bluestore - Bug #36482

High amount of Read I/O on BlueFS/DB when listing omap keys

10/17/2018 11:54 AM - Wido den Hollander

Status:	Resolved	**% Done:**	0%
Priority:	High		
Assignee:	Igor Fedotov		
Category:			
Target version:			
Source:	Community (dev)	**Affected Versions:**	v12.2.4, v12.2.5, v12.2.8
Tags:	osd,bluestore,bluefs,read_random,omap, rbd	**ceph-qa-suite:**	
Backport:	nautilus	**Pull request ID:**	27627
Regression:	No	**Crash signature (v1):**	
Severity:	2 - major	**Crash signature (v2):**	
Reviewed:			

Description

I don't know how to describe this issue the best, but I've been observing various issues with Luminous 12.2.4 ~ 12.2.8 OSDs.

It started with the stupidallocator dumping messages as described here:
http://lists.ceph.com/pipermail/ceph-users-ceph.com/2018-October/030546.html

```
2018-10-10 21:52:04.019037 7f90c2f0f700  0 stupidalloc 0x0x55828ae047d0
dump  0x15cd2078000~34000
2018-10-10 21:52:04.019038 7f90c2f0f700  0 stupidalloc 0x0x55828ae047d0
dump  0x15cd22cc000~24000
2018-10-10 21:52:04.019038 7f90c2f0f700  0 stupidalloc 0x0x55828ae047d0
dump  0x15cd2300000~20000
2018-10-10 21:52:04.019039 7f90c2f0f700  0 stupidalloc 0x0x55828ae047d0
dump  0x15cd2324000~24000
2018-10-10 21:52:04.019040 7f90c2f0f700  0 stupidalloc 0x0x55828ae047d0
dump  0x15cd26c0000~24000
2018-10-10 21:52:04.019041 7f90c2f0f700  0 stupidalloc 0x0x55828ae047d0
dump  0x15cd2704000~30000
```

After we offloaded data (CRUSH migration) to other OSDs these messages went away.

A few days later we observed OSDs utilizing their disks (Samsung PM863a 1.92TB SSDs) for ~90% and reading a lot from them, like this:

```
Device:         rrqm/s   wrqm/s     r/s     w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await r_awai
t w_await  svctm  %util
loop0             0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00    0.0
0    0.00   0.00   0.00
sda               0.00     0.00   13.00   43.00    60.00   236.00    10.57     0.00    0.00    0.0
0    0.00   0.00   0.00
sdb               0.00     4.00   12.00   98.00    52.00   820.00    15.85     0.00    0.04    0.0
0    0.04   0.04   0.40
sdc               0.00     1.00   25.00  145.00   120.00  1336.00    17.13     0.04    0.24    0.0
0    0.28   0.12   2.00
sdd               0.00     7.00   27.00  117.00   112.00  1076.00    16.50     0.01    0.06    0.0
0    0.07   0.06   0.80
sde               0.00     0.00 6839.00   69.00 53464.00   388.00    15.59     0.93    0.13    0.1
4    0.12   0.13  92.80
sdf               0.00     0.00   17.00  102.00   192.00   756.00    15.93     0.00    0.03    0.2
```

```
4    0.00   0.03   0.40
sdg               0.00     0.00   11.00   89.00    44.00   516.00    11.20     0.02    0.20    0.0
0    0.22   0.12   1.20
sdh               0.00     0.00   11.00  121.00    68.00   676.00    11.27     0.01    0.06    0.0
0    0.07   0.06   0.80
sdi               0.00    11.00    0.00    5.00     0.00    64.00    25.60     0.00    0.80    0.0
0    0.80   0.80   0.40
```

/dev/sde is osd.246 in this case and it's showing this in it's logs:

```
2018-10-17 13:32:09.050155 7f54713d7700  1 heartbeat_map is_healthy 'OSD::osd_op_tp thread 0x7f545
31aa700' had timed out after 60
2018-10-17 13:32:09.050160 7f54713d7700  1 heartbeat_map is_healthy 'OSD::osd_op_tp thread 0x7f545
71b2700' had timed out after 60
2018-10-17 13:32:09.050714 7f54713d7700  1 heartbeat_map is_healthy 'OSD::osd_op_tp thread 0x7f545
31aa700' had timed out after 60
2018-10-17 13:32:09.050719 7f54713d7700  1 heartbeat_map is_healthy 'OSD::osd_op_tp thread 0x7f545
71b2700' had timed out after 60
2018-10-17 13:32:09.073876 7f5470bd6700  1 heartbeat_map is_healthy 'OSD::osd_op_tp thread 0x7f545
31aa700' had timed out after 60
```

I increased this value from 15 to 60 as the OSDs were committing suicide.

