Ceph - Bug #16127

OSDMonitor: drop pg temps from not the current primary

06/02/2016 05:18 PM - Samuel Just

Status:	Resolved	**Start date:**	06/02/2016
Priority:	Urgent	**Due date:**	
Assignee:	Samuel Just	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:	other	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:	jewel,hammer	**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor		

Description

Otherwise, the vagaries of pg->osd->mon->other mon message ordering could result in a previous interval pg temp request being processed after a current interval request causing the pg to get stuck.

sjust@teuthology:/a/samuelj-2016-06-01_11:28:31-rados-wip-sam-testing-distro-basic-smithi/228382/remote

<sjusthm> sage: hmm
<sjusthm> pg temps from two osds raced
<sjusthm> on the same osd
<sjusthm> osd.3 sent a request for [0,3] at pg_epoch 16
<sjusthm> which went out at osd epoch 16
<sjusthm> and then again at 17
<sjusthm> (before the pg processed the map)
<sjusthm> however
<sjusthm> 17 also changed the acting set to [0,3]

- kefu has quit (Quit: My Mac has gone to sleep. ZZZzzz…)
  <sjusthm> and the new primary requested an empty temp mapping
  <sjusthm> also at pg/osd epoch 17
  <sjusthm> the mons processed the one from the new primary
  <sjusthm> and then the stale one from the old primary
  <sjusthm> resulting in the acting set remaining at [0,3] and the pg being stuck
  <sjusthm> the osd epoch part seems to bge a red herring since it's not used in the OSDMonitor
  <sjusthm> I think we need to include with each mapping the interval start epoch
  <sjusthm> and remember that in the OSDMonitor
  <sjusthm> that would allow us to dicard pg temp mappings based on previous intervals
  <sjusthm> hmm
  <sjusthm> wouldn't actually help here
  <sjusthm> since the empty mapping wouldn't be remembered explicitely
- sahid has quit (Quit: Lost terminal)
  <sjusthm> maybe we can just ignore if it comes from not the current primary from the mon's point of view?
- rzarzynski has quit (Quit: This computer has gone to sleep)
- dgurtner_ has quit (Ping timeout: 480 seconds)
- swami2 has quit (Ping timeout: 480 seconds)
- rzarzynski (~rzarzynsk@80.87.33.6) has joined #ceph-devel
  <sjusthm> I guess that should be safe
  <sjusthm> can I do that in preprocess?

Related issues:

Copied to Ceph - Backport #16429: jewel: OSDMonitor: drop pg temps from not t...	**Resolved**
Copied to Ceph - Backport #16430: hammer: OSDMonitor: drop pg temps from not ...	**Resolved**

History

#1 - 06/11/2016 12:52 AM - John Coyle

Maybe unrelated but I'm curious if this issue would cause cephtool-test-mon.sh to get stuck at:

../qa/workunits/cephtool/test.sh:1401: test_mon_pg: ceph osd primary-affinity osd.0 1

- DEVELOPER MODE: setting PATH, PYTHONPATH and LD_LIBRARY_PATH ***
  2016-06-10 17:36:58.182222 7fbb90cfeab0 -1 WARNING: the following dangerous and experimental features are enabled: *
  2016-06-10 17:36:58.182267 7fbb90cfeab0 0 lockdep start
  2016-06-10 17:36:58.187814 7fbb90cfeab0 -1 WARNING: the following dangerous and experimental features are enabled: *
  set osd.0 primary-affinity to 1 (8655362)
  2016-06-10 17:36:58.370347 7fbb90cfeab0 0 lockdep stop
  ../qa/workunits/cephtool/test.sh:1403: test_mon_pg: ceph osd pg-temp 0.0 0 1 2
- DEVELOPER MODE: setting PATH, PYTHONPATH and LD_LIBRARY_PATH ***
  2016-06-10 17:36:58.454619 7fc3e7279ab0 -1 WARNING: the following dangerous and experimental features are enabled: *
  2016-06-10 17:36:58.454625 7fc3e7279ab0 0 lockdep start
  2016-06-10 17:36:58.464076 7fc3e7279ab0 -1 WARNING: the following dangerous and experimental features are enabled: *
  2016-06-10 17:37:13.384954 7fc3d78c2ab0 0 monclient: hunting for new mon
  2016-06-10 17:37:33.464059 7fc3d78c2ab0 0 monclient: hunting for new mon

Thanks!

#2 - 06/13/2016 07:49 PM - Samuel Just

sjust@teuthology:/a/teuthology-2016-06-12_18:15:02-upgrade:hammer-x-jewel-distro-basic-vps/255160/remote as well I think.

#3 - 06/22/2016 05:28 PM - Samuel Just

- Status changed from New to Pending Backport

https://github.com/ceph/ceph/pull/9875

#4 - 06/22/2016 09:09 PM - Nathan Cutler

- Copied to Backport #16429: jewel: OSDMonitor: drop pg temps from not the current primary added

#5 - 06/22/2016 09:09 PM - Nathan Cutler

- Copied to Backport #16430: hammer: OSDMonitor: drop pg temps from not the current primary added

#6 - 08/08/2016 08:42 AM - Loic Dachary

- Status changed from Pending Backport to Resolved