_pg_creates,
send_pg_creates commands
# possible first bad commit: [c1f6eec94b5788d88fb12fca3b824f8c90768758] messages/MOSDPGCreate: make it more re
adable
# possible first bad commit: [d3eba9b0afe6cf2085ee5704ce7d1eb03c580ad9] osd: subscribe to all pg creates, not
just once on start
# possible first bad commit: [dd91837a8e945a6088a9aab899971ea2b90303e5] mon/PGMonitor: track creating_pgs_by_o
sd_epoch
# possible first bad commit: [2754007c4b4036c5a42aee625f4abe6d8947fe34] mon/PGMap: assert our pg counts don't
go negative
# possible first bad commit: [b3b0a95e43a0cca255c8cb6ed0ca1125b05d7c60] mon/OSDMonitor: do not prime pg_temp f
or creating pgs
# possible first bad commit: [242bf504f1c1e1924ef3a4ac74407c0aa120418e] mon/PGMonitor: note mapping_epoch for
creating pgs
# possible first bad commit: [39e06ef8f070e136e54452bdea3f6105cd79bb73] mon: let peon mons send the osdmap rep
lies
# possible first bad commit: [05aaa60eb53557d157f587da97580e74edecde61] msg/simple/Pipe: show keepalives at le
vel 2
# possible first bad commit: [6557b76f845d8f235259a8f23c4d3dbcd228343f] mon: set mon_subscribe_interval to a d
ay
# possible first bad commit: [26496b907758f0f4ac5b33bbc74795ab65227693] mon: only ack subscriptions (and renew
) if client or mon is old
# possible first bad commit: [ae9d5ee65c1a0cd2444671bddfb83fd3a2667552] mon: remove old subscribe renewal-base
d timeouts
# possible first bad commit: [6f30002485009bca046b70817190924f778e6ba5] mon: small cleanup in _ms_dispatch
# possible first bad commit: [e5fc790329e5209acd5218bf78ab3aa704f8a1e0] mon: new session_timeout mechanism tha
t is not subscribe-based
# possible first bad commit: [536c70281a8952358e8d88a6ff8d7cd9b8db5a76] msg: make last_keepalive[_ack] lock sa
fe
# possible first bad commit: [fb9dfada02e61928b3b63e0c2794c1885021180f] msg: track stamp of last keepalive[2]
received
# possible first bad commit: [d781f48438c896c9a4b636a9772420bd718db90d] common: mirror leveldb default tuning
```

