Linux kernel client - Bug #18807

I/O error on rbd device after adding new OSD to Crush map

02/03/2017 11:53 AM - Nikita Shalnov

Status:	Duplicate	**Start date:**	02/03/2017
Priority:	High	**Due date:**	
Assignee:	Ilya Dryomov	**% Done:**	0%
Category:	libceph	**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:		**Severity:**	2 - major
Tags:		**Reviewed:**	
Backport:		**Affected Versions:**	v10.2.3
Regression:	No	**ceph-qa-suite:**	

Description

Hello.

I run Ceph Jewel and KVM hypervisor.

```
ceph version 10.2.3 (ecc23778eb545d8dd55e2e4735b53cc93f92e65b)
qemu-kvm  1:2.1+dfsg-12+deb8u6
```

I have a virtual machine **test-ceph-13.g01.i-free.ru**, which uses one mapped rbd image.

```
root@test-hoster-kvm-buffer-01a:~# rbd showmapped
id pool        image                          snap device
0  rbdkvm_sata test-ceph-13.g01.i-free.ru-var -    /dev/rbd0
```

On guest a disk is mounted on /var.

My ceph osd tree looks like:

```
root@test-storage-ceph-01x:~# ceph osd tree
ID WEIGHT   TYPE NAME                     UP/DOWN REWEIGHT PRIMARY-AFFINITY
-1 21.78587 root default
-2 10.89294     host test-storage-ceph-01y
 0  1.81549         osd.0                      up  1.00000          1.00000
 2  1.81549         osd.2                      up  1.00000          1.00000
 4  1.81549         osd.4                      up  1.00000          1.00000
 6  1.81549         osd.6                      up  1.00000          1.00000
 8  1.81549         osd.8                      up  1.00000          1.00000
10  1.81549         osd.10                     up  1.00000          1.00000
-3 10.89293     host test-storage-ceph-01x
 3  1.81549         osd.3                      up  1.00000          1.00000
 5  1.81549         osd.5                      up  1.00000          1.00000
 7  1.81549         osd.7                      up  1.00000          1.00000
 9  1.81549         osd.9                      up  1.00000          1.00000
11  1.81549         osd.11                     up  1.00000          1.00000
 1  1.81548         osd.1                      up  1.00000          1.00000
```

Status of a cluster:

```
root@test-storage-ceph-01x:~# ceph -s
    cluster 63b92a66-8523-4adc-9e3a-ee267e5be456
```

```
     health HEALTH_OK
     monmap e1: 3 mons at {test-hoster-kvm-buffer-01a=192.168.103.89:6789/0,test-hoster-kvm-buffer
-01b=192.168.103.80:6789/0,test-hoster-kvm-buffer-01e=192.168.103.22:6789/0}
            election epoch 58, quorum 0,1,2 test-hoster-kvm-buffer-01e,test-hoster-kvm-buffer-01b,
test-hoster-kvm-buffer-01a
     osdmap e1705: 12 osds: 12 up, 12 in
            flags sortbitwise
      pgmap v8540009: 1024 pgs, 1 pools, 4064 GB data, 868 kobjects
            8132 GB used, 14176 GB / 22309 GB avail
                1024 active+clean
  client io 630 kB/s rd, 30977 kB/s wr, 1317 op/s rd, 1620 op/s wr
```

I remove osd.1 from the cluster and Crush map:

```
systemctl stop ceph-osd@1.service; ceph osd rm osd.1; ceph osd crush remove osd.1
removed osd.1
removed item id 1 name 'osd.1' from crush map
```

Recovery begins. osd.1 disappears from cluster and map.
I can still read/write to my mapped disk.

