Ceph - Bug #2199

mon: get_bl osdmap_full/9583 No such file or directory

03/21/2012 08:26 PM - Yehuda Sadeh

Status:	Resolved	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:	Greg Farnum		
Category:	Monitor		
Target version:	v0.45		
Source:	Development	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

Happened on congress (afair, off 0.41). One monitor is out for more than a month. Following network outage, both monitors were restarted, however, for some reason they stopped accepting new connections on 6789. Following another restart one of the monitors crashed:

```
2012-03-21 20:26:31.116315 3a9badf8700 mon.beta@1(leader) e1 handle_subscribe mon_subscribe({monma
p=2+,osdmap=9583}) v2
2012-03-21 20:26:31.116321 3a9badf8700 mon.beta@1(leader) e1 check_sub monmap next 2 have 1
2012-03-21 20:26:31.116329 3a9badf8700 mon.beta@1(leader).osd e9585 send_incremental [9583..9585]
to osd.139 [2607:f298:4:2243::6241]:6810/10739
2012-03-21 20:26:31.116333 3a9badf8700 mon.beta@1(leader).osd e9585 build_incremental [9583..9585]
2012-03-21 20:26:31.116345 3a9badf8700 store(/var/ceph/mon) reading at off 0 of 1996
2012-03-21 20:26:31.116356 3a9badf8700 store(/var/ceph/mon) get_bl osdmap/9585 = 1996 bytes
2012-03-21 20:26:31.116364 3a9badf8700 mon.beta@1(leader).osd e9585 build_incremental    inc 9585
1996 bytes
2012-03-21 20:26:31.116379 3a9badf8700 store(/var/ceph/mon) reading at off 0 of 1291
2012-03-21 20:26:31.116390 3a9badf8700 store(/var/ceph/mon) get_bl osdmap/9584 = 1291 bytes
2012-03-21 20:26:31.116398 3a9badf8700 mon.beta@1(leader).osd e9585 build_incremental    inc 9584
1291 bytes
2012-03-21 20:26:31.116419 3a9badf8700 store(/var/ceph/mon) get_bl osdmap/9583 No such file or dir
ectory
2012-03-21 20:26:31.116435 3a9badf8700 store(/var/ceph/mon) get_bl osdmap_full/9583 No such file o
r directory
mon/OSDMonitor.cc: In function 'MOSDMap* OSDMonitor::build_incremental(epoch_t, epoch_t)', in thre
ad '3a9badf8700'
mon/OSDMonitor.cc: 961: FAILED assert(0)
 ceph version 0.40-18-g1299e47 (commit:1299e478f468e6b89e3bcf7fc6a2da4a6b05178d)
 1: (OSDMonitor::build_incremental(unsigned int, unsigned int)+0xa91) [0x4a1581]
 2: (OSDMonitor::send_incremental(unsigned int, entity_inst_t&, bool)+0x84) [0x4a1694]
 3: (OSDMonitor::check_sub(Subscription*)+0x98) [0x4a1d48]
 4: (Monitor::handle_subscribe(MMonSubscribe*)+0x8ba) [0x47f38a]
 5: (Monitor::_ms_dispatch(Message*)+0xda8) [0x480e08]
 6: (Monitor::ms_dispatch(Message*)+0x8e) [0x48d70e]
 7: (SimpleMessenger::dispatch_entry()+0x869) [0x535da9]
 8: (SimpleMessenger::DispatchThread::entry()+0x1c) [0x466d7c]
 9: (()+0x68ba) [0x3a9c67da8ba]
 10: (clone()+0x6d) [0x3a9c506302d]
```

History

#1 - 03/21/2012 08:41 PM - Yehuda Sadeh

kept logs for the failing monitor under /var/log/ceph/2199

#2 - 03/22/2012 10:00 AM - Greg Farnum

My guess/hope is that this is one of the issues solved by the monitor slurp and other fixes since 0.41, but I haven't checked the logs. Let me know if I need to do that...

#3 - 03/22/2012 11:00 AM - Greg Farnum

- Category set to Monitor

- Assignee set to Greg Farnum

#4 - 03/22/2012 10:00 PM - Greg Farnum

- Status changed from New to In Progress

Looks like the problem is that the Monitor got elected leader, and while it collected all the state it didn't write it to disk properly:

```
2012-03-21 17:05:40.148323 387c13e3700 store(/var/ceph/mon) erase_ss osdmap/3418
2012-03-21 17:05:40.148358 387c13e3700 store(/var/ceph/mon) erase_ss osdmap_full/3418
2012-03-21 17:05:40.150286 387c13e3700 store(/var/ceph/mon) set_int osdmap/first_committed = 3419
2012-03-21 17:05:40.195212 387c13e3700 store(/var/ceph/mon) put_bl osdmap/latest = 114349 bytes
2012-03-21 17:05:40.228514 387c13e3700 store(/var/ceph/mon) set_int osdmap/last_consumed = 9583
2012-03-21 17:05:40.245084 387c13e3700 store(/var/ceph/mon) set_int osdmap/first_committed = 9583
2012-03-21 17:05:40.261516 387c13e3700 store(/var/ceph/mon) set_int osdmap/last_committed = 9583
2012-03-21 17:05:40.277923 387c13e3700 store(/var/ceph/mon) exists_bl osdmap/9584
```

Notice that it writes to osdmap/latest, but not to osdmap_full/9583! This appears to be because Paxos::store_state got broken when applying a stash (which prevents the sequence that would normally write out everything to disk in proper order from ever executing).
I'll make a patch for this in the morning.

#5 - 03/23/2012 11:04 AM - Greg Farnum

- Status changed from In Progress to Fix Under Review

- Target version set to v0.45

I believe this is fixed in misc-fixes-for-review commit:e08b489d094efe384c3db639af0be765665bee23. Sage needs to review it (and somebody needs to write a Monitor disk store that isn't completely opaque).

#6 - 03/26/2012 09:59 AM - Greg Farnum

- Status changed from Fix Under Review to In Progress

Sage pointed out the stash data structure isn't necessarily the same as the other stored data structures, so this needs a bit more adjustment.

#7 - 03/26/2012 01:34 PM - Greg Farnum

- Status changed from In Progress to Fix Under Review

Re-pushed misc-fixes-for-review.

#8 - 03/27/2012 02:03 PM - Greg Farnum

- Status changed from Fix Under Review to Resolved

Merged to master in 1814aac17593dee0fa4c774d5b462f277f6698da, reviewed by Sage — even though I forgot to add the tag. :(