Ceph - Bug #15073

osd: osdmap write not ordered before pgs consume map

03/11/2016 02:02 PM - Sage Weil

Status:	Resolved	**Start date:**	03/11/2016
Priority:	Urgent	**Due date:**	
Assignee:	Sage Weil	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:	Q/A	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor		

Description

/a/sage-2016-03-10_19:53:19-rados:thrash-wip-bluestore---basic-mira/52021

- osdmap write txn is queued
- pg consumes map, writes pg info
- pg info commits
- **crash before osdmap is written**

This was triggered with BlueStore. FileStore orders the transactions due to its journaling even though different sequencers are used.

Associated revisions

Revision b839a06c - 03/17/2016 04:58 PM - Sage Weil

osd: commit osdmaps before exposing them to PGs

handle_osd_map and the PGs use different sequencers when writing
their updates. We therefore need to make sure new osdmaps are
committed to disk before we expose them to PGs, lest they update
their info to reference a new osdmap that hasn't actually
committed yet.

This doesn't happen with FileStore because transactions are
ordered when they are queued, but it does affect BlueStore.

Fix by splitting handle_osd_map into two phases, one that just
persists stuff, and the second half that publishes the new maps to
the rest of the OSD.

Fixes: #15073
Signed-off-by: Sage Weil <sage@redhat.com>

History

#1 - 03/11/2016 05:29 PM - Sage Weil

- Status changed from Verified to Testing

- Assignee set to Sage Weil

https://github.com/liewegas/ceph/commit/wip-15073

#2 - 03/14/2016 04:49 PM - Sage Weil

https://github.com/ceph/ceph/pull/8096

#3 - 03/17/2016 09:56 PM - Sage Weil

- Status changed from Testing to Resolved