Linux kernel client - Bug #1086

rbd: iozone failure

05/12/2011 10:00 AM - Sage Weil

Status:	Resolved	**Start date:**	05/12/2011
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:	rbd	**Estimated time:**	0.00 hour
Target version:	v2.6.39	**Spent time:**	0.00 hour
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Crash signature:**	
Severity:	3 - minor		

Description

I was able to reproduce Fyodor's problem on rbd (latest kernel) and ext2:

```
root@uml:~# mke2fs /dev/rbd0
mke2fs 1.41.12 (17-May-2010)
Filesystem label=
OS type: Linux
Block size=4096 (log=2)
Fragment size=4096 (log=2)
Stride=0 blocks, Stripe width=0 blocks
6406144 inodes, 25600000 blocks
1280000 blocks (5.00%) reserved for the super user
First data block=0
Maximum filesystem blocks=4294967296
782 block groups
32768 blocks per group, 32768 fragments per group
8192 inodes per group
Superblock backups stored on blocks: 
        32768, 98304, 163840, 229376, 294912, 819200, 884736, 1605632, 2654208, 
        4096000, 7962624, 11239424, 20480000, 23887872

Writing inode tables: done                            
Writing superblocks and filesystem accounting information: done

This filesystem will be automatically checked every 39 mounts or
180 days, whichever comes first.  Use tune2fs -c or -i to override.
root@uml:~# mount /dev/rbd0 mnt
cd mnt
root@uml:~# cd mnt
root@uml:~/mnt# iozone -a -n4g -g20g
        Iozone: Performance Test of File I/O
                Version $Revision: 3.308 $
                Compiled for 64 bit mode.
                Build: linux 

        Contributors:William Norcott, Don Capps, Isom Crawford, Kirby Collins
                     Al Slater, Scott Rhine, Mike Wisner, Ken Goss
                     Steve Landherr, Brad Smith, Mark Kelly, Dr. Alain CYR,
                     Randy Dunlap, Mark Montague, Dan Million, Gavin Brebner,
                     Jean-Marc Zucconi, Jeff Blomberg, Benny Halevy,
                     Erik Habbinga, Kris Strecker, Walter Wong, Joshua Root.

        Run began: Thu May 12 16:20:59 2011

        Auto Mode
        Using minimum file size of 4194304 kilobytes.
        Using maximum file size of 20971520 kilobytes.
        Command line used: iozone -a -n4g -g20g
        Output is in Kbytes/sec
        Time Resolution = 0.000001 seconds.
        Processor cache size set to 1024 Kbytes.
        Processor cache line size set to 32 bytes.
        File stride size set to 17 * record size.
                                                            random  random    bkwd   record   stri
de
              KB  reclen   write rewrite    read    reread    read   write    read  rewrite     re
ad   fwrite frewrite   fread  freread
         4194304      64    3570    3040    12640    13725

Error in file: Found ?aaaaaaaaaaaaaaaa? Expecting ?3838383838383838? addr 40a00000
Error in file: Position 2813329408
Record # 42928 Record size 64 kb
where 40a00000 loop 0
```

Not sure yet if this is an osd or rbd problem.

History

#1 - 05/12/2011 10:06 AM - Sage Weil

strangely, the file looks correct (before and after a remount):

```
root@uml:~/mnt# dd if=iozone.tmp of=/tmp/foo skip=2813329408 bs=1 count=1000
1000+0 records in
1000+0 records out
1000 bytes (1.0 kB) copied, 0.028639 s, 34.9 kB/s
root@uml:~/mnt# hexdump -C /tmp/foo
00000000  38 38 38 38 38 38 38 38  00 00 00 00 00 00 00 00  |88888888........|
00000010  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
*
000003e0
```

#2 - 05/12/2011 11:12 AM - Fyodor Ustinov

My comment.

I have the starting file size 4G (-n4g) because on this server 4G memory. Into smaller files on the server, this error does not occur. After reducing the memory up to 2G problem began to emerge on the file 2G.

#3 - 05/13/2011 06:29 AM - Yehuda Sadeh

The problem is that our use of blk_end_request is wrong, as it assumes ordering on the requests completion. In most requests there's just a single rados operation so there's no issue, but when requests cross rados objects then we have multiple requests and there's a chance that the responses will arrive in a different order.

#4 - 05/13/2011 09:10 AM - Sage Weil

i just noticed this comment didn't post yesterday, no wonder yehuda didn't know what i was talking about :)

here:

for my second failure, the last two osd ops were spanning an object boundary

```
2011-05-12 11:30:14.931086 7f784a7d0710 -- 10.0.1.252:6801/26928 <== client4311 10.0.1.219:0/2123522985 430307
 ==== osd_op(client4311.1:430307 rb.0.0.00000000000c [read 4161536~32768] 3.50cf) ==== 127+0+0 (810368824 0 0)
 0x7404240 con 0x1bf7dc0
2011-05-12 11:30:14.969848 7f784a7d0710 -- 10.0.1.252:6801/26928 <== client4311 10.0.1.219:0/2123522985 430308
 ==== osd_op(client4311.1:430308 rb.0.0.00000000000d [read 0~32768] 3.734e) ==== 127+0+0 (4216669042 0 0) 0x33
0a240 con 0x1bf7dc0
2011-05-12 11:30:14.970464 7f78486cb710 -- 10.0.1.252:6801/26928 --> 10.0.1.219:0/2123522985 -- osd_op_reply(4
30308 rb.0.0.00000000000d [read 0~32768] = 0) v1 -- ?+32768 0x84bc8c0 con 0x1bf7dc0
2011-05-12 11:30:15.071018 7f7847eca710 -- 10.0.1.252:6801/26928 --> 10.0.1.219:0/2123522985 -- osd_op_reply(4
30307 rb.0.0.00000000000c [read 4161536~32768] = 0) v1 -- ?+32768 0x9d6c540 con 0x1bf7dc0
```

when i unmounted, remounted, and re-read that data, i saw

```
2011-05-12 11:55:00.029320 7f7847eca710 -- 10.0.1.252:6801/26928 --> 10.0.1.219:0/2123522985 -- osd_op_reply(4
30354 rb.0.0.00000000000c [read 4161536~4096] = 0) v1 -- ?+4096 0x9deca80 con 0x34653c0
```

(and got the correct data)

So... I wonder if the read over a stripe boundary could be broken somehow?

Notably, reading that same file range *after* iozone failed also returned correct data. So maybe it's a race where rbd is completing the IO to the upper layers PRIOR to the second piece coming in?

#5 - 05/13/2011 01:51 PM - Sage Weil

- Project changed from Ceph to Linux kernel client

- Target version deleted (v0.29)

#6 - 05/13/2011 01:51 PM - Sage Weil

- Category set to rbd

- Target version set to v2.6.39

fixed by commit:1fec70932d867416ffe620dd17005f168cc84eb5

#7 - 05/13/2011 01:51 PM - Sage Weil

- Status changed from New to Resolved