RADOS - Bug #52583

partial recovery become whole object recovery after restart osd

09/13/2021 09:43 AM - jianwei zhang

Status:	Pending Backport	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:			
Category:	Backfill/Recovery		
Target version:			
Source:	Community (user)	**Affected Versions:**	v15.2.8
Tags:		**ceph-qa-suite:**	rados
Backport:	pacific, octopus	**Component(RADOS):**	OSD
Regression:	No	**Pull request ID:**	43146
Severity:	3 - minor	**Crash signature (v1):**	
Reviewed:	08/25/2021	**Crash signature (v2):**	

Description

Problem: After the osd that is undergoing partial recovery is restarted, the data recovery is rolled back from the partial recovery to the whole object recovery.

```
Steps to reproduce:
1. ceph osd set noout
2. vdbench 5000 iops 8K 7:3 randrw
3. after running 3 minutes, stop osd.0 running for 10 minutes
4. start osd.0   --> Note 1: At this time, it is a real partial recovery
5. During the parital recovery, again to stop osd.0
6. After stopping osd.0, immediately start osd.0 again --> Note 2
: At this time, it becomes a whole object recovery
```

```
1. ceph osd set noout
2. vdbench 5000 iops 8K 7:3 randrw
3. after 3 minutes, stop osd.0 running for 30 minutes
4. start osd.0   --> Note 1: At this time, it is a real partial recovery

>>> first pg peering
/// /var/log/ceph/ceph-osd.0.log  primary   osd.0 calc pg_missing_item, clean_offsets is not empty
2021-08-20T15:30:22.229+0800 7fcb19056700  1 add_next_event MY-DEBUG:log=1743'10560 (1528'9407
) modify   3:efff9aef:::rbd_data.8a5ca0e29cea.000000000000ad15:head by client.1855642.0:67946 2021
-08-20T15:20:16.144439+0800 0 ObjectCleanRegions clean_offsets: [0~1064960,1073152~
18446744073708478463], clean_omap: 1, new_object: 0, missing_item=1743'10560(1528'9407
) flags = none clean_offsets: [0~1064960,1073152~18446744073708478463], clean_omap: 1
, new_object: 0
2021-08-20T15:30:26.172+0800 7fcb19056700  1 osd.0 pg_epoch: 1746 pg[3.1f7( v 1743'11859 (1526'8
788,1743'11859] local-lis/les=1745/1746 n=956 ec=229/108 lis/c=1742/1724 les/c/f=1743/1725/0 sis=
1745) [2,0,5]/[0,5] async=[2] r=0 lpr=1745 pi=[1724,1745)/1 crt=1743'11859 lcod 1743'11858 mlcod 0
'0 active+undersized+degraded+remapped mbc={255={(2+0)=721
}}] state<Started/Primary/Active>: react(AllReplicasActivated) MY-DEBUG: osd.2
 pg_missing_t sobject: 3:efff9aef:::rbd_data.8a5ca0e29cea.000000000000ad15:head, pg_missing_item:
1743'10560(1528'9407) flags = none clean_offsets: [0~1064960,1073152~18446744073708478463
], clean_omap: 1, new_object: 0

/// /var/log/ceph/ceph-osd.2log replica     osd.2 calc pg_missing_item, clean_offsets is not empty
2021-08-20T15:30:25.274+0800 7f135b750700  1 add_next_event MY-DEBUG:log=1743'10560 (1528'9407
) modify   3:efff9aef:::rbd_data.8a5ca0e29cea.000000000000ad15:head by client.1855642.0:67946 2021
-08-20T15:20:16.144439+0800 0 ObjectCleanRegions clean_offsets: [0~1064960,1073152~
18446744073708478463], clean_omap: 1, new_object: 0, missing_item=1743'10560(1528'9407
) flags = none clean_offsets: [0~1064960,1073152~18446744073708478463], clean_omap: 1
, new_object: 0
2021-08-20T15:30:25.332+0800 7f135b750700  1 merge_log MY-DEBUG-2: pg_missing_t sobject=3
:efff9aef:::rbd_data.8a5ca0e29cea.000000000000ad15:head, missing_item=1743'10560(1528'9407
) flags = none clean_offsets: [0~1064960,1073152~18446744073708478463], clean_omap: 1
, new_object: 0

5. During the parital recovery, again to stop osd.0
6. After stopping osd.0, immediately start osd.0 again --> Note 2
: At this time, it becomes a whole object recovery

>>> second pg peering
/// /var/log/ceph/ceph-osd.2.log replica       osd.2 read missing from rocksdb, but clean_offsets
is empty
2021-08-20T15:32:17.332+0800 7f728062dbc0  1
 read_log_and_missing MY-DEBUG-I: pg_missing_t sobject=3:efff9aef:::rbd_data.8a5ca0e29cea
.000000000000ad15:head, missing_item=1743'10560(1528'9407
) flags = none clean_offsets: [], clean_omap: 0, new_object: 1
2021-08-20T15:32:17.363+0800 7f728062dbc0  1
 read_log_and_missing MY-DEBUG-II: pg_missing_t sobject=3:efff9aef:::rbd_data.8a5ca0e29cea
.000000000000ad15:head, missing_item=1743'10560(1528'9407
) flags = none clean_offsets: [], clean_omap: 0, new_object: 1

/// /var/log/ceph/ceph-osd.0.log  primary    osd.0 get missing from osd.2 by network, so clean_off
sets is empty
2021-08-20T15:32:38.469+0800 7fcb19056700  1 proc_replica_log MY-DEBUG-A: pg_missing_t sobject=3
:efff9aef:::rbd_data.8a5ca0e29cea.000000000000ad15:head, missing_item=1743'10560(1528'9407
) flags = none clean_offsets: [], clean_omap: 0, new_object: 1
2021-08-20T15:32:39.392+0800 7fcb19056700  1 osd.0 pg_epoch: 1775 pg[3.1f7( v 1743'11859 (1526'8
788,1743'11859] local-lis/les=1774/1775 n=956 ec=229/108 lis/c=1770/1724 les/c/f=1771/1725/0 sis=
1774) [2,0,5]/[0,5] async=[2] r=0 lpr=1774 pi=[1724,1774)/1 crt=1743'11859 lcod 1743'11858 mlcod 0
'0 active+undersized+degraded+remapped mbc={255={(2+0)=721
}}] state<Started/Primary/Active>: react(AllReplicasActivated) MY-DEBUG: osd.2
 pg_missing_t sobject: 3:efff9aef:::rbd_data.8a5ca0e29cea.000000000000ad15:head, pg_missing_item:
1743'10560(1528'9407) flags = none clean_offsets: [], clean_omap: 0, new_object: 1
```

