RADOS - Bug #38184

osd: recovery does not preserve copy-on-write allocations between object clones after 'rbd revert'

02/05/2019 04:48 PM - Vitaliy Filippov

Status:	New	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:			
Category:			
Target version:			
Source:		**Affected Versions:**	
Tags:		**ceph-qa-suite:**	
Backport:		**Component(RADOS):**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature (v1):**	
Reviewed:		**Crash signature (v2):**	

Description

Hi. I've already reported it in issue 36614, but here is a more concrete case.

- Start with a bluestore Ceph cluster
- Create an RBD image
- Fill it with data
- Remember disk space used by the image as X
- Create a snapshot of it
- Immediately revert to it (rbd snap revert)
- After revert finishes you'll see that there was still X space used, but object count in the cluster is doubled
- Trigger a massive rebalance in the cluster
- After rebalance finishes you'll see that the image's objects residing in moved PGs now use 2*X disk space. This is because virtual clones stop being virtual after their data is moved
- Now run rbd snap revert again
- You'll see the space usage drop. This is because "virtual clones" become "virtual" again.

I think it's a bug and should be fixed. It had led to a bad situation in our cluster once, described in issue 36614.

History

#1 - 02/27/2019 12:03 AM - Vitaliy Filippov

Anyone?

#2 - 02/28/2019 03:25 PM - Sage Weil

- Project changed from bluestore to RADOS

- Subject changed from Virtual clones break and begin to eat space after rebalancing to osd: recovery does not preserve copy-on-write allocations between object clones after 'rbd revert'

- Status changed from New to 12

This is indeed the current behavior. The OSD isn't clever enough to preserve the shared allocations across recovery. It is a large effort to change this.

#3 - 12/05/2019 09:36 PM - Patrick Donnelly

- Status changed from 12 to New