RADOS - Feature #42321

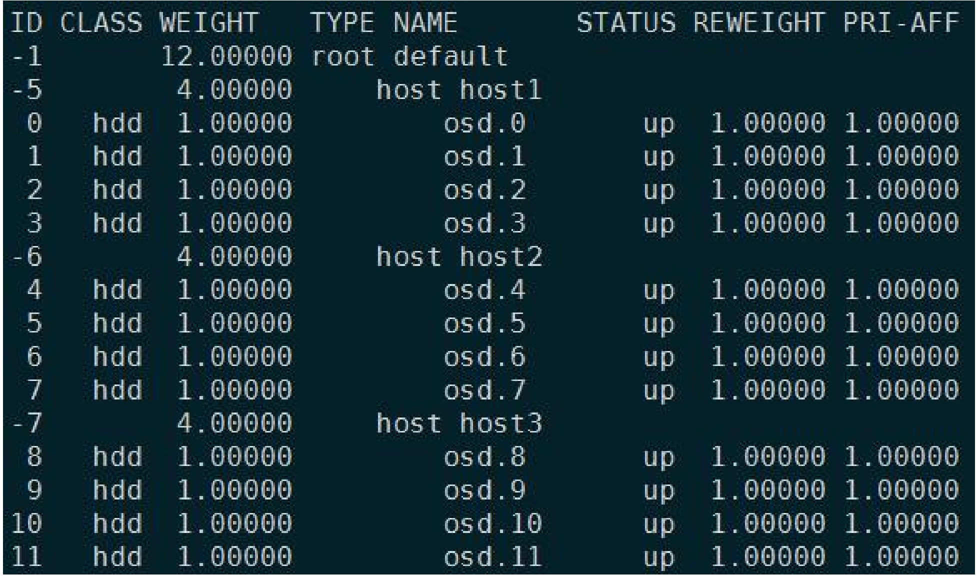
Add a new mode to balance pg layout by primary osds

10/15/2019 08:54 AM - rosin luo

Status:	Fix Under Review	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:			
Category:			
Target version:	v14.2.5		
Source:		**Affected Versions:**	
Tags:		**Component(RADOS):**	
Backport:	luminous,mimic,nautilus	**Pull request ID:**	30929
Reviewed:			