```
w/ rocksdb
# possible first bad commit: [73bdf0fc044dbdc2780dccb6d9b201256817c0d6] mon/MonClient: don't send log if we're
 reconnecting
# possible first bad commit: [a12dd1b61274ef7c6f63ab601f1318d5bb8e13d6] mon: disabled rocksdb compression when
 used as the backend
# possible first bad commit: [7489ec484908188964a66220425df575296b989a] osd: cap adjusted max mon report inter
val at 2/3 of timeout
# possible first bad commit: [39c1495406cce3f24c427a3d7a53fb7b26a509c9] osd: protect mon reporting with mon_re
port_lock
# possible first bad commit: [e31b69514a8aafddd0f4e50482d185ddcaa11d5e] osd: fix reconnect behavior from booti
ng state
# possible first bad commit: [8b5b6c85cc6225d961debfb98621927fc50e81b5] osd: move the monitor report to OSD::t
ick_without_osd_lock
# possible first bad commit: [7bc4763ed720705c549bcc24d6d6c24878562a5f] osd: _got_mon_epochs - refactor the lo
ck scope to avoid a race (which fail make check)
# possible first bad commit: [21ca0b591aa495586f771b11a2ad2f5d9b920dfb] osd: don't send dup subscribes so much
# possible first bad commit: [d4f813b37576992803c950da0faf0c98d64e9561] osd: introduce explicit preboot stage
# possible first bad commit: [2af422a5af8d9295167d48e6451b4a87162d4488] osd: skip osdmap version query if we c
an
# possible first bad commit: [605e188003699c7cb4923acf5634a4b45af2dd5a] osd: make [_]maybe_boot lockless varia
nt
# possible first bad commit: [21e95c2dac9bceb70619ccfc6e466f4626211112] osd: only send boot if booting on getv
ersion completion
# possible first bad commit: [894eb2af5b3729dfe9f7492f9ab6d8c2042f57ae] osd: do not resend pg_temp requests
# possible first bad commit: [c9534dfb15a595387f2daf2795fa03e6330efb7b] osd: do not send dup failure reports
# possible first bad commit: [865ddcac41069f7857a3066b3b04a5c62cc1db8c] osd: resend pending failure reports wi
th a new mon session
# possible first bad commit: [b3ca828ae8ebc9068073494c46faf3e8e1443ada] osd: fix send_failures() locking
# possible first bad commit: [5e10de4cc02e497a92c7359315c30c30bbe2ebc7] osd: backoff the max reporting interva
l, too
# possible first bad commit: [12c7e54fd9255a0a05796f0a9163e3d6177c83f5] osd: no need for regular send_pg_temps
# possible first bad commit: [19b714f519123a014dc81634e13b89694b507a9d] osd: just send alive when it is queue
# possible first bad commit: [d5a2f9a6c7e1eec2bae7facc670860729bc06408] osd: fix pg stat reporting
# possible first bad commit: [2d2e6b2ea89233f97d92fdf757297861ba621b44] osd: inline do_mon_report
# possible first bad commit: [c131c81511ca766f96238b1942e8fad566bd7413] osd: limit nubmer of pg stat updates i
n flight
# possible first bad commit: [093478afa96bf0c99e2ce4ee98e4ab04bdccc54c] osd: fix pg_stats_queue lock protectio
n
# possible first bad commit: [e2756f9ab3c394a98c078c42cb6aa9b085b90181] osd: scale mon report interval with ti
meout backoff
# possible first bad commit: [ae1cae027df1ecf3c66dc13def87c95f7abd8207] osd: keep count of outstanding pg stat
 updates to mon
# possible first bad commit: [75e28c425422bb2bcb98272eff1a23766ebbf881] osd: no stats outstanding when we rese
t the session
# possible first bad commit: [17d24292b8121d5d13ddd27179342ef99b9de895] osd: remove old stats backoff mechanis
m
# possible first bad commit: [facd36fc14bb855e950fd55a764444f8887e0828] osd: exponential backoff on pg stats a
ck timeout
# possible first bad commit: [25888bb7f5be5825c3a0d8a6fec3a459e1d678ef] message/MLog: include seq in print
# possible first bad commit: [56dbf7a63cb39999c2eda288e97fcf2b2c778052] osd/OSDMap: cache values for in, up os
ds
# possible first bad commit: [28138c65a6b02ba0dd0e65105303870bb8d2b86e] mon/PGMonitor: avoid iterating over al
l pgs to find stale
```

#20 - 12/11/2015 09:36 AM - Loic Dachary

```
bash-4.2$ git bisect start # initialize the search
git bisect start # initialize the search
bash-4.2$ git bisect bad 9aabc8a9b8d7775337716c4e0fa3cc53938acb45
git bisect bad 9aabc8a9b8d7775337716c4e0fa3cc53938acb45
bash-4.2$ git bisect good 28138c65a6b02ba0dd0e65105303870bb8d2b86e
git bisect good 28138c65a6b02ba0dd0e65105303870bb8d2b86e
Bisecting: 32 revisions left to test after this (roughly 5 steps)
[fb9dfada02e61928b3b63e0c2794c1885021180f] msg: track stamp of last keepalive[2] received
bash-4.2$ git bisect run $(pwd)/try.sh
git bisect run $(pwd)/try.sh
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-745-gfb9dfad.log
Bisecting: 15 revisions left to test after this (roughly 4 steps)
[865ddcac41069f7857a3066b3b04a5c62cc1db8c] osd: resend pending failure reports with a new mon session
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-729-g865ddca.log
Bisecting: 7 revisions left to test after this (roughly 3 steps)
[7bc4763ed720705c549bcc24d6d6c24878562a5f] osd: _got_mon_epochs - refactor the lock scope to avoid a race (whi
ch fail make check)
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-737-g7bc4763.log
Bisecting: 3 revisions left to test after this (roughly 2 steps)
[605e188003699c7cb4923acf5634a4b45af2dd5a] osd: make [_]maybe_boot lockless variant
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-733-g605e188.log
Bisecting: 1 revision left to test after this (roughly 1 step)
[d4f813b37576992803c950da0faf0c98d64e9561] osd: introduce explicit preboot stage
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-735-gd4f813b.log
Bisecting: 0 revisions left to test after this (roughly 0 steps)
[2af422a5af8d9295167d48e6451b4a87162d4488] osd: skip osdmap version query if we can
running /home/loic/ceph-centos-7-loic/try.sh
v9.2.0-734-g2af422a.log
d4f813b37576992803c950da0faf0c98d64e9561 is the first bad commit
commit d4f813b37576992803c950da0faf0c98d64e9561
Author: Sage Weil <sage@redhat.com>
Date:   Wed Sep 23 17:58:15 2015 -0400

    osd: introduce explicit preboot stage

    We want to separate the stage where we do a bunch of work
    prior to booting (but intend to eventually boot), like when we
    get maps and wait to be healthy, from the point after we've sent
    the boot message while we are just waiting for a response (so that
    we can avoid resending that boot message needlessly).

