# Ceph - Bug #1636

## reweight-by-utilization does not choose good weights

10/20/2011 04:11 AM - pille palle

| | | | |
|---|---|---|---|
| **Status:** | Resolved | **% Done:** | 0% |
| **Priority:** | Normal | **Spent time:** | 0.00 hour |
| **Assignee:** | Josh Durgin | | |
| **Category:** | OSD | | |
| **Target version:** | v0.38 | | |
| **Source:** | | **Reviewed:** | |
| **Tags:** | | **Affected Versions:** | |
| **Backport:** | | **ceph-qa-suite:** | |
| **Regression:** | No | **Pull request ID:** | |
| **Severity:** | 3 - minor | **Crash signature:** | |

**Description**

there's a problem distributing the data evenly over all devices.
i'm using v0.36 and have a test setup with two hosts (each 7 devices).
i started with one host and added the second to test capacity increase.
i'm using a generated crushmap:

```
# begin crush map

# devices
device 0 device0
device 1 device1
device 2 device2
device 3 device3
device 4 device4
device 5 device5
device 6 device6
device 7 device7
device 8 device8
device 9 device9
device 10 device10
device 11 device11
device 12 device12
device 13 device13

# types
type 0 osd
type 1 domain
type 2 pool

# buckets
domain root {
        id -1           # do not change unnecessarily
        # weight 14.000
        alg straw
        hash 0  # rjenkins1
        item device0 weight 1.000
        item device1 weight 1.000
        item device2 weight 1.000
        item device3 weight 1.000
        item device4 weight 1.000
        item device5 weight 1.000
        item device6 weight 1.000
        item device7 weight 1.000
        item device8 weight 1.000
```

```
        item device9 weight 1.000
        item device10 weight 1.000
        item device11 weight 1.000
        item device12 weight 1.000
        item device13 weight 1.000
}

# rules
rule data {
        ruleset 0
        type replicated
        min_size 1
        max_size 10
        step take root
        step choose firstn 0 type osd
        step emit
}
rule metadata {
        ruleset 1
        type replicated
        min_size 1
        max_size 10
        step take root
        step choose firstn 0 type osd
        step emit
}
rule rbd {
        ruleset 2
        type replicated
        min_size 1
        max_size 10
        step take root
        step choose firstn 0 type osd
        step emit
}

# end crush map
```

