rgw - Bug #20612

radosgw ceases responding to list requests

07/13/2017 01:44 AM - Bob Bobington

Status:	Can't reproduce	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:			
Category:			
Target version:	v12.1.0		
Source:	Community (user)	**Reviewed:**	
Tags:		**Affected Versions:**	v12.1.0
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

I have 4 OSDs set up with --dmcrypt and --bluestore. Occasionally, likely due to http://tracker.ceph.com/issues/20545 they crash.

I'm running radosgw with CivetWeb (the el7 package doesn't seem to have fastcgi enabled) and I've configured a pool with erasure coding:

```
ceph osd erasure-code-profile set myprofile k=2 m=1 ruleset-failure-domain=osd
ceph osd pool create default.rgw.buckets.data 256 256 erasure myprofile
systemctl start ceph-radosgw@rgw.radosgw.service
```

I'm running a copy with the open source rclone tool:

```
rclone copy -v --transfers 4 --stats 10s /place/on/disk/ ceph:bucket
```

And after a while, an OSD will always stop responding to queries, resulting in things like this turning up in the logs:

```
2017-07-12 16:23:21.121534 osd.2 osd.2 192.168.122.132:6808/75831 27539 : cluster [WRN] 5022 slow
requests, 5 included below; oldest blocked for > 4871.736939 secs
2017-07-12 16:23:21.121544 osd.2 osd.2 192.168.122.132:6808/75831 27540 : cluster [WRN] slow reque
st 960.787066 seconds old, received at 2017-07-12 16:07:20.333473: osd_op(client.14171.0:714555 10
.7 10.4322fa9f (undecoded) ondisk+write+kn
own_if_redirected e124) currently queued_for_pg
```

After this, I generally end up manually restarting the OSD then restarting radosgw. When I make subsequent requests, I see the following repeated in the radosgw logs until rclone gives up:

```
2017-07-12 18:08:44.918031 7f0cc3acb700  1 ====== starting new request req=0x7f0cc3ac55d0 =====
2017-07-12 18:08:45.011691 7f0cc3acb700  0 WARNING: set_req_state_err err_no=36 resorting to 500
2017-07-12 18:08:45.011862 7f0cc3acb700  1 ====== req done req=0x7f0cc3ac55d0 op status=-36 http_s
tatus=500 ======
2017-07-12 18:08:45.011941 7f0cc3acb700  1 civetweb: 0x7f0cf84ba000: 192.168.122.1 - - [12/Jul/201
7:18:07:57 -0700] "GET /bucket?delimiter=%2F&max-keys=1024&prefix= HTTP/1.1" 1 0 - rclone/v1.36
```

The data pool has ~172k objects and the bucket ~13k, so I expect a list request to be somewhat expensive but that doesn't seem to be what this is.

I'm able to retrieve objects fine, it seems to only be list requests that have issues.

History

#1 - 07/13/2017 03:15 AM - Greg Farnum

- Project changed from Ceph to rgw

- Category deleted (22)

#2 - 07/13/2017 03:50 AM - Matt Benjamin

@greg,

does this really look like a radosgw issue to you?

Matt

#3 - 07/20/2017 06:07 PM - Matt Benjamin

could we have radosgw logs at --debug-rgw=20 --debug-ms=1, please?

#4 - 07/20/2017 06:07 PM - Orit Wasserman

- Status changed from New to Need More Info

#5 - 07/21/2017 12:33 AM - Bob Bobington

I've since found another solution for my storage needs and allocated my hardware to that, so I can't reproduce I'm afraid.

#6 - 07/27/2017 05:55 PM - Orit Wasserman

- Status changed from Need More Info to Can't reproduce

#7 - 10/07/2017 05:47 AM - Wei Wu

I had same issue.