    - start at PREBOOT in start_boot()
    - transition to BOOTING in _send_boot()
    - only call _preboot() while in PREBOOT state

    Signed-off-by: Sage Weil <sage@redhat.com>

:040000 040000 4c9a7682b8ca2ec94396cc32e95922f244e5b4d5 330149ade7872bc05cc9c8f171e16ca61f51aff6 M      src
bisect run success
bash-4.2$ cat try.sh
cat try.sh
#!/bin/bash
cd src
log=$(git describe)
echo $log.log
make -j12 ceph-mon ceph-osd >& $log.log
bash ../osd-reuse-id.sh > $log.out 2>&1
```

#21 - 12/11/2015 10:07 AM - Wei-Chung Cheng

Loic Dachary wrote:

> @Vicente : I do not see an assertion in http://jenkins.ceph.dachary.org/job/ceph/LABELS=centos-7&&x86_64/10060/consoleText, only the fact that the OSD cannot join. Also, the test osd-reuse-id.sh removes the directory in which the original OSD files are, so I don't think this part of the code is involved. Am I missing something ?

Hi Loic,

Due to issue #13989, I follow your step, use the same uuid and id to recreate osd.
I will hit the assertion described on #13989.

I try your repo with **dachary:wip-13988-reuse-osd-id-infernalis** and above method of repeation to confirm assertion.
And try to modify the check condition to avoid these situation.
It seems work.

When I run **run-make-check.sh**, it could not hit the assertion.
I think the reason is that we would not use the same uuid to create osd. right?

For the problem the osd could not be `up` state you hit, I could not reproduce everytime.
So I try to resolve the #13989? Or Try to figure this(#13988) problems?

thanks
vicente

#22 - 12/11/2015 11:36 AM - Wei-Chung Cheng

Hi Loic,

I also fail on the master branch with resue-osd-id test cases.
(but not always fail on infernalis)

Maybe #13988 do not have the same root cause with #13989?

thanks!
vicente

#23 - 12/11/2015 03:43 PM - Loic Dachary

Before removal:

```
epoch 15
fsid 8441dd6a-940e-4e1b-a209-a155dcaf7d9b
created 2015-12-09 14:06:12.811944
modified 2015-12-09 16:00:20.432352
flags sortbitwise

pool 0 'rbd' replicated size 2 min_size 1 crush_ruleset 0 object_hash rjenkins pg_num 8 pgp_num 8 last_change
1 flags hashpspool stripe_width 0

max_osd 3
osd.0 up   in  weight 1 up_from 3 up_thru 3 down_at 0 last_clean_interval [0,0) 167.114.253.182:6800/9101 167.
114.253.182:6801/9101 167.114.253.182:6802/9101 167.114.253.182:6803/9101 exists,up c2210d09-3b33-404e-863e-86
4d3eb5048d
osd.1 up   in  weight 1 up_from 3 up_thru 13 down_at 0 last_clean_interval [0,0) 167.114.253.182:6804/9102 167
.114.253.182:6805/9102 167.114.253.182:6806/9102 167.114.253.182:6807/9102 exists,up e755450f-da23-489a-97de-6
23791205759
osd.2 down in  weight 1 up_from 10 up_thru 11 down_at 13 last_clean_interval [0,0) 167.114.253.183:6800/11225
167.114.253.183:6801/11225 167.114.253.183:6802/11225 167.114.253.183:6803/11225 exists 8800c444-1a12-4cc5-a60
1-48fb583545a5

pg_temp 0.0 [2,1]
pg_temp 0.1 [2,1]
pg_temp 0.2 [2,1]
pg_temp 0.3 [2,1]
pg_temp 0.4 [2,1]
pg_temp 0.5 [2,1]
pg_temp 0.6 [2,1]
pg_temp 0.7 [2,1]
```

after removal