Setting debug_bluefs to 10 shows me:

```
2018-10-17 13:33:10.680782 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x235c56c~ecf from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
2018-10-17 13:33:10.680932 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x235d43b~f75 from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
2018-10-17 13:33:10.681075 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x235e3b0~eb7 from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
2018-10-17 13:33:10.681229 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x235f267~e8f from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
2018-10-17 13:33:10.681370 7f54571b2700 10 bluefs _read_random h 0x55b695996b80 0x3782d9f~fb8 from
 file(ino 214919 size 0x445eac0 mtime 2018-10-16 01:55:01.588751 bdev 1 allocated 4500000 extents
[1:0xe3f4b00000+4500000])
2018-10-17 13:33:10.681523 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x23600f6~ea6 from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
2018-10-17 13:33:10.681654 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x2360f9c~edd from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
2018-10-17 13:33:10.681798 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x2361e79~f57 from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
2018-10-17 13:33:10.681940 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x2362dd0~ed1 from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
2018-10-17 13:33:10.682088 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x2363ca1~f7b from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
2018-10-17 13:33:10.682232 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x2364c1c~f1c from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
2018-10-17 13:33:10.682393 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x2365b38~f48 from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
```

```
2018-10-17 13:33:10.682551 7f54571b2700 10 bluefs _read_random h 0x55b695997100 0x2366a80~fdd from
 file(ino 211403 size 0x41af2a4 mtime 2018-10-15 18:41:40.702314 bdev 1 allocated 4200000 extents
[1:0xe753f00000+4200000])
```

I spend a few hours with Igor from SUSE to debug this problem as it is showing up on multiple OSDs.

We debugged with osd.240 (not 246) and we could trigger the problem using this command:

```
root@mon01:~# time rados -p rbd-ssd-c03-p02 listomapkeys rbd_header.0904ed238e1f29
features
object_prefix
order
size
snap_seq
snapshot_0000000000000431

real    2m16.678s
user    0m0.064s
sys    0m0.028s
root@mon01:~#
```

As you can see, it takes 2 minutes (!!) to list the omap values for this RBD header. In these two minutes the OSD becomes unusable and causes massive slow requests as it's simply blocking.

The image in this case:

```
root@mon01:~# rbd -p rbd-ssd-c03-p02 info fa2d5398-08de-4f9b-b519-28e6258bc5f3
rbd image 'fa2d5398-08de-4f9b-b519-28e6258bc5f3':
    size 50GiB in 12800 objects
    order 22 (4MiB objects)
    block_name_prefix: rbd_data.0904ed238e1f29
    format: 2
    features: layering, exclusive-lock, object-map, fast-diff, deep-flatten
    flags:
root@mon01:~#
```

```
root@mon01:~# ceph osd map rbd-ssd-c03-p02 rbd_header.0904ed238e1f29
osdmap e623992 pool 'rbd-ssd-c03-p02' (29) object 'rbd_header.0904ed238e1f29' -> pg 29.9fd75a23 (2
9.223) -> up ([240,183,176], p240) acting ([240,183,176], p240)
root@mon01:~#
```

Doing this triggers the OSD (240) to be jump to 90% util on it's SSD and cause massive slow/blocked requests for more then 60 seconds.

