Ceph - Bug #11332

mon failed to read inc osdmap

04/06/2015 04:31 PM - Yuri Weinstein

Status:	Resolved	**Start date:**	04/06/2015
Priority:	Urgent	**Due date:**	
Assignee:		**% Done:**	0%
Category:	Monitor	**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:	Q/A	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:	luminous	**ceph-qa-suite:**	upgrade/giant-x
Regression:	No	**Pull request ID:**	
Severity:	3 - minor		

Description

Run: http://pulpito.ceph.com/teuthology-2015-04-01_17:05:02-upgrade:giant-x-hammer-distro-basic-vps/
Job: ['832492']
Logs: http://qa-proxy.ceph.com/teuthology/teuthology-2015-04-01_17:05:02-upgrade:giant-x-hammer-distro-basic-vps/832492/

```
Assertion: mon/OSDMonitor.cc: 2097: FAILED assert(0)
ceph version 0.93-217-g28787d2 (28787d2184754a2988a26abb23eb3fa91d1d46fd)
 1: (OSDMonitor::build_incremental(unsigned int, unsigned int)+0xb06) [0x60c436]
 2: (OSDMonitor::send_incremental(unsigned int, MonSession*, bool)+0x97) [0x60c647]
 3: (OSDMonitor::check_sub(Subscription*)+0xef) [0x60d09f]
 4: (Monitor::handle_subscribe(MMonSubscribe*)+0x797) [0x57d217]
 5: (Monitor::dispatch(MonSession*, Message*, bool)+0x39b) [0x5a08fb]
 6: (Monitor::_ms_dispatch(Message*)+0x1f6) [0x5a1096]
 7: (Monitor::ms_dispatch(Message*)+0x32) [0x5be012]
 8: (DispatchQueue::entry()+0x4fa) [0x7df52a]
 9: (DispatchQueue::DispatchThread::entry()+0xd) [0x7d03cd]
 10: (()+0x7851) [0x7f2e132fe851]
 11: (clone()+0x6d) [0x7f2e1208890d]
```

Related issues:

Related to Ceph - Bug #11267: 2015-03-27T19:11:41.127 INFO:tasks.rados.rados....	**Resolved**	**03/29/2015**
Related to Ceph - Bug #11428: "FAILED assert(0)" in rados-giant-distro-basic-...	**Rejected**	**04/19/2015**
Copied to Ceph - Backport #23626: mon failed to read inc osdmap	**Resolved**	

History

#1 - 06/02/2015 08:42 PM - Sage Weil

- Subject changed from "FAILED assert(0)" in upgrade:giant-x-hammer-distro-basic-vps run to mon failed to read inc osdmap

