RADOS - Bug #54263

cephadm upgrade pacific to quincy autoscaler is scaling pgs from 32 -> 32768 for cephfs meta pool

02/11/2022 10:39 PM - Vikhyat Umrao

Status:	Resolved	**% Done:**	0%
Priority:	High	**Spent time:**	0.00 hour
Assignee:	Kamoltat Sirivadhna		
Category:			
Target version:			
Source:		**Affected Versions:**	v17.0.0
Tags:		**ceph-qa-suite:**	
Backport:	quincy, pacific	**Component(RADOS):**	
Regression:	No	**Pull request ID:**	45200
Severity:	3 - minor	**Crash signature (v1):**	
Reviewed:		**Crash signature (v2):**	

Description

Pacific version - 16.2.7-34.el8cp
Quincy version - 17.0.0-10315-ga00e8b31

After doing some analysis it looks like during the upgrade to the quincy version autoscaler TARGET RATIO got set as 4.0.

- After upgrade some commands output:

```
# ceph osd pool ls detail
pool 1 '.mgr' replicated size 3 min_size 2 crush_rule 0 object_hash rjenkins pg_num 1 pgp_num 1 au
toscale_mode on last_change 25006 flags hashpspool,backfillfull stripe_width 0 pg_num_min 1 applic
ation mgr,mgr_devicehealth
pool 2 'rbd' replicated size 3 min_size 2 crush_rule 0 object_hash rjenkins pg_num 256 pgp_num 256
 autoscale_mode on last_change 25006 lfor 0/0/1324 flags hashpspool,backfillfull,selfmanaged_snaps
 stripe_width 0 application rbd
pool 3 'cephfs.cephfs.meta' replicated size 3 min_size 2 crush_rule 0 object_hash rjenkins pg_num
32768 pgp_num 32768 autoscale_mode on last_change 25006 lfor 0/0/9281 flags hashpspool,backfillful
l stripe_width 0 pg_num_min 16 recovery_priority 5 target_size_ratio 4 application cephfs
pool 4 'cephfs.cephfs.data' replicated size 3 min_size 2 crush_rule 0 object_hash rjenkins pg_num
4435 pgp_num 4214 pg_num_target 32 pgp_num_target 32 autoscale_mode on last_change 25817 lfor 0/25
815/25813 flags hashpspool,backfillfull stripe_width 0 application cephfs
```

```
# ceph osd pool autoscale-status
POOL                     SIZE  TARGET SIZE  RATE  RAW CAPACITY   RATIO  TARGET RATIO  EFFECTIVE RATIO
  BIAS  PG_NUM  NEW PG_NUM  AUTOSCALE  BULK
.mgr                   448.5k                3.0        12506G  0.0000
   1.0       1              on         False
rbd                    40985M                3.0        12506G  0.0096
   1.0     256              on         False
cephfs.cephfs.meta     102.9M                3.0        12506G  1.0000        4.0000           1.0000
   1.0   32768              on         False
cephfs.cephfs.data      1733G                3.0        12506G  0.4158
   1.0      32              on         False
```

From MGR and system logs:

```
Before upgrade:
```

```
2634769 Feb 11 00:48:44 f03-h02-000-r640 conmon[2849344]: debug 2022-02-11T00:48:44.028+0000 7f3ba
b474700  0 [pg_autoscaler INFO root] effective_target_ratio 0.0 0.0 0 13428844396544
```

After upgrade:

```
2022-02-11T00:57:14.734+0000 7f4ceec03000  0 ceph version 17.0.0-10315-ga00e8b31 (a00e8b315af02865
380634f8100dc7d18a18af4f) quincy (dev), process ceph-mgr, pid 7
2022-02-11T00:58:57.186+0000 7f4add690700  0 [pg_autoscaler INFO root] effective_target_ratio 0.0
4.0 0 13428844396544
```

Related issues:

Related to RADOS - Bug #54238: cephadm upgrade pacifc to quincy -> causing os...	**New**
Related to RADOS - Backport #54412: pacific:osd:add pg_num_max value	**Rejected**
Copied to RADOS - Backport #54526: pacific: cephadm upgrade pacific to quincy...	**Resolved**
Copied to RADOS - Backport #54527: quincy: cephadm upgrade pacific to quincy ...	**Resolved**