Then I start osd.1 but don`t add it to Crush:

```
systemctl start ceph-osd@1.service
```

osd.1 is registering by itself.

```
root@test-storage-ceph-01x:~# ceph -s
    cluster 63b92a66-8523-4adc-9e3a-ee267e5be456
     health HEALTH_WARN
            249 pgs backfill_wait
            8 pgs backfilling
            170 pgs degraded
            30 pgs stuck unclean
            170 pgs undersized
            recovery 147154/1886901 objects degraded (7.799%)
            recovery 357258/1886901 objects misplaced (18.934%)
     monmap e1: 3 mons at {test-hoster-kvm-buffer-01a=192.168.103.89:6789/0,test-hoster-kvm-buffer
-01b=192.168.103.80:6789/0,test-hoster-kvm-buffer-01e=192.168.103.22:6789/0}
            election epoch 58, quorum 0,1,2 test-hoster-kvm-buffer-01e,test-hoster-kvm-buffer-01b,
test-hoster-kvm-buffer-01a
     osdmap e1712: 12 osds: 12 up, 12 in; 257 remapped pgs
            flags sortbitwise
      pgmap v8541913: 1024 pgs, 1 pools, 4064 GB data, 869 kobjects
            8151 GB used, 14157 GB / 22309 GB avail
            147154/1886901 objects degraded (7.799%)
            357258/1886901 objects misplaced (18.934%)
                 767 active+clean
                 170 active+undersized+degraded+remapped+wait_backfill
                  79 active+remapped+wait_backfill
                   8 active+remapped+backfilling
recovery io 190 MB/s, 40 objects/s
  client io 138 kB/s rd, 14709 kB/s wr, 377 op/s rd, 715 op/s wr
```

I can still read/write to my mapped disk.
Now I add this osd back to the Crush:

```
root@test-storage-ceph-01x:~# ceph osd crush add osd.1 1.81549 host=test-storage-ceph-01x
add item id 1 name 'osd.1' weight 1.81549 at location {host=test-storage-ceph-01x} to crush map
```

And I get I/O Errors on the mapped disk instantly (kern.log from hoster below):

```
Feb  3 15:29:39 test-hoster-kvm-buffer-01a kernel: libceph: osd1 down
Feb  3 15:32:41 test-hoster-kvm-buffer-01a kernel: libceph: osd1 up
Feb  3 15:32:41 test-hoster-kvm-buffer-01a kernel: libceph: osd1 weight 0x10000 (in)
Feb  3 15:36:07 test-hoster-kvm-buffer-01a kernel: rbd: rbd0: write 6e000 at a9c00000 (0)\x0a
Feb  3 15:36:07 test-hoster-kvm-buffer-01a kernel: rbd: rbd0:   result -6 xferred 6e000\x0a
Feb  3 15:36:07 test-hoster-kvm-buffer-01a kernel: end_request: I/O error, dev rbd0, sector 556236
8
Feb  3 15:36:07 test-hoster-kvm-buffer-01a kernel: rbd: rbd0: write 80000 at a9c6e000 (6e000)\x0a
Feb  3 15:36:07 test-hoster-kvm-buffer-01a kernel: rbd: rbd0:   result -6 xferred 80000\x0a
Feb  3 15:36:07 test-hoster-kvm-buffer-01a kernel: end_request: I/O error, dev rbd0, sector 556324
8
```