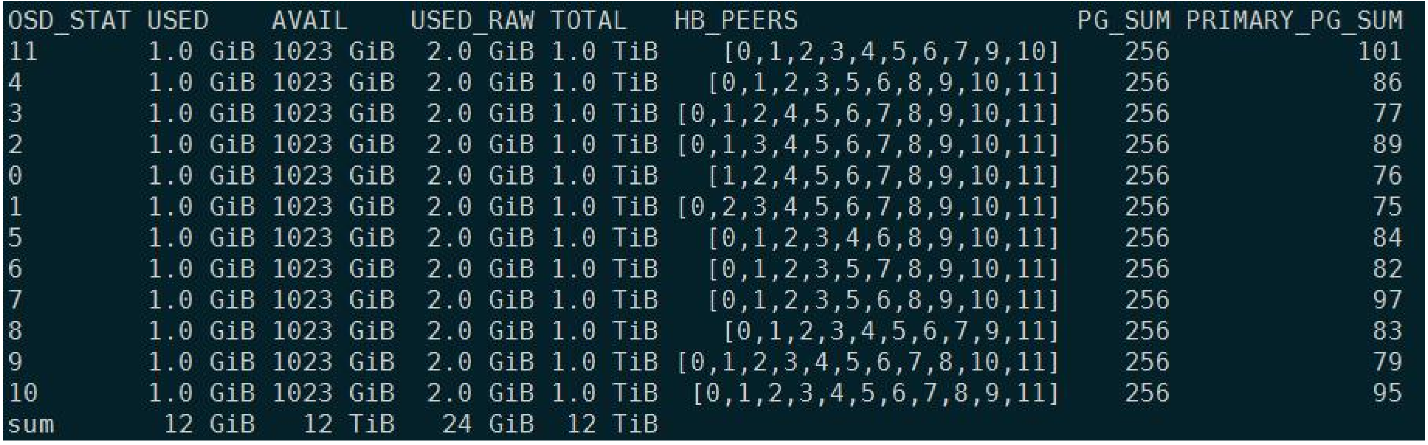
Description

There already have upmap optimizer since Luminous version. The upmap optimizer is help for balancing PGs across OSDs, it can get a “perfect” distribution, each OSD have equal number of PGs. But it is not balanced in primary PGs.
The upmap-by-primary-osd optimizer balance primary PG and replica PG in turn. The implementation of upmap-by-primary-osd refers to upmap. It’s behavior is just like upmap does to get a balanced distribution both primary PGs and total PGs. The optimizer balance PGs distribution in the same failure domain. As PG’s primary osd handles the read/write operations, the unbalanced OSDs result in unbalanced load. The OSD have more primary PGs will be the performance bottleneck especially for reading operation.We use fio to do 4M read test in rbd pools, it have about 20%-30% bandwidth improvement vs upmap.
We have a ceph cluster which contain 3 host,4 osds per host.We create a pool with 1024 pgs to do pg balance.
ceph osd tree looks like:



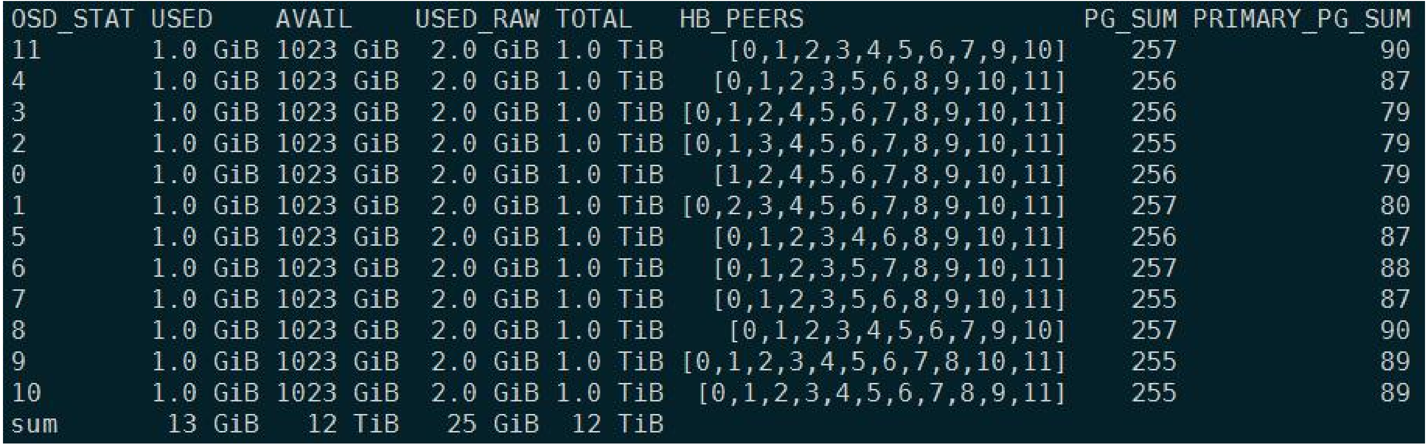
```
ID CLASS WEIGHT   TYPE NAME       STATUS REWEIGHT PRI-AFF
-1       12.00000 root default
-5        4.00000     host host1
 0   hdd  1.00000         osd.0       up  1.00000 1.00000
 1   hdd  1.00000         osd.1       up  1.00000 1.00000
 2   hdd  1.00000         osd.2       up  1.00000 1.00000
 3   hdd  1.00000         osd.3       up  1.00000 1.00000
-6        4.00000     host host2
 4   hdd  1.00000         osd.4       up  1.00000 1.00000
 5   hdd  1.00000         osd.5       up  1.00000 1.00000
 6   hdd  1.00000         osd.6       up  1.00000 1.00000
 7   hdd  1.00000         osd.7       up  1.00000 1.00000
-7        4.00000     host host3
 8   hdd  1.00000         osd.8       up  1.00000 1.00000
 9   hdd  1.00000         osd.9       up  1.00000 1.00000
10   hdd  1.00000         osd.10      up  1.00000 1.00000
11   hdd  1.00000         osd.11      up  1.00000 1.00000
```

