RADOS - Bug #38931

osd does not proactively remove leftover PGs

03/25/2019 11:14 AM - Dan van der Ster

Status:	New	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:			
Category:			
Target version:			
Source:		**Affected Versions:**	v12.2.11
Tags:		**ceph-qa-suite:**	
Backport:		**Component(RADOS):**	
Regression:	No	**Pull request ID:**	
Severity:	2 - major	**Crash signature (v1):**	
Reviewed:		**Crash signature (v2):**	

Description

(Context: cephfs cluster running v12.2.11)

We had an osd go nearfull this weekend. I reweighted it to move out some PGs, but when looking today it's still holding much more data than it should.

The osd currently has 34 PGs mapped to it:

```
  74   hdd 5.45609  1.00000 5.46TiB 3.86TiB 1.60TiB 70.77 1.37  34
```

But the OSD itself reports 20 more:

```
{
    "whoami": 74,
    "state": "active",
    "oldest_map": 46992,
    "newest_map": 47738,
    "num_pgs": 54
}
```

When I restart the OSD, it reloads those 20, e.g. here is a PG it loads but which is mapped to [22,129,14]. That PG is currently active+clean.

```
2019-03-25 11:09:26.655955 7fe3fdb23d80 10 osd.74 47719 load_pgs loaded pg[2.6d( v 47685'27177090
(47637'27175587,47685'27177090] lb MIN (bitwise) local-lis/les=47680/47681 n=0 ec=371/371 lis/c 47
683/47497 les/c/f 47684/47498/0 47686/47688/43553) [22,129,14] r=-1 lpr=47689 pi=[47497,47688)/1 c
rt=47685'27177090 lcod 0'0 unknown NOTIFY mbc={}] log((47637'27175587,47685'27177090], crt=47685'2
7177090)
```

I found a way to remove those leftover PGs (without using ceph-objectstore-tool): If the PG re-peers, then osd.74 notices he's not in the up/acting set then starts deleting the PG.
So at the moment I'm restarting those former peers to trim this OSD.

Is this all an expected behaviour?

Shouldn't the OSD start removing leftover PGs at boot time?

History

#1 - 03/30/2019 07:14 PM - Neha Ojha

https://github.com/ceph/ceph/pull/27205/commits/f7c5b01e181630bb15e8b923b0334eb6adfdf50a

#2 - 04/04/2019 08:37 PM - Greg Farnum

So should we backport part of that PR, Neha?

To answer your question more directly, Dan: OSDs don't delete PGs themselves because they don't know if the data is still needed; they wait for the primary to tell them to remove it. Based on the linked commit, apparently there's a bug where in some circumstances the primary will erroneously mark the stray OSD as having deleted the PG already though, and you seem to have fallen victim to that.

#3 - 04/04/2019 09:29 PM - Neha Ojha

Greg Farnum wrote:

> So should we backport part of that PR, Neha?
>
> To answer your question more directly, Dan: OSDs don't delete PGs themselves because they don't know if the data is still needed; they wait for the primary to tell them to remove it. Based on the linked commit, apparently there's a bug where in some circumstances the primary will erroneously mark the stray OSD as having deleted the PG already though, and you seem to have fallen victim to that.

I think so, guess I wasn't sure since Xie Xingguo used "Related-to" instead of "Fixes:" in that commit.
As a matter of fact, the parent PR https://github.com/ceph/ceph/pull/27205 is also pending backport.