then i disabled all OSDs on host1 (1-7) and waited for data-movement to settle.
after that everything was fine (all data accessable, nothing degraded), but the last OSD (which has more capacity than all the others) wasn't used for data.
i did multiple auto reweights, while copying more data, but OSD14 seems to be unused:

```
pille@mp2 ~ % ceph osd reweight-by-utilization
failed to open log file '/var/log/ceph/client.admin.log': error 13: Permission denied
2011-10-20 09:39:23.589350 mon <- [osd,reweight-by-utilization]
2011-10-20 09:39:23.590567 mon.0 -> 'SUCCESSFUL reweight-by-utilization: average_full: 0.214111, o
verload_full: 0.256933. overloaded osds: , 8 [0.577489], 9 [0.528660], 10 [0.612519], 11 [0.553863
], 12 [0.622734], 13 [0.515446]' (0)
pille@mp2 ~ % df -Ph |fgrep osd
/dev/mapper/vg--data-ceph--disk8   10G  5.8G  2.3G  72% /ceph/osd.8
/dev/mapper/vg--data-ceph--disk9   10G  6.1G  2.0G  76% /ceph/osd.9
/dev/mapper/vg--data-ceph--disk10   10G  5.5G  2.6G  69% /ceph/osd.10
/dev/mapper/vg--data-ceph--disk11   10G  5.9G  2.2G  74% /ceph/osd.11
/dev/mapper/vg--data-ceph--disk12   10G  5.4G  2.7G  68% /ceph/osd.12
/dev/mapper/vg--data-ceph--disk13   10G  6.2G  1.9G  78% /ceph/osd.13
/dev/mapper/vg--data-ceph--disk14  100G   19M   98G   1% /ceph/osd.14
pille@mp2 ~ % ceph osd dump
failed to open log file '/var/log/ceph/client.admin.log': error 13: Permission denied
2011-10-20 09:39:49.619388 mon <- [osd,dump]
2011-10-20 09:39:49.621114 mon.0 -> 'dumped osdmap epoch 789' (0)
epoch 789
fsid 0583d0bf-7324-f4e3-c147-27365578d8a8
```

```
created 2011-10-19 10:39:11.797594
modifed 2011-10-20 08:47:44.190611
flags

pg_pool 0 'data' pg_pool(rep pg_size 3 crush_ruleset 0 object_hash rjenkins pg_num 512 pgp_num 512
 lpg_num 2 lpgp_num 2 last_change 619 owner 0)
        removed_snaps [4~4]
pg_pool 1 'metadata' pg_pool(rep pg_size 5 crush_ruleset 1 object_hash rjenkins pg_num 512 pgp_num
 512 lpg_num 2 lpgp_num 2 last_change 615 owner 0)
pg_pool 2 'rbd' pg_pool(rep pg_size 2 crush_ruleset 2 object_hash rjenkins pg_num 512 pgp_num 512
lpg_num 2 lpgp_num 2 last_change 1 owner 0)

max_osd 15
osd.1 down out weight 0 up_from 400 up_thru 550 down_at 783 last_clean_interval 125-172 10.1.11.1:
6813/11012 10.1.11.1:6814/11012 10.1.11.1:6815/11012
osd.2 down in  weight 0.980072 up_from 127 up_thru 221 down_at 784 last_clean_interval 8-116 10.1.
11.1:6804/27250 10.1.11.1:6805/27250 10.1.11.1:6806/27250
osd.3 down out weight 0 up_from 129 up_thru 348 down_at 785 last_clean_interval 9-120 10.1.11.1:68
07/27394 10.1.11.1:6808/27394 10.1.11.1:6809/27394
osd.4 down in  weight 0.981155 up_from 132 up_thru 584 down_at 789 last_clean_interval 9-120 10.1.
11.1:6810/27799 10.1.11.1:6811/27799 10.1.11.1:6812/27799
osd.5 down in  weight 0.976013 up_from 179 up_thru 179 down_at 787 last_clean_interval 133-139 10.
1.11.1:6801/360 10.1.11.1:6802/360 10.1.11.1:6803/360
osd.6 down in  weight 0.99585 up_from 545 up_thru 548 down_at 786 last_clean_interval 132-207 10.1
.11.1:6816/18591 10.1.11.1:6817/18591 10.1.11.1:6818/18591
osd.7 down in  weight 0.998474 up_from 132 up_thru 506 down_at 788 last_clean_interval 26-121 10.1
.11.1:6819/28017 10.1.11.1:6820/28017 10.1.11.1:6821/28017
osd.8 up   in  weight 0.577484 up_from 58 up_thru 777 down_at 56 last_clean_interval 32-57 10.1.11
.2:6815/24370 10.1.11.2:6801/24370 10.1.11.2:6802/24370
osd.9 up   in  weight 0.528656 up_from 728 up_thru 740 down_at 718 last_clean_interval 32-702 10.1
.11.2:6814/25968 10.1.11.2:6818/25968 10.1.11.2:6820/25968
osd.10 up   in  weight 0.612518 up_from 726 up_thru 781 down_at 713 last_clean_interval 30-702 10.
1.11.2:6800/25605 10.1.11.2:6803/25605 10.1.11.2:6804/25605
osd.11 up   in  weight 0.553848 up_from 728 up_thru 773 down_at 718 last_clean_interval 32-702 10.
1.11.2:6805/25671 10.1.11.2:6806/25671 10.1.11.2:6807/25671
osd.12 up   in  weight 0.622726 up_from 728 up_thru 772 down_at 701 last_clean_interval 32-700 10.
1.11.2:6808/25739 10.1.11.2:6811/25739 10.1.11.2:6813/25739
osd.13 up   in  weight 0.515442 up_from 59 up_thru 772 down_at 56 last_clean_interval 32-58 10.1.1
1.2:6809/24130 10.1.11.2:6810/24130 10.1.11.2:6819/24130
osd.14 up   in  weight 1 up_from 57 up_thru 0 down_at 56 last_clean_interval 32-56 10.1.11.2:6812/
24214 10.1.11.2:6816/24214 10.1.11.2:6817/24214
```