dmesg from client:

```
[ 2543.766483] end_request: I/O error, dev vdc, sector 5563248
[ 2543.767832] EXT4-fs warning (device vdc): ext4_end_bio:317: I/O error -5 writing to inode 13368
0 (offset 226492416 size 8388608 starting block 695532)
[ 2543.767837] Buffer I/O error on device vdc, logical block 695406
[ 2543.768863] Buffer I/O error on device vdc, logical block 695407
[ 2543.769778] Buffer I/O error on device vdc, logical block 695408
[ 2543.770473] Buffer I/O error on device vdc, logical block 695409
[ 2543.770473] Buffer I/O error on device vdc, logical block 695410
[ 2543.770473] Buffer I/O error on device vdc, logical block 695411
[ 2543.770473] Buffer I/O error on device vdc, logical block 695412
[ 2543.770473] Buffer I/O error on device vdc, logical block 695413
[ 2543.770473] Buffer I/O error on device vdc, logical block 695414
[ 2543.770473] Buffer I/O error on device vdc, logical block 695415
[ 2543.776123] end_request: I/O error, dev vdc, sector 5564256
[ 2543.777018] EXT4-fs warning (device vdc): ext4_end_bio:317: I/O error -5 writing to inode 13368
0 (offset 226492416 size 8388608 starting block 695658)
[ 2543.777086] end_request: I/O error, dev vdc, sector 5565264
[ 2543.777776] EXT4-fs warning (device vdc): ext4_end_bio:317: I/O error -5 writing to inode 13368
0 (offset 226492416 size 8388608 starting block 695784)
[ 2543.777831] end_request: I/O error, dev vdc, sector 5566272

.......(skipping)
[ 2554.794988] EXT4-fs (vdc): This should not happen!! Data will be lost

[ 2554.812962] EXT4-fs error (device vdc) in ext4_writepages:2580: Journal has aborted
[ 2554.876906] EXT4-fs (vdc): Delayed block allocation failed for inode 133680 at logical offset 1
67937 with max blocks 2048 with error 30
[ 2554.878468] EXT4-fs (vdc): This should not happen!! Data will be lost

[ 2554.879626] EXT4-fs error (device vdc) in ext4_writepages:2580: Journal has aborted
[ 2584.927488] EXT4-fs error (device vdc): ext4_journal_check_start:56: Detected aborted journal
[ 2584.930459] EXT4-fs (vdc): Remounting filesystem read-only
[ 2584.971757] EXT4-fs (vdc): ext4_writepages: jbd2_start: 5120 pages, ino 133680; err -30
```

The moment when I added osd.1 to cluster:

```
2017-02-03 14:25:56.763297 mon.1 [INF] from='client.? 192.168.103.1:0/1194126355' entity='client.a
dmin' cmd=[{"prefix": "osd crush add", "args": ["host=test-storage-ceph-01x"], "id": 1, "weight":
1.81549}]: dispatch
2017-02-03 14:25:56.764477 mon.0 [INF] from='client.8535301 :/0' entity='client.admin' cmd=[{"pref
ix": "osd crush add", "args": ["host=test-storage-ceph-01x"], "id": 1, "weight": 1.81549}]: dispat
ch
2017-02-03 14:25:57.597892 mon.0 [INF] pgmap v8542082: 1024 pgs: 170 active+undersized+degraded+re
mapped+wait_backfill, 6 active+remapped+backfilling, 76 active+remapped+wait_backfill, 772 active+
```

```
clean; 4064 GB data, 8144 GB used, 14164 GB / 22309 GB avail; 2294 kB/s rd, 38304 kB/s wr, 1447 op
/s; 147153/1882465 objects degraded (7.817%); 350032/1882465 objects misplaced (18.594%); 176 MB/s
, 36 objects/s recovering
2017-02-03 14:25:57.695348 mon.0 [INF] from='client.8535301 :/0' entity='client.admin' cmd='[{"pre
fix": "osd crush add", "args": ["host=test-storage-ceph-01x"], "id": 1, "weight": 1.81549}]': fini
shed
2017-02-03 14:25:57.719138 mon.0 [INF] osdmap e1723: 12 osds: 12 up, 12 in
2017-02-03 14:25:57.769683 mon.0 [INF] pgmap v8542083: 1024 pgs: 170 active+undersized+degraded+re
mapped+wait_backfill, 6 active+remapped+backfilling, 76 active+remapped+wait_backfill, 772 active+
clean; 4064 GB data, 8144 GB used, 14164 GB / 22309 GB avail; 523 kB/s rd, 23070 kB/s wr, 891 op/s
; 147153/1882465 objects degraded (7.817%); 350032/1882465 objects misplaced (18.594%)
2017-02-03 14:25:58.929068 mon.0 [INF] osdmap e1724: 12 osds: 12 up, 12 in
2017-02-03 14:25:58.953220 mon.0 [INF] pgmap v8542084: 1024 pgs: 13 peering, 170 active+undersized
+degraded+remapped+wait_backfill, 5 active+remapped+backfilling, 64 active+remapped+wait_backfill,
 772 active+clean; 4064 GB data, 8144 GB used, 14164 GB / 22309 GB avail; 7137 B/s rd, 12401 kB/s
wr, 240 op/s; 147153/1867074 objects degraded (7.881%); 319996/1867074 objects misplaced (17.139%)
; 51393 kB/s, 8 objects/s recovering
2017-02-03 14:25:59.003858 mon.0 [INF] pgmap v8542085: 1024 pgs: 13 peering, 170 active+undersized
+degraded+remapped+wait_backfill, 5 active+remapped+backfilling, 64 active+remapped+wait_backfill,
 772 active+clean; 4064 GB data, 8144 GB used, 14164 GB / 22309 GB avail; 8295 B/s rd, 14414 kB/s
wr, 279 op/s; 147153/1867074 objects degraded (7.881%); 319996/1867074 objects misplaced (17.139%)
; 59736 kB/s, 9 objects/s recovering
2017-02-03 14:25:58.932489 osd.9 [INF] 6.4f starting backfill to osd.6 from (0'0,0'0] MIN to 1722'
5041303
2017-02-03 14:26:00.196595 mon.0 [INF] osdmap e1725: 12 osds: 12 up, 12 in
```

