fs - Bug #18166

monitor cannot start because of "FAILED assert(info.state == MDSMap::STATE_STANDBY)"

12/07/2016 09:43 AM - guotao Yao

Status:	Resolved	**Start date:**	12/07/2016
Priority:	Normal	**Due date:**	
Assignee:	John Spray	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			
Source:	Community (user)	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:	jewel, kraken	**ceph-qa-suite:**	
Regression:	No	**Component(FS):**	MDSMonitor
Severity:	3 - minor	**Labels (FS):**	

Description

ceph version: v10.2.3
operation system: ubuntu 14.04
linux kernel version: 3.13.0

Description:

I test for cephfs, and I have two mds, when I start by 'start ceph-mds-all', the mds cluster is OK, two mds is OK, one is active, and another is standby. But when I test the option of hot-standby in command, problem has arisen.

I have three nodes:
ceph1: monitor + mds + osd
ceph2: monitor + mds + osd
ceph3: monitor + osd

Procedure:

1. start two mds
2. stop the mds in standby state
3. perform command: “ceph-mds --cluster=ceph -i ceph2 --setuser ceph --setgroup ceph --hot-standby 0”
at this step, mds cannot start, so I do next:
4. restart monitor service
at this time, I found that a monitor(ceph1) is down. When I restart the down monitor, the monitor is up, but, the other monitor(ceph2) is down.

monitor log:

```
-132> 2016-12-07 08:59:38.803492 7f9a7a2f3480 10 mon.ceph3@-1(probing).paxosservice(mdsmap 1..438)
 refresh
  -131> 2016-12-07 08:59:38.803506 7f9a7a2f3480 10 mon.ceph3@-1(probing).mds e0 update_from_paxos
version 438, my e 0
  -130> 2016-12-07 08:59:38.803642 7f9a7a2f3480 10 mon.ceph3@-1(probing).mds e0 update_from_paxos
got 438
  -129> 2016-12-07 08:59:38.803686 7f9a7a2f3480  4 mon.ceph3@-1(probing).mds e438 new map



-128> 2016-12-07 08:59:38.803723 7f9a7a2f3480  0 mon.ceph3@-1(probing).mds e438 print_map
e438
enable_multiple, ever_enabled_multiple: 0,0
compat: compat={},rocompat={},incompat={1=base v0.20,2=client writeable ranges,3=default file layo
uts on dirs,4=dir inode in separate object,5=mds uses versioned encoding,6=dirfrag is stored in om
ap,8=file layout v2}
```

Filesystem 'first_cephfs' (1)
fs_name first_cephfs
epoch 436
flags 8
created 2016-11-16 06:06:34.376616
modified 2016-11-23 11:37:26.768424
tableserver 0
root 0
session_timeout 60
session_autoclose 300
max_file_size 1099511627776
last_failure 0
last_failure_osd_epoch 68
compat compat={},rocompat={},incompat={1=base v0.20,2=client writeable ranges,3=default file layouts on dirs,4=dir inode in separate object,5=mds uses versioned encoding,6=dirfrag is stored in omap,8=file layout v2}
max_mds 1
in 0
up {0=58523}
failed
damaged
stopped
data_pools 1
metadata_pool 2
inline_data disabled
58523: 10.10.38.40:6806/13750 'ceph1' mds.0.433 up:active seq 70

Standby daemons:

39073: 10.10.38.41:6805/12990 'ceph2' mds.-1.0 up:standby seq 1 (standby for rank 0)

```
-8> 2016-12-07 08:59:38.825925 7f9a7a2f3480 10 mon.ceph3@0(leader).paxosservice(mdsmap 1..438) ele
ction_finished
    -7> 2016-12-07 08:59:38.825926 7f9a7a2f3480 10 mon.ceph3@0(leader).paxosservice(mdsmap 1..438)
 _active
    -6> 2016-12-07 08:59:38.825928 7f9a7a2f3480  7 mon.ceph3@0(leader).paxosservice(mdsmap 1..438)
 _active creating new pending
    -5> 2016-12-07 08:59:38.825939 7f9a7a2f3480 10 mon.ceph3@0(leader).mds e438 create_pending e43
9
    -4> 2016-12-07 08:59:38.825943 7f9a7a2f3480 10 mon.ceph3@0(leader).mds e438 e438: 1/1/1 up {0=
ceph1=up:active}, 1 up:standby
    -3> 2016-12-07 08:59:38.825960 7f9a7a2f3480 20 mon.ceph3@0(leader).mds e438 gid 39073 is stand
by and following nobody
    -2> 2016-12-07 08:59:38.825966 7f9a7a2f3480 10 mon.ceph3@0(leader).mds e438   setting to follo
w mds rank 0
    -1> 2016-12-07 08:59:38.826000 7f9a6ffff700  5 asok(0x7f9a839eb480) AdminSocket: request 'get_
command_descriptions' '' to 0x7f9a83a44270 returned 2165 bytes
     0> 2016-12-07 08:59:38.828473 7f9a7a2f3480 -1 mon/MDSMonitor.cc: In function 'bool MDSMonitor
::maybe_promote_standby(std::shared_ptr&lt;Filesystem&gt;)' thread 7f9a7a2f3480 time 2016-12-07 08
:59:38.825971
mon/MDSMonitor.cc: 2797: FAILED assert(info.state == MDSMap::STATE_STANDBY)



ceph version 10.2.3 (ecc23778eb545d8dd55e2e4735b53cc93f92e65b)
 1: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0x8b) [0x7f9a79e329fb]
 2: (MDSMonitor::maybe_promote_standby(std::shared_ptr&lt;Filesystem&gt;)+0x976) [0x7f9a79b73016]
 3: (MDSMonitor::tick()+0x397) [0x7f9a79b76fa7]
 4: (MDSMonitor::on_active()+0x15) [0x7f9a79b6d0a5]
 5: (PaxosService::_active()+0x1df) [0x7f9a79af92ef]
 6: (PaxosService::election_finished()+0x67) [0x7f9a79af9a47]
 7: (Monitor::win_election(unsigned int, std::set&lt;int, std::less&lt;int&gt;, std::allocator&lt;
int&gt; >&, unsigned long, MonCommand const*, int, std::set&lt;int, std::less&lt;int&gt;, std::all
ocator&lt;int&gt; > const*)+0x236) [0x7f9a79abb336]
 8: (Monitor::win_standalone_election()+0x158) [0x7f9a79abb738]
 9: (Monitor::bootstrap()+0xa03) [0x7f9a79abc283]
 10: (Monitor::init()+0x190) [0x7f9a79abc590]
 11: (main()+0x24ca) [0x7f9a79a32a4a]
```

```
 12: (__libc_start_main()+0xf5) [0x7f9a76ff6ec5]
 13: (()+0x2609ca) [0x7f9a79a849ca]
 NOTE: a copy of the executable, or `objdump -rdS &lt;executable&gt;` is needed to interpret this.
```

Related issues:

Copied to fs - Backport #18282: jewel: monitor cannot start because of "FAILE...	**Resolved**
Copied to fs - Backport #18283: kraken: monitor cannot start because of "FAIL...	**Closed**

History

#1 - 12/07/2016 05:01 PM - Greg Farnum

- Project changed from Ceph to fs

- Component(FS) MDSMonitor added

So this cluster is freshly-created with version 10.2.3?
Can you upload the monitor log with ceph-post-file? (Preferably one with "debug mon = 20" set.)

Maybe the assert is just bad because the MDS was in standby but got marked as failed, but I'd like to see the log to make sure.

#2 - 12/08/2016 06:01 AM - guotao Yao

- File ceph-mon.ceph3.log added

The attachment is the log of crash monitor.
Thanks!

#3 - 12/08/2016 04:23 PM - John Spray

- Status changed from New to Verified

It looks like MDSMonitor::maybe_promote_standby is iterating over pending_fsmap.standby_daemons, but inside the loop calling try_standby_replay, which modifies standby_daemons (via assign_standby_replay).

#4 - 12/08/2016 05:38 PM - John Spray

- Status changed from Verified to Need Review

- Assignee set to John Spray

- Backport set to jewel

https://github.com/ceph/ceph/pull/12395

#5 - 12/14/2016 12:56 PM - John Spray

- Status changed from Need Review to Pending Backport

- Backport changed from jewel to jewel kraken

#6 - 12/16/2016 02:42 PM - Nathan Cutler

- Copied to Backport #18282: jewel: monitor cannot start because of "FAILED assert(info.state == MDSMap::STATE_STANDBY)" added

#7 - 12/16/2016 02:42 PM - Nathan Cutler

- Copied to Backport #18283: kraken: monitor cannot start because of "FAILED assert(info.state == MDSMap::STATE_STANDBY)" added

#8 - 01/31/2017 01:13 PM - Nathan Cutler

- Backport changed from jewel kraken to jewel, kraken

#9 - 04/14/2017 09:39 PM - Nathan Cutler

kraken backport is unnecessary (fix already in v11.2.0)

#10 - 04/14/2017 09:39 PM - Nathan Cutler

- Status changed from Pending Backport to Resolved

Files

ceph-mon.ceph3.log	105 KB	12/08/2016	guotao Yao