Running Ceph

```
$ ceph --version
ceph version 12.2.1 (3e7492b9ada8bdc9a5cd0feafd42fbca27f9c38e) luminous (stable)

$ uname -a
Linux data-buw1-ceph-storage-data-data-01-02 3.10.0-693.2.2.el7.x86_64 #1 SMP Tue Sep 12 22:26:13 UTC 2017 x86
_64 x86_64 x86_64 GNU/Linux

$ cat /etc/centos-release
CentOS Linux release 7.4.1708 (Core)
```

```
2017-10-07 01:42:40.473378 7fb5b4cb5700 10 RGWRados::cls_bucket_list: got /external/data1/google_images/airpor
t_runway/_gallery_rx-12_aa_takeoff_2002.jpg[]
2017-10-07 01:42:40.473380 7fb5b4cb5700 10 RGWRados::cls_bucket_list: got /external/data1/google_images/airpor
t_runway/_gawker-media_image_upload_tasrko6vvs0e0euvirll.jpg[]
2017-10-07 01:42:40.473391 7fb5b4cb5700 20 get_obj_state: rctx=0x7fb5b4cad810 obj=my-new-bucket:/external/data
1/google_images/airport_runway/_gc_154529266-looking-down-an-empty-airport-runway-gettyimages.jpg%253Fv%253D1%
2526c%253DIWSAsset%2526k%253D2%2526d%253D96J4IMAdRiLTGCtfsVU3Ir4NB5Kb12dztTnxi5EyDwkzWf5RlzvyAJwHmCg3x44Q stat
e=0x5626488398a0 s->prefetch_data=0
2017-10-07 01:42:40.473428 7fb5b4cb5700  1 -- 10.10.6.52:0/1910691890 --> 10.10.6.53:6800/13199 -- osd_op(unkn
own.0.0:9 15.1595 15:a9aef266:::7c0eb44f-296e-42ea-b020-312f85ada593.154102.1_%2fexternal%2fdata1%2fgoogle_ima
ges%2fairport_runway%2f_gc_154529266-looking-down-an-empty-airport-runway-gettyimages.jpg%25253Fv%25253D1%2525
```

```
26c%25253DIWSAsset%252526k%25253D2%252526d%25253D96J4IMAdRiLTGCtfsVU3Ir4NB5Kb12dztTnxi5EyDwkzWf5RlzvyAJwHmCg3x
44Q:head [getxattrs,stat] snapc 0=[] ondisk+read+known_if_redirected e787) v8 -- 0x562649361800 con 0
2017-10-07 01:42:40.473996 7fb5e4bb8700  1 -- 10.10.6.52:0/1910691890 <== osd.0 10.10.6.53:6800/13199 4 ==== o
sd_op_reply(9 7c0eb44f-296e-42ea-b020-312f85ada593.154102.1_/external/data1/google_images/airport_runway/_gc_1
54529266-looking-down-an-empty-airport-runway-gettyimages.jpg%253Fv%253D1%2526c%253DIWSAsset%2526k%253D2%2526d
%253D96J4IMAdRiLTGCtfsVU3Ir4NB5Kb12dztTnxi5EyDwkzWf5RlzvyAJwHmCg3x44Q [getxattrs,stat] v0'0 uv0 ondisk = -36 (
(36) File name too long)) v8 ==== 461+0+0 (2335629423 0 0) 0x562649361800 con 0x562649440000
2017-10-07 01:42:40.474448 7fb5b4cb5700  2 req 23:0.045474:s3:GET /my-new-bucket:list_bucket:completing
2017-10-07 01:42:40.474457 7fb5b4cb5700  0 WARNING: set_req_state_err err_no=36 resorting to 500
2017-10-07 01:42:40.474522 7fb5b4cb5700  2 req 23:0.045549:s3:GET /my-new-bucket:list_bucket:op status=-36
2017-10-07 01:42:40.474526 7fb5b4cb5700  2 req 23:0.045553:s3:GET /my-new-bucket:list_bucket:http status=500
2017-10-07 01:42:40.474529 7fb5b4cb5700  1 ====== req done req=0x7fb5b4caf180 op status=-36 http_status=500 ==
====
```

#8 - 10/08/2017 01:04 AM - Bob Bobington

I've got my Ceph cluster back up and I can confirm that I encounter the same issue as Wei Wu.

#9 - 10/08/2017 01:24 AM - Bob Bobington

It turns out my problem was the result of an old "osd max object name len = 256" entry in my configuration. I removed it and the error went away.

#10 - 10/08/2017 10:03 PM - Wei Wu

It did work when I remove `osd max object name len` from my config.

Thanks