I can reproduce this many times. After that I rebooted my VM and it went in emergency mode. I had to destroy the VM and unmap rbd-device.

```
root@test-hoster-kvm-buffer-01a:~# virsh destroy test-ceph-13.g01.i-free.ru
Domain test-ceph-13.g01.i-free.ru destroyed

root@test-hoster-kvm-buffer-01a:~# rbd unmap /dev/rbd0
```

Then I could map it again and start the VM. Only after these actions VM became available.

This happens ONLY AFTER adding new osd to Crush map.
Could you please explain what happens and how can I avoid this behavior?
Thank you.

Tell me if you need more info.

Related issues:

Duplicates Linux kernel client - Bug #14901: misdirected requests on 4.2 duri...	**Resolved**	**02/26/2016**

History

#1 - 02/03/2017 11:58 AM - Nikita Shalnov

Distributor ID: Debian
Description: Debian GNU/Linux 8.6 (jessie)
Release: 8.6
Codename: jessie

#2 - 02/03/2017 01:56 PM - Jason Dillaman

- Project changed from rbd to Linux kernel client

- Subject changed from rbd map: I/O error on rbd device after adding new OSD to Crush map to I/O error on rbd device after adding new OSD to Crush map

#3 - 02/06/2017 12:49 PM - Ilya Dryomov

- Assignee set to Ilya Dryomov

Hi Nikita,

Which kernel are you running on test-hoster-kvm-buffer-01a?

#4 - 02/06/2017 12:50 PM - Ilya Dryomov

- Category set to libceph

#5 - 02/06/2017 01:18 PM - Nikita Shalnov

Hi Ilya,

I don't completely understand, what do you need - either a version of kernel of the VM or of the hoster, so I would give you both:

Hoster: Linux test-hoster-kvm-buffer-01a 3.16.0-4-amd64 #1 SMP Debian 3.16.36-1+deb8u1 (2016-09-03) x86_64 GNU/Linux

VM: Linux test-ceph-13 3.16.0-4-amd64 #1 SMP Debian 3.16.7-ckt25-1 (2016-03-06) x86_64 GNU/Linux

#6 - 02/06/2017 01:42 PM - Ilya Dryomov

I wanted the host kernel -- looks like it's 3.16.36. This should be fixed in 3.16.39, which is in jessie AFAICT. Could you please try the upgraded kernel and report back?

#7 - 02/06/2017 03:22 PM - Nikita Shalnov

Yes, I can. I will try it.
Thank you.

#8 - 02/07/2017 08:41 AM - Nikita Shalnov

Hi Ilya.

I have upgraded the kernel and it looks like, the bug was fixed - I can't reproduce it.

Thank you.

#9 - 02/07/2017 09:07 AM - Ilya Dryomov

- Duplicates Bug #14901: misdirected requests on 4.2 during rebalancing added

#10 - 02/07/2017 09:08 AM - Ilya Dryomov

- Status changed from New to Duplicate