RADOS - Bug #40081

mon: luminous crash attempting to decode maps after nautilus quorum has been formed

05/30/2019 10:39 AM - Joao Eduardo Luis

Status:	In Progress	**Start date:**	05/30/2019
Priority:	High	**Due date:**	
Assignee:	Joao Eduardo Luis	**% Done:**	0%
Category:	Correctness/Safety	**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:	Development	**Affected Versions:**	v14.2.2
Tags:		**ceph-qa-suite:**	
Backport:		**Component(RADOS):**	Monitor
Regression:	No	**Pull request ID:**	28671
Severity:	2 - major	**Crash signature:**	
Reviewed:			

Description

While upgrading, we found a rather annoying corner case:

Assuming we start with 3 luminous ceph-mon, upgrading from luminous to nautilus,

1. shutdown mon.a
2. upgrade mon.a
3. shutdown mon.b
4. upgrade mon.b
5. shutdown mon.c
6. wait for a bit
7. bring back mon.c as luminous, without upgrading

I.e., we have a 2 nautilus + 1 luminous ceph-mons.

Due to #38850, the luminous monitor would not be able to join the quorum unless forced, but that causes issues of itself by breaking quorum during upgrade.
With the backport of 90e4c5f in https://github.com/ceph/ceph/pull/28262 we would not need to force the monitor to join the quorum, and it would automatically be accepted.

However, regardless of which approach is taken ('quorum enter' or using 90e4c5f), the result is the same: mon.c will crash upon decoding nautilus-encoded maps.

This makes sense given the nautilus monitors are encoding maps, during their time as the only participants in the quorum, with features that luminous does not understand. And it seems rather difficult (if not at all impossible with major code changes) to ensure those maps are encoded with backwards compatibility - given we use the quorum's connection features to encode the maps, those are established by the minimum set of connection features supported by members of the quorum (not to be confused with monmap's features); in this case, the minimum set of features would be nautilus (given mon.c would be shutdown and thus not able to participate during probing).

The only way I see out of this is to force the new quorum not to accept a new monitor if it doesn't support all the connection features existing in the quorum. We are still guaranteeing a rolling upgrade, but this effectively means that once we have a nautilus quorum, no other monitors will be able to join it until they are upgraded (exposing us to tolerance of zero failures).

Either this or we'd have to take into account the monmap's features to encode maps; but this seems a pretty significant change to backport to nautilus, and I'm assuming it would affect clients as well. So the previous solution seems easier, and cleaner, despite its drawbacks.

History

#1 - 05/30/2019 02:49 PM - Joao Eduardo Luis

https://github.com/ceph/ceph/pull/28323 (closed; see Pull Request ID field for the real PR)

This actually has us ending up in the same behavior as before-ish, with the luminous monitor unable to join the quorum - but this time is by design.

I think I'll have it graciously commit suicide with a pretty message explaining why.

#2 - 06/05/2019 09:03 PM - Neha Ojha

- Status changed from New to In Progress

#3 - 06/20/2019 11:42 AM - Kefu Chai

- Pull request ID set to 28671

#4 - 07/17/2019 08:54 PM - Sage Weil

https://github.com/ceph/ceph/pull/28672 (nautilus backport PR)

#5 - 08/20/2019 08:43 PM - Greg Farnum

- Priority changed from Urgent to High