after some time more and more OSDs of host2 where taken down (probably because they ran full):

```
pille@mp2 ~ % ceph osd dump
failed to open log file '/var/log/ceph/client.admin.log': error 13: Permission denied
2011-10-20 11:04:50.349637 mon <- [osd,dump]
2011-10-20 11:04:50.351508 mon.0 -> 'dumped osdmap epoch 805' (0)
epoch 805
fsid 0583d0bf-7324-f4e3-c147-27365578d8a8
created 2011-10-19 10:39:11.797594
modifed 2011-10-20 10:24:11.692047
flags

pg_pool 0 'data' pg_pool(rep pg_size 3 crush_ruleset 0 object_hash rjenkins pg_num 512 pgp_num 512
 lpg_num 2 lpgp_num 2 last_change 619 owner 0)
        removed_snaps [4~4]
pg_pool 1 'metadata' pg_pool(rep pg_size 5 crush_ruleset 1 object_hash rjenkins pg_num 512 pgp_num
 512 lpg_num 2 lpgp_num 2 last_change 615 owner 0)
pg_pool 2 'rbd' pg_pool(rep pg_size 2 crush_ruleset 2 object_hash rjenkins pg_num 512 pgp_num 512
lpg_num 2 lpgp_num 2 last_change 1 owner 0)

max_osd 15
```

```
osd.1 down out weight 0 up_from 400 up_thru 550 down_at 783 last_clean_interval 125-172 10.1.11.1:
6813/11012 10.1.11.1:6814/11012 10.1.11.1:6815/11012
osd.2 down out weight 0 up_from 127 up_thru 221 down_at 784 last_clean_interval 8-116 10.1.11.1:68
04/27250 10.1.11.1:6805/27250 10.1.11.1:6806/27250
osd.3 down out weight 0 up_from 129 up_thru 348 down_at 785 last_clean_interval 9-120 10.1.11.1:68
07/27394 10.1.11.1:6808/27394 10.1.11.1:6809/27394
osd.4 down out weight 0 up_from 132 up_thru 584 down_at 789 last_clean_interval 9-120 10.1.11.1:68
10/27799 10.1.11.1:6811/27799 10.1.11.1:6812/27799
osd.5 down out weight 0 up_from 179 up_thru 179 down_at 787 last_clean_interval 133-139 10.1.11.1:
6801/360 10.1.11.1:6802/360 10.1.11.1:6803/360
osd.6 down out weight 0 up_from 545 up_thru 548 down_at 786 last_clean_interval 132-207 10.1.11.1:
6816/18591 10.1.11.1:6817/18591 10.1.11.1:6818/18591
osd.7 down out weight 0 up_from 132 up_thru 506 down_at 788 last_clean_interval 26-121 10.1.11.1:6
819/28017 10.1.11.1:6820/28017 10.1.11.1:6821/28017
osd.8 down out weight 0 up_from 58 up_thru 794 down_at 798 last_clean_interval 32-57 10.1.11.2:681
5/24370 10.1.11.2:6801/24370 10.1.11.2:6802/24370
osd.9 down out weight 0 up_from 728 up_thru 740 down_at 798 last_clean_interval 32-702 10.1.11.2:6
814/25968 10.1.11.2:6818/25968 10.1.11.2:6820/25968
osd.10 down out weight 0 up_from 726 up_thru 790 down_at 794 last_clean_interval 30-702 10.1.11.2:
6800/25605 10.1.11.2:6803/25605 10.1.11.2:6804/25605
osd.11 up   in  weight 0.553848 up_from 728 up_thru 800 down_at 718 last_clean_interval 32-702 10.
1.11.2:6805/25671 10.1.11.2:6806/25671 10.1.11.2:6807/25671
osd.12 down out weight 0 up_from 728 up_thru 795 down_at 800 last_clean_interval 32-700 10.1.11.2:
6808/25739 10.1.11.2:6811/25739 10.1.11.2:6813/25739
osd.13 down out weight 0 up_from 59 up_thru 772 down_at 790 last_clean_interval 32-58 10.1.11.2:68
09/24130 10.1.11.2:6810/24130 10.1.11.2:6819/24130
osd.14 up   in  weight 1 up_from 57 up_thru 0 down_at 56 last_clean_interval 32-56 10.1.11.2:6812/
24214 10.1.11.2:6816/24214 10.1.11.2:6817/24214

pille@mp2 ~ % df -Ph |fgrep osd
/dev/mapper/vg--data-ceph--disk8   10G  7.7G  399M  96% /ceph/osd.8
/dev/mapper/vg--data-ceph--disk9   10G  8.1G  1.1M 100% /ceph/osd.9
/dev/mapper/vg--data-ceph--disk10   10G  7.3G  741M  91% /ceph/osd.10
/dev/mapper/vg--data-ceph--disk11   10G  7.9G  195M  98% /ceph/osd.11
/dev/mapper/vg--data-ceph--disk12   10G  7.3G  808M  91% /ceph/osd.12
/dev/mapper/vg--data-ceph--disk13   10G  8.1G  412K 100% /ceph/osd.13
/dev/mapper/vg--data-ceph--disk14  100G   19M   98G   1% /ceph/osd.14
```

now the cluster is degraded and my mountpoint hangs, while there's still plenty of capacity unused ;-(

### Associated revisions

**Revision f94a44e6 - 10/21/2011 12:20 AM - Josh Durgin**

OSDMonitor: reweight towards average utilization

The existing reweight-by-utilization calculation did not take into account the current weight of an OSD, and depended in part on the threshold given by the user. Also send the user both the old and new weights.

Fixes: #1636
Signed-off-by: Josh Durgin <josh.durgin@dreamhost.com>

### History

**#1 - 10/20/2011 04:18 PM - Josh Durgin**

*- Subject changed from distributing accross devices with different capacity to reweight-by-utilization does not choose good weights*

*- Category set to OSD*

*- Assignee set to Josh Durgin*

*- Target version set to v0.38*

#### #2 - 10/20/2011 05:45 PM - Josh Durgin

*- Status changed from New to Resolved*

The existing reweight-by-utilization code didn't make sense - f94a44e688883f2db0971435a5333a8b60c77dec fixes this.

When you have osds with different amounts of storage, you're better off setting the weights in the crushmap [[ http://permalink.gmane.org/gmane.comp.file-systems.ceph.devel/2858 like this]]. This way you don't have to wait until some osds fill up. reweight-by-utilization is meant more for correcting for unfortunate random distribution.