rgw - Bug #17574

multisite: many duplicate mdlog entries cause race to sync and result in ECANCELED

10/13/2016 07:03 PM - Casey Bodley

Status:	New	**Start date:**	10/13/2016
Priority:	Normal	**Due date:**	
Assignee:	Casey Bodley	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:	other	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

Seen in multiple teuthology runs:

http://qa-proxy.ceph.com/teuthology/cbodley-2016-09-16_10:00:36-rgw-wip-cbodley-testing---basic-mira/419161/teuthology.log
- osd.2 is the culprit here, but the log doesn't include 'debug cls'
http://qa-proxy.ceph.com/teuthology/cbodley-2016-09-16_10:00:36-rgw-wip-cbodley-testing---basic-mira/419161/remote/mira041/log/ceph-osd.2.log.gz

http://qa-proxy.ceph.com/teuthology/owasserm-2016-09-26_16:09:01-rgw-wip-orit-testing---basic-mira/438706/teuthology.log

Related issues:

Related to rgw - Bug #17465: multisite: coroutine deadlock in RGWMetaSyncCR a...	**Resolved**	**09/30/2016**
Related to rgw - Bug #17996: rgw: ECANCELED in rgw_get_system_obj() leads to ...	**Resolved**	**11/22/2016**

History

#1 - 10/13/2016 07:05 PM - Casey Bodley

- Related to Bug #17465: multisite: coroutine deadlock in RGWMetaSyncCR after ECANCELED errors added

#2 - 10/24/2016 01:52 PM - Abhishek Lekshmanan

Seeing similar errors in jewel; in delete bucket operations where the OSD returns -ECANCELLED and we return a 500

#3 - 10/25/2016 03:38 PM - Casey Bodley

I pushed a branch to ceph-qa-suite [1] that sets debug_objclass=20 and scheduled a run of the rgw/verify suite [2].

[1] https://github.com/ceph/ceph-qa-suite/commits/wip-rgw-ecanceled
[2] http://pulpito.ceph.com/cbodley-2016-10-25_11:34:59-rgw:verify-master---basic-mira/

#4 - 10/26/2016 03:59 PM - Casey Bodley

This issue reproduced in job 494193. The osd log [1] indicates that the majority of ECANCELED errors are coming from 'cls_version: failed condition check'. The osd operations look like '.bucket.meta.test-client.0-ba5gfwy6dpmh9u4-271:r0z1.4136.272 [delete,create 0~0,call version.check_conds,call version.set,writefull 0~277,setxattr user.rgw.acl (185)]' and '[call version.check_conds,call version.read,read 0~524288]'.

An interesting note about the gateway issuing these requests, its log [2] shows many duplicated mdlog entries:

```
2016-10-26 00:00:28.103066 35c45700 20 meta sync: remote mdlog, shard_id=47 num of shard entries: 70
2016-10-26 00:00:28.103747 35c45700 20 cr:s=0x73da9930:op=0x782087a0:24RGWCloneMetaLogCoroutine: operate()
2016-10-26 00:00:28.103801 35c45700 20 meta sync: operate: shard_id=47: storing mdlog entries
2016-10-26 00:00:28.103840 35c45700 20 meta sync: entry: name=test-client.0-ba5gfwy6dpmh9u4-271:r0z1.4136.272
2016-10-26 00:00:28.103954 35c45700 20 meta sync: entry: name=test-client.0-ba5gfwy6dpmh9u4-271:r0z1.4136.272
```

```
2016-10-26 00:00:28.104042 35c45700 20 meta sync: entry: name=test-client.0-ba5gfwy6dpmh9u4-271
2016-10-26 00:00:28.104122 35c45700 20 meta sync: entry: name=test-client.0-ba5gfwy6dpmh9u4-271
2016-10-26 00:00:28.104200 35c45700 20 meta sync: entry: name=test-client.0-ba5gfwy6dpmh9u4-271:r0z1.4136.272
2016-10-26 00:00:28.104288 35c45700 20 meta sync: entry: name=test-client.0-ba5gfwy6dpmh9u4-271:r0z1.4136.272
2016-10-26 00:00:28.104376 35c45700 20 meta sync: entry: name=test-client.0-ba5gfwy6dpmh9u4-271:r0z1.4136.272
2016-10-26 00:00:28.104463 35c45700 20 meta sync: entry: name=test-client.0-ba5gfwy6dpmh9u4-271:r0z1.4136.272
```

```
... repeated 62 more times ...
```

The gateway goes on to spawn a bunch of RGWMetaSyncSingleEntryCRs for these, so they're all fetching the bucket instance metadata from the master and racing to store the entry via RGWMetaStoreEntryCR.

[1] http://qa-proxy.ceph.com/teuthology/cbodley-2016-10-25_11:34:59-rgw:verify-master---basic-mira/494193/remote/mira023/log/ceph-osd.2.log.gz
[2] http://qa-proxy.ceph.com/teuthology/cbodley-2016-10-25_11:34:59-rgw:verify-master---basic-mira/494193/remote/mira032/log/rgw.client.1.log.gz

#5 - 10/27/2016 06:16 PM - Yehuda Sadeh

- Priority changed from Normal to High

#6 - 11/10/2016 06:48 PM - Casey Bodley

- Assignee set to Casey Bodley

#7 - 12/08/2016 07:16 PM - Casey Bodley

- Related to Bug #17996: rgw: ECANCELED in rgw_get_system_obj() leads to infinite loop added

#8 - 12/22/2016 07:04 PM - Casey Bodley

- Subject changed from single osd starts failing cls operations with ECANCELED to multisite: many duplicate mdlog entries cause race to sync and result in ECANCELED

#9 - 02/09/2017 06:53 PM - Yehuda Sadeh

Discussed in bug scrub. We think that what we see there is not necessarily a bug, but s3-tests doing a lot of metadata changes. We can probably have a squash mechanism in the metadata sync similar to the one that we have in the data sync.

#10 - 03/23/2017 05:44 PM - Yehuda Sadeh

- Priority changed from High to Normal