Note 1: In the first peering, osd.0 is primary, and add_next_event() is performed to build pg_missing_item, log transfer to osd.2 of the replica pg, and osd.2 does add_next_event() to build its pg_missing_item, the result is indeed an partial clean_offsets

Note 2: During the second peering, the clean_offsets of the pg_missing_item read from osd.2 is empty, that is, it is marked as whole object recovery

Question: Does this mean that the partial clean_offsets of pg_missing_item build on osd.2 is not write to disk?

Related issues:

Copied to RADOS - Backport #52620: pacific: partial recovery become whole obj...	**Resolved**
Copied to RADOS - Backport #52710: octopus: partial recovery become whole obj...	**In Progress**

History

#1 - 09/13/2021 10:24 AM - jianwei zhang

FIX URL:
https://github.com/ceph/ceph/pull/43146
https://github.com/ceph/ceph/pull/42904

#2 - 09/15/2021 03:18 PM - Kefu Chai

- Status changed from New to Pending Backport

- Backport set to pacific

#3 - 09/15/2021 03:18 PM - Kefu Chai

- Pull request ID changed from 42904 to 43146

#4 - 09/15/2021 03:20 PM - Backport Bot

- Copied to Backport #52620: pacific: partial recovery become whole object recovery after restart osd added

#5 - 09/22/2021 10:42 PM - Neha Ojha

- Backport changed from pacific to pacific, octopus

#6 - 09/22/2021 10:45 PM - Backport Bot

- Copied to Backport #52710: octopus: partial recovery become whole object recovery after restart osd added