fs - Bug #23519

mds: mds got laggy because of MDSBeacon stuck in mqueue

03/30/2018 03:15 AM - dongdong tao

Status:	Resolved	**Start date:**	03/30/2018
Priority:	Normal	**Due date:**	
Assignee:	dongdong tao	**% Done:**	0%
Category:	Correctness/Safety	**Estimated time:**	0.00 hour
Target version:	v14.0.0		
Source:	Community (dev)	**Affected Versions:**	
Tags:		**ceph-qa-suite:**	
Backport:	mimic,luminous	**Component(FS):**	MDS, MDSMonitor
Regression:	No	**Labels (FS):**	
Severity:	2 - major	**Pull request ID:**	
Reviewed:			

Description

the MDSBeacon message from monitor may stuck for long time in mqueue.
because DispatcherQueue is currently dispatching a MDSMap(rejoin) message, most of the time is spend in process_imported_caps during rejoin.

Below is the log, see the first line and last line.

```
2018-03-29 12:51:13.600301 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d0b5e6080 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 29 v8293) v7
2018-03-29 12:51:13.600344 7fc0c8b06700  1 mds.0.8291 handle_mds_map i am now mds.0.8291
2018-03-29 12:51:13.600352 7fc0c8b06700  1 mds.0.8291 handle_mds_map state change up:reconnect -->
 up:rejoin
2018-03-29 12:51:13.600362 7fc0c8b06700  1 mds.0.8291 rejoin_start
2018-03-29 12:51:17.020712 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d0b5e63c0 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 30 v8293) v7
2018-03-29 12:51:18.420713 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd mdsmap enqueue
0x557d266d58c0 mdsmap(e 8295) v1
2018-03-29 12:51:21.041255 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557cfc186340 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 31 v8295) v7
2018-03-29 12:51:25.005205 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6acd00 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 32 v8295) v7
2018-03-29 12:51:29.024516 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6ad040 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 33 v8295) v7
2018-03-29 12:51:33.032054 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6ad380 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 34 v8295) v7
2018-03-29 12:51:37.025607 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6ad6c0 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 35 v8295) v7
2018-03-29 12:51:41.025065 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6ada00 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 36 v8295) v7
2018-03-29 12:51:45.024202 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6add40 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 37 v8295) v7
2018-03-29 12:51:49.137766 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
```

```
0x557d2d6ae080 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 38 v8295) v7
2018-03-29 12:51:53.141025 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6ae3c0 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 39 v8295) v7
2018-03-29 12:51:57.024013 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6ae700 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 40 v8295) v7
2018-03-29 12:52:01.031026 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6aea40 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 41 v8295) v7
2018-03-29 12:52:05.027415 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6aed80 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 42 v8295) v7
2018-03-29 12:52:09.129624 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6af0c0 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 43 v8295) v7
2018-03-29 12:52:13.124805 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6af400 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 44 v8295) v7
2018-03-29 12:52:17.039225 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6af740 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 45 v8295) v7
2018-03-29 12:52:21.025163 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d2d6afa80 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 46 v8295) v7
2018-03-29 12:52:25.025566 7fc0cb280700  0 -- 10.19.248.31:6804/731296555 >> 10.19.248.31:6789/0 c
onn(0x557cfc0dc800 :-1 s=STATE_OPEN pgs=7475863 cs=1 l=1).taodd beacon enqueue
0x557d87ba0080 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 47 v8295) v7
2018-03-29 12:52:25.337149 7fc0c8b06700  1 mds.0.8291 rejoin_joint_start
2018-03-29 12:52:25.337199 7fc0c8b06700  0 -- 10.19.248.31:6804/731296555 taodd outqueue beacon 0x
557d0b5e6080 mdsbeacon(31079818/enn-yc-31 up:rejoin seq 29 v8293) v7
```

Related issues:

Related to fs - Bug #19706: Laggy mon daemons causing MDS failover (symptom: ...	**Can't reproduce** 04/20/2017
Copied to fs - Backport #26923: mimic: mds: mds got laggy because of MDSBeaco...	**Resolved**
Copied to fs - Backport #26924: luminous: mds: mds got laggy because of MDSBe...	**Resolved**

History

#1 - 03/30/2018 03:16 AM - dongdong tao

- Assignee set to dongdong tao

#2 - 03/30/2018 03:35 AM - dongdong tao

currently, there is no incoming mds messages can be fast dispatch.
so, and since the MDSBeacon's handler handle_mds_beacon is really simple.
i'm wondering maybe we can make MDSBeacon fast dispatch to avoid this kind laggy ?
@zheng, @Patrick, what's your opinion ?

#3 - 04/05/2018 11:36 PM - Patrick Donnelly

- Related to Bug #19706: Laggy mon daemons causing MDS failover (symptom: failed to set counters on mds daemons: set(['mds.dir_split'])) added

#4 - 04/09/2018 01:39 PM - Patrick Donnelly

Dongdong, I think fast dispatch may not be the answer here. We're not yet sure on the cause. Do you have ideas?

#5 - 04/09/2018 02:27 PM - dongdong tao

Patrick Donnelly wrote:

> Dongdong, I think fast dispatch may not be the answer here. We're not yet sure on the cause. Do you have ideas?

Because Beacon message is waiting in the mqueue too long for mds taking care of the previous MDSMap message.

#6 - 04/09/2018 08:12 PM - Patrick Donnelly

- Status changed from New to In Progress

- Target version set to v13.0.0

- Source set to Community (dev)

- Backport set to luminous

- Severity changed from 3 - minor to 2 - major

- Component(FS) MDS, MDSMonitor added

#7 - 04/09/2018 10:30 PM - Patrick Donnelly

- Description updated

- Category set to Correctness/Safety

#8 - 04/09/2018 10:34 PM - Patrick Donnelly

dongdong tao wrote:

> Patrick Donnelly wrote:
>
> > Dongdong, I think fast dispatch may not be the answer here. We're not yet sure on the cause. Do you have ideas?
>
> Because Beacon message is waiting in the mqueue too long for mds taking care of the previous MDSMap message.

Sorry, I didn't have my coffee yet when I was processing your original description. Yes, I think fast dispatch might make sense.

Am I correct that #23530 is a fix for this problem but you're looking for a more robust solution?

#9 - 04/10/2018 02:24 PM - dongdong tao

the two issue are not the same, but they are caused by the same reason: mds take too much time to handle MDSMap message(rejoin)
#23530 is another consequence. #23530's solution can not solve this issue.

#10 - 05/22/2018 10:15 PM - Patrick Donnelly

- Target version changed from v13.0.0 to v14.0.0

- Backport changed from luminous to mimic,luminous

#11 - 08/10/2018 09:47 AM - Zheng Yan

- Status changed from In Progress to Need Review

https://github.com/ceph/ceph/pull/23527

#12 - 08/12/2018 09:14 PM - Patrick Donnelly

- Status changed from Need Review to Pending Backport

We've merged the fast dispatch fix but I want to point out for the record that the beacon replies from the monitors are actually high priority. For this reason, the beacons are never actually "stuck" in the message queue behind less important messages. The real problem, for this issue, is that the MDS was taking too long during rejoin.

With that said, faster processing of the beacon replies outside of the mds_lock (via normal dispatch) is still a positive change worth merging.

#13 - 08/12/2018 09:17 PM - Patrick Donnelly

- Copied to Backport #26923: mimic: mds: mds got laggy because of MDSBeacon stuck in mqueue added

#14 - 08/12/2018 09:17 PM - Patrick Donnelly

- Copied to Backport #26924: luminous: mds: mds got laggy because of MDSBeacon stuck in mqueue added

#15 - 09/24/2018 11:34 AM - Nathan Cutler

- Status changed from Pending Backport to Resolved