```
epoch 16
fsid 8441dd6a-940e-4e1b-a209-a155dcaf7d9b
created 2015-12-09 14:06:12.811944
modified 2015-12-09 16:00:21.699903
flags sortbitwise

pool 0 'rbd' replicated size 2 min_size 1 crush_ruleset 0 object_hash rjenkins pg_num 8 pgp_num 8 last_change
1 flags hashpspool stripe_width 0

max_osd 3
osd.0 up   in  weight 1 up_from 3 up_thru 3 down_at 0 last_clean_interval [0,0) 167.114.253.182:6800/9101 167.
114.253.182:6801/9101 167.114.253.182:6802/9101 167.114.253.182:6803/9101 exists,up c2210d09-3b33-404e-863e-86
4d3eb5048d
osd.1 up   in  weight 1 up_from 3 up_thru 13 down_at 0 last_clean_interval [0,0) 167.114.253.182:6804/9102 167
.114.253.182:6805/9102 167.114.253.182:6806/9102 167.114.253.182:6807/9102 exists,up e755450f-da23-489a-97de-6
23791205759

pg_temp 0.0 [2,1]
pg_temp 0.1 [2,1]
pg_temp 0.2 [2,1]
pg_temp 0.3 [2,1]
pg_temp 0.4 [2,1]
pg_temp 0.5 [2,1]
pg_temp 0.6 [2,1]
pg_temp 0.7 [2,1]
```

after creating an osd

```
epoch 17
fsid 8441dd6a-940e-4e1b-a209-a155dcaf7d9b
created 2015-12-09 14:06:12.811944
modified 2015-12-09 16:01:13.189546
flags sortbitwise

pool 0 'rbd' replicated size 2 min_size 1 crush_ruleset 0 object_hash rjenkins pg_num 8 pgp_num 8 last_change
```

```
1 flags hashpspool stripe_width 0

max_osd 3
osd.0 up   in  weight 1 up_from 3 up_thru 3 down_at 0 last_clean_interval [0,0) 167.114.253.182:6800/9101 167.
114.253.182:6801/9101 167.114.253.182:6802/9101 167.114.253.182:6803/9101 exists,up c2210d09-3b33-404e-863e-86
4d3eb5048d
osd.1 up   in  weight 1 up_from 3 up_thru 13 down_at 0 last_clean_interval [0,0) 167.114.253.182:6804/9102 167
.114.253.182:6805/9102 167.114.253.182:6806/9102 167.114.253.182:6807/9102 exists,up e755450f-da23-489a-97de-6
23791205759
osd.2 down in  weight 1 up_from 10 up_thru 11 down_at 13 last_clean_interval [0,0) 167.114.253.183:6800/11225
167.114.253.183:6801/11225 167.114.253.183:6802/11225 167.114.253.183:6803/11225 exists,new 5e99a402-9b2a-4032
-bd0e-70c4d5d0f4d1

pg_temp 0.0 [2,1]
pg_temp 0.1 [2,1]
pg_temp 0.2 [2,1]
pg_temp 0.3 [2,1]
pg_temp 0.4 [2,1]
pg_temp 0.5 [2,1]
pg_temp 0.6 [2,1]
pg_temp 0.7 [2,1]
```

#24 - 12/11/2015 05:00 PM - Loic Dachary

https://github.com/ceph/ceph/pull/6900

#25 - 12/11/2015 05:01 PM - Loic Dachary

- Status changed from In Progress to Testing

#26 - 12/15/2015 12:56 PM - Loic Dachary

- Blocks Bug #14080: ceph-disk: use blkid instead of sgdisk -i added

#27 - 12/15/2015 12:56 PM - Loic Dachary

- Blocks Bug #13970: ceph-disk list fails on /dev/cciss!c0d0 added

#28 - 12/18/2015 08:08 PM - Sage Weil

- Status changed from Testing to Resolved

#29 - 03/01/2017 04:24 PM - Ilya Dryomov

- Related to Bug #19119: pre-jewel "osd rm" incrementals are misinterpreted added

Files

ceph-mon.a.log.gz	249 KB	12/09/2015	Loic Dachary

ceph-osd.2.log.gz	87.2 KB	12/09/2015	Loic Dachary
osdmap.15.plain	1.11 KB	12/09/2015	Loic Dachary
osdmap.16.plain	904 Bytes	12/09/2015	Loic Dachary
osdmap.17.plain	1.12 KB	12/09/2015	Loic Dachary
l.out	175 KB	12/11/2015	Loic Dachary