History

#1 - 02/11/2022 10:39 PM - Vikhyat Umrao

- Subject changed from cephadm upgrade pacific to quincy autoscaler is scaling pgs from 32 -> 32768 to cephadm upgrade pacific to quincy autoscaler is scaling pgs from 32 -> 32768 for cephfs meta pool

#2 - 02/11/2022 10:40 PM - Vikhyat Umrao

- Related to Bug #54238: cephadm upgrade pacifc to quincy -> causing osd's FULL/cascading failure added

#3 - 02/11/2022 10:48 PM - Vikhyat Umrao

The following path has MGR logs, Mon logs, Cluster logs, audit logs, and system logs.

```
/home/core/tracker54263
```

#4 - 02/15/2022 11:21 PM - Kamoltat Sirivadhna

In summary,
the root cause of the problem is after the upgrade to quincy, cephfs meta data pool was somehow given a 4.0 target _size_ratio. This should not happen when we only have 4 pools in the same root of the cluster, especially, when total_target_byte is also 0 for cephfs.cephfs.meta , it is guaranteed that effective ratio will be 1.0 for that of cephfs.cephfs.meta, hence it will take precedence over capacity_ratio and this means it will give cephfs.cephfs.meta the maximum number of PGs it is allow to give, in this case, 32768 PGs.

Here is a link to my findings:
https://docs.google.com/document/d/1lpNTXlrgtcQ6tQylHqfRHkeLijU5Af1xjkYa_u7ZmbY/edit#

#5 - 02/17/2022 05:00 PM - Kamoltat Sirivadhna

Update:

From the monitor sides of things of pool creation, target_size_ratio cannot be more than 1.0 or less than 0.0. As it was specified herein /src/mon/MonCommands.h,
therefore, We can rule out the possibility of `target_size_ratio` getting set off by the command `ceph osd pool create <pool-name> --target_size_ratio

<ratio>` However,
`ceph osd pool set <pool-name> target_size_ratio <ratio>` is able to set the target_size_ratio to be out of 0.0-1.0 range.

Note:

target_size_ratio can be more than 1.0 and the bound that was set during pool creation in /src/mon/MonCommands.h, should be changed.

#6 - 02/28/2022 03:50 PM - Kamoltat Sirivadhna

- Related to Backport #54412: pacific:osd:add pg_num_max value added

#7 - 03/02/2022 12:38 AM - Vikhyat Umrao

- Status changed from New to In Progress

- Pull request ID set to 45200

#8 - 03/02/2022 06:46 PM - Neha Ojha

- Status changed from In Progress to Fix Under Review

#9 - 03/02/2022 11:14 PM - Kamoltat Sirivadhna

Update:

After recreating the problem by tweaking the upgrade/pacific-x/parallel suite and adding additional logs, we conclude that the problem lies in the declaration of `opt_mapping` in src/osd/osd_types.cc. https://github.com/ceph/ceph/pull/44054 added PG_NUM_MAX to the middle of the list, which we found out that the order of the list is important and we should always add to the end of list to preserve the order of options during upgrade. For more information regarding bug analysis please see:
https://docs.google.com/document/d/10PJDwU2H7uY2o7_1lwTtQHUFoF9Fx7jguKeT-mKR2sA/edit?usp=sharing

#10 - 03/10/2022 09:57 PM - Kamoltat Sirivadhna

- Status changed from Fix Under Review to Pending Backport

- Backport set to quincy, pacific

#11 - 03/10/2022 10:00 PM - Backport Bot

- Copied to Backport #54526: pacific: cephadm upgrade pacific to quincy autoscaler is scaling pgs from 32 -> 32768 for cephfs meta pool added

#12 - 03/10/2022 10:00 PM - Backport Bot

- Copied to Backport #54527: quincy: cephadm upgrade pacific to quincy autoscaler is scaling pgs from 32 -> 32768 for cephfs meta pool added

#13 - 03/28/2022 03:48 PM - Yuri Weinstein

https://github.com/ceph/ceph/pull/45173 merged

#14 - 03/29/2022 06:45 PM - Kamoltat Sirivadhna

- Status changed from Pending Backport to Resolved