Ceph - Fix #11590

MDSMonitor: handle MDSBeacon messages properly

05/12/2015 05:00 AM - Greg Farnum

Status:	Resolved	**Start date:**	05/12/2015
Priority:	Normal	**Due date:**	
Assignee:	Kefu Chai	**% Done:**	100%
Category:	Monitor	**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:	Q/A	**Affected Versions:**	
Tags:		**ceph-qa-suite:**	
Backport:	firefly, hammer	**Pull request ID:**	
Reviewed:			

Description

We discovered while investigating #11481 that the MDSMonitor simply does not handle MDSBeacon messages appropriately. It's supposed to send back an MMDSBeacon message in response to every one it receives, but in fact it only sends back responses to those that are ignored! (This is in preprocess_beacon)

The other paths generally stick the incoming MMDSBeacon inside of a C_Updated context that waits until a map commit has happened, but this context doesn't do anything useful with them. :(

This is made significantly worse because beacon messages have to go to the leader, and so are often forwarded. On an election, the peon will then forward the message to the leader again, and if the leader accepts it then we can do some horrible warps back in time. (That said, these outdated beacons **ought** to be rejected based on the seq and info.state_seq values, but it seems they aren't always, as in the referenced ticket.)

Related issues:

Copied to Ceph - Backport #11979: MDSMonitor: handle MDSBeacon messages properly	**Resolved**	**05/12/2015**
Copied to Ceph - Backport #11980: MDSMonitor: handle MDSBeacon messages properly	**Resolved**	**05/12/2015**

Associated revisions

Revision b3555e9c - 05/29/2015 07:21 AM - Kefu Chai

mon: always reply mdsbeacon

the MDS (Beacon) is always expecting the reply for the mdsbeacon messages from
the lead mon, and it uses the delay as a metric for the laggy-ness of the
Beacon. when it comes to the MDSMonitor on a peon, it will remove the route
session at seeing a reply (route message) from leader, so a reply to
mdsbeacon will stop the peon from resending the mdsbeacon request to the
leader.

if the MDSMonitor re-forwards the unreplied requests after they are
outdated, there are chances that the requests reflecting old and even wrong
state of the MDSs mislead the lead monitor. for example, the MDSs which sent
the outdated messages could be dead.

Fixes: #11590
Signed-off-by: Kefu Chai <kchai@redhat.com>

Revision 72a37b3a - 06/03/2015 06:14 AM - Kefu Chai

mon: send no_reply() to peon to drop ignored mdsbeacon

so the peon can remove the ignored mdsbeacon request from the routed_requets at seeing this reply, and hence no longer resend the request.

Fixes: #11590
Signed-off-by: Kefu Chai <kchai@redhat.com>

Revision a03968ad - 07/01/2015 05:56 PM - Kefu Chai

mon: send no_reply() to peon to drop ignored mdsbeacon

so the peon can remove the ignored mdsbeacon request from the routed_requets at seeing this reply, and hence no longer resend the request.

Fixes: #11590
Signed-off-by: Kefu Chai <kchai@redhat.com>
(cherry picked from commit 72a37b3a8e145d8522ea67fc14ce2c5510b6852b)

Revision 524f4a52 - 07/01/2015 05:59 PM - Kefu Chai

mon: always reply mdsbeacon

the MDS (Beacon) is always expecting the reply for the mdsbeacon messages from the lead mon, and it uses the delay as a metric for the laggy-ness of the Beacon. when it comes to the MDSMonitor on a peon, it will remove the route session at seeing a reply (route message) from leader, so a reply to mdsbeacon will stop the peon from resending the mdsbeacon request to the leader.

if the MDSMonitor re-forwards the unreplied requests after they are outdated, there are chances that the requests reflecting old and even wrong state of the MDSs mislead the lead monitor. for example, the MDSs which sent the outdated messages could be dead.

Fixes: #11590
Signed-off-by: Kefu Chai <kchai@redhat.com>
(cherry picked from commit b3555e9c328633c9e1fbc27d652c004b30535e5b)

Revision 329da091 - 07/10/2015 08:08 PM - Kefu Chai

mon: send no_reply() to peon to drop ignored mdsbeacon

so the peon can remove the ignored mdsbeacon request from the
routed_requets at seeing this reply, and hence no longer resend the
request.

Fixes: #11590
Signed-off-by: Kefu Chai <kchai@redhat.com>
(cherry picked from commit 72a37b3a8e145d8522ea67fc14ce2c5510b6852b)

Revision dc128758 - 07/10/2015 08:16 PM - Kefu Chai

mon: always reply mdsbeacon

the MDS (Beacon) is always expecting the reply for the mdsbeacon messages from
the lead mon, and it uses the delay as a metric for the laggy-ness of the
Beacon. when it comes to the MDSMonitor on a peon, it will remove the route
session at seeing a reply (route message) from leader, so a reply to
mdsbeacon will stop the peon from resending the mdsbeacon request to the
leader.

if the MDSMonitor re-forwards the unreplied requests after they are
outdated, there are chances that the requests reflecting old and even wrong
state of the MDSs mislead the lead monitor. for example, the MDSs which sent
the outdated messages could be dead.

Fixes: #11590
Signed-off-by: Kefu Chai <kchai@redhat.com>
(cherry picked from commit b3555e9c328633c9e1fbc27d652c004b30535e5b)

History

#1 - 05/12/2015 05:00 AM - Greg Farnum

This might also be the cause of the very small number of leaked messages I think Sam or Joao mentioned to me.

#2 - 05/15/2015 03:44 PM - Kefu Chai

> That said, these outdated beacons ought to be rejected based on the seq and info.state_seq values, but it seems they aren't always, as in the referenced ticket.

after the old MDS died, a new stand-by MDS replaced it. so MDSMonitor erased all the history (e.g. pending_mdsmap.mds_info, last_beacon) related to the old gid when it did the housekeeping. that's why after 4123 replaced the 4111. MDSMonitor forgot everything about 4111, including the last msg seq# from it. and the peon monitor kept resending the mdsbeacon message (mdsbeacon(4111/a-s up:boot seq 1 v0)) from the dead MDS tirelessly. this outdated message misled the lead mon and eventually brought down the innocent new MDS.

#3 - 05/15/2015 03:51 PM - Greg Farnum

D'oh, well spotted. We could maybe check for matching entity_inst_t when handling beacons, but that gets complicated with some of our naming and takeover logic.

For now just replying to beacons and stopping the retransmission is fine.

#4 - 05/18/2015 04:28 PM - Kefu Chai

- Status changed from New to Need Review

https://github.com/ceph/ceph/pull/4702

#5 - 05/29/2015 07:27 AM - Kefu Chai

see the discussions on https://github.com/ceph/ceph/pull/4702

per greg,

> So we need some kind of state message to tell peons to drop forwarded messages because they've been received and aren't getting a response.

so we can not reply to mdsbeacons ignored by the leader.

#6 - 06/03/2015 11:48 AM - Kefu Chai

- % Done changed from 0 to 60

the 2nd pull request for this issue

https://github.com/ceph/ceph/pull/4825

#7 - 06/05/2015 06:15 PM - Greg Farnum

- Status changed from Need Review to Resolved

5e99022fc68ba5305f8d69d663746d6567168ff9

#8 - 06/12/2015 01:41 AM - Kefu Chai

- Status changed from Resolved to Pending Backport

- Backport set to firefly, hammer

#9 - 09/09/2015 04:06 AM - Nathan Cutler

- Status changed from Pending Backport to Resolved

- % Done changed from 60 to 100