The upmap optimizer to balance pg,result is blow:



```
OSD_STAT USED    AVAIL    USED_RAW TOTAL   HB_PEERS                  PG_SUM PRIMARY_PG_SUM
11        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB    [0,1,2,3,4,5,6,7,9,10]    256            101
4         1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB   [0,1,2,3,5,6,8,9,10,11]    256             86
3         1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB [0,1,2,4,5,6,7,8,9,10,11]    256             77
2         1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB [0,1,3,4,5,6,7,8,9,10,11]    256             89
0         1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB   [1,2,4,5,6,7,8,9,10,11]    256             76
1         1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB [0,2,3,4,5,6,7,8,9,10,11]    256             75
5         1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB   [0,1,2,3,4,6,8,9,10,11]    256             84
6         1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB   [0,1,2,3,5,7,8,9,10,11]    256             82
7         1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB   [0,1,2,3,5,6,8,9,10,11]    256             97
8         1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB    [0,1,2,3,4,5,6,7,9,11]    256             83
9         1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB [0,1,2,3,4,5,6,7,8,10,11]    256             79
10        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB  [0,1,2,3,4,5,6,7,8,9,11]    256             95
sum        12 GiB   12 TiB   24 GiB  12 TiB
```

The upmap-by-primary-osd optimizer to balance pg,result is blow pic,pg primary osds is not balanced between hosts, host1 has less primary pg and so osd0,osd1,osd2,osd3 has less primary pg nums.



```
OSD_STAT USED    AVAIL    USED_RAW TOTAL   HB_PEERS                    PG_SUM PRIMARY_PG_SUM
11       1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB    [0,1,2,3,4,5,6,7,9,10]    257             90
4        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB   [0,1,2,3,5,6,8,9,10,11]    256             87
3        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB [0,1,2,4,5,6,7,8,9,10,11]    256             79
2        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB [0,1,3,4,5,6,7,8,9,10,11]    255             79
0        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB   [1,2,4,5,6,7,8,9,10,11]    256             79
1        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB [0,2,3,4,5,6,7,8,9,10,11]    257             80
5        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB   [0,1,2,3,4,6,8,9,10,11]    256             87
6        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB   [0,1,2,3,5,7,8,9,10,11]    257             88
7        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB   [0,1,2,3,5,6,8,9,10,11]    255             87
8        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB    [0,1,2,3,4,5,6,7,9,10]    257             90
9        1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB [0,1,2,3,4,5,6,7,8,10,11]    255             89
10       1.0 GiB 1023 GiB  2.0 GiB 1.0 TiB  [0,1,2,3,4,5,6,7,8,9,11]    255             89
sum       13 GiB   12 TiB   25 GiB  12 TiB
```

The usage is just like upmap:
osdmaptool osdmap.file --upmap-by-primary-osd out.txt [--upmap-pool <pool>] [--upmap-max <max-count>] [--upmap-deviation <max-deviation>]

History

#1 - 10/21/2019 09:04 PM - Greg Farnum

- Project changed from Ceph to RADOS

- Category deleted (OSDMap)

- Status changed from New to Fix Under Review

Files

ceph_osd_tree.png	18 KB	10/15/2019	rosin luo
pg_balance_use_upmap_by_primary_osd.png	28 KB	10/15/2019	rosin luo
pg_balance_use_upmap.png	28.6 KB	10/15/2019	rosin luo