- Regression set to No

#2 - 06/02/2015 08:42 PM - Sage Weil

- Category set to Monitor

#3 - 06/29/2015 11:33 AM - Kefu Chai

- Assignee set to Kefu Chai

will try to understand this and give it a try

#4 - 06/30/2015 04:10 PM - Kefu Chai

```
 -1125> 2015-04-04 20:13:07.501795 7f2e0cbd3700 10 mon.a@0(leader).osd e2538  min_last_epoch_clean 2387
 -1124> 2015-04-04 20:13:07.501797 7f2e0cbd3700 10 mon.a@0(leader).paxosservice(osdmap 1787..2538) maybe_trim
trimming to 2038, 251 states
 -1123> 2015-04-04 20:13:07.501801 7f2e0cbd3700 10 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim from 1
787 to 2038
 -1122> 2015-04-04 20:13:07.501803 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim 1787
 -1121> 2015-04-04 20:13:07.501874 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim full_1
787
...
  -988> 2015-04-04 20:13:07.513568 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim 1854
  -987> 2015-04-04 20:13:07.515632 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim full_1
854
  -986> 2015-04-04 20:13:07.515671 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim 1855
  -985> 2015-04-04 20:13:07.515717 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim full_1
855
  -984> 2015-04-04 20:13:07.515741 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim 1856
  -983> 2015-04-04 20:13:07.515780 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim full_1
856
...
  -619> 2015-04-04 20:13:11.639359 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim 2037
  -618> 2015-04-04 20:13:11.639375 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538) trim full_2
037
  -617> 2015-04-04 20:13:11.639378 7f2e0cbd3700 20 mon.a@0(leader).paxosservice(osdmap 1787..2538)  compacting
 prefix osdmap
  -616> 2015-04-04 20:13:11.639389 7f2e0cbd3700 10 mon.a@0(leader).osd e2538 encode_trim_extra including full
map for e 2038
...
  -271> 2015-04-04 20:13:11.742236 7f2e0cbd3700 10 mon.a@0(leader).paxosservice(osdmap 1787..2538) propose_pen
ding
...
   -38> 2015-04-04 20:13:11.757384 7f2e0c1d2700 10 mon.a@0(leader).osd e2538 check_sub 0x5451f00 next 1736 (on
etime)
   -37> 2015-04-04 20:13:11.757387 7f2e0c1d2700  5 mon.a@0(leader).osd e2538 send_incremental [1736..2538] to
mds.0 10.214.130.132:6835/7883
   -36> 2015-04-04 20:13:11.763668 7f2e0c1d2700 20 mon.a@0(leader).osd e2538 send_incremental starting with ba
se full 1787 13507 bytes
   -35> 2015-04-04 20:13:11.763685 7f2e0c1d2700  1 -- 10.214.130.132:6789/0 --> 10.214.130.132:6835/7883 -- os
d_map(1787..1787 src has 1787..2538) v3 -- ?+0 0x55c5340 con 0x56948e0
   -34> 2015-04-04 20:13:11.763999 7f2e0c1d2700 10 mon.a@0(leader).osd e2538 build_incremental [1788..1888]
   -33> 2015-04-04 20:13:11.780216 7f2e0c1d2700 20 mon.a@0(leader).osd e2538 build_incremental    inc 1888 709
 bytes
   -32> 2015-04-04 20:13:11.780265 7f2e0c1d2700 20 mon.a@0(leader).osd e2538 build_incremental    inc 1887 190
 bytes
...
    -1> 2015-04-04 20:13:11.791732 7f2e0c1d2700 20 mon.a@0(leader).osd e2538 build_incremental    inc 1856 549
 bytes
     0> 2015-04-04 20:13:13.426753 7f2e0c1d2700 -1 mon/OSDMonitor.cc: In function 'MOSDMap* OSDMonitor::build_
incremental(epoch_t, epoch_t)' thread 7f2e0c1d2700 time 2015-04-04 20:13:11.838080
mon/OSDMonitor.cc: 2097: FAILED assert(0)
```

seems epoch inc 1855 was erased, but somehow epoch inc 1856 was available. they are supposed to be erased in a single transaction! weird.

#5 - 08/20/2015 11:15 AM - Kefu Chai

- Assignee deleted (Kefu Chai)

#6 - 08/25/2015 02:29 PM - Sage Weil

- Status changed from New to Can't reproduce

#7 - 11/23/2017 09:51 AM - Xuehan Xu

Hi, everyone. We also encountered this problem, and we found that this seems to be caused by the lack of mutual exclusion between applying "trim" and handling subscriptions. Since "build_incremental" operations doesn't go through the "PAXOS" procedure, and applying "trim" contains two phases, which are modifying "mondbstore" and updating "cached_first_committed", there could be a chance for "send_incremental" operations to happen between them. What's more, "build_incremental" operations also contain two phases, getting "cached_first_committed" and getting actual incrementals for MonDBStore. So, if "build_incremental" do happens concurrently with applying "trim", it could get an out-dated "cached_first_committed" and try to read a full map whose already trimmed.

Is this right?

#8 - 01/15/2018 03:20 AM - Kefu Chai

- Status changed from Can't reproduce to Need Review

https://github.com/ceph/ceph/pull/19397

#9 - 04/10/2018 01:24 PM - Kefu Chai

- Status changed from Need Review to Pending Backport

- Backport set to luminous

#10 - 04/10/2018 01:26 PM - Kefu Chai

- Copied to Backport #23626: mon failed to read inc osdmap added

#11 - 04/16/2018 08:16 PM - Nathan Cutler

- Status changed from Pending Backport to Resolved