In the logs this is shown:

```
2018-10-16 21:07:15.112839 7f37d26f2700 10 bluestore(/var/lib/ceph/osd/ceph-240) get_omap_iterator
 29.223_head #29:c45aebf9:::rbd_header.0904ed238e1f29:head#
2018-10-16 21:07:15.112840 7f37d26f2700 10 bluestore(/var/lib/ceph/osd/ceph-240) get_omap_iterator
 has_omap = 1
2018-10-16 21:07:15.112904 7f37d26f2700 10 bluefs _read_random h 0x55ddaca32680 0x22df502~f09 from
 file(ino 201593 size 0x41a812d mtime 2018-10-16 11:22:59.508274 bdev 1 allocated 4200000 extents
[1:0xe048900000+4200000])
2018-10-16 21:07:15.113079 7f37d26f2700 10 bluefs _read_random h 0x55ddaca32680 0x22e040b~efd from
 file(ino 201593 size 0x41a812d mtime 2018-10-16 11:22:59.508274 bdev 1 allocated 4200000 extents
```

```
[1:0xe048900000+4200000])
2018-10-16 21:07:15.113246 7f37d26f2700 10 bluefs _read_random h 0x55ddaca32680 0x22e1308~fa2 from
 file(ino 201593 size 0x41a812d mtime 2018-10-16 11:22:59.508274 bdev 1 allocated 4200000 extents
[1:0xe048900000+4200000])
2018-10-16 21:07:15.113414 7f37d26f2700 10 bluefs _read_random h 0x55ddaca32680 0x22e22aa~f37 from
 file(ino 201593 size 0x41a812d mtime 2018-10-16 11:22:59.508274 bdev 1 allocated 4200000 extents
[1:0xe048900000+4200000])
2018-10-16 21:07:15.113570 7f37d26f2700 10 bluefs _read_random h 0x55ddaca32680 0x22e31e1~f5a from
 file(ino 201593 size 0x41a812d mtime 2018-10-16 11:22:59.508274 bdev 1 allocated 4200000 extents
[1:0xe048900000+4200000])
2018-10-16 21:07:15.113733 7f37d26f2700 10 bluefs _read_random h 0x55ddaca32680 0x22e413b~ecc from
 file(ino 201593 size 0x41a812d mtime 2018-10-16 11:22:59.508274 bdev 1 allocated 4200000 extents
[1:0xe048900000+4200000])
2018-10-16 21:07:15.113897 7f37d26f2700 10 bluefs _read_random h 0x55ddaca32680 0x22e5007~f84 from
 file(ino 201593 size 0x41a812d mtime 2018-10-16 11:22:59.508274 bdev 1 allocated 4200000 extents
[1:0xe048900000+4200000])
2018-10-16 21:07:15.114033 7f37d26f2700 10 bluefs _read_random h 0x55dd99ea9f00 0xfc0~fd0 from fil
e(ino 201588 size 0x44a99d7 mtime 2018-10-16 11:22:53.020935 bdev 1 allocated 4500000 extents [1:0
xe010800000+4500000])
...
...
...
2018-10-16 21:09:06.190232 7f37f091f700  1 heartbeat_map is_healthy 'OSD::osd_op_tp thread 0x7f37d
26f2700' had timed out after 60
2018-10-16 21:09:06.190259 7f37f011e700  1 heartbeat_map is_healthy 'OSD::osd_op_tp thread 0x7f37d
26f2700' had timed out after 60
...
...
2018-10-16 21:09:29.761930 7f37d26f2700 10 bluefs _read_random h 0x55ddf45bcc80 0x2adbe12~fc9 from
 file(ino 201517 size 0x44c87af mtime 2018-10-16 06:57:28.419862 bdev 1 allocated 4500000 extents
[1:0xe66dd00000+4500000])
2018-10-16 21:09:29.762090 7f37d26f2700 10 bluefs _read_random h 0x55dd9422b400 0x2b812c2~fb6 from
 file(ino 69377 size 0x41b57a8 mtime 2018-07-10 15:29:54.444569 bdev 1 allocated 4200000 extents [
1:0xe057600000+4200000])
2018-10-16 21:09:29.762166 7f37f091f700  1 heartbeat_map is_healthy 'OSD::osd_op_tp thread 0x7f37d
26f2700' had timed out after 60
...
2018-10-16 21:09:30.812418 7f37d26f2700 10 bluefs _read_random h 0x55ddeb9a3e00 0x8d6201~fc9 from
file(ino 201734 size 0xd6c969 mtime 2018-10-16 21:02:55.145115 bdev 1 allocated e00000 extents [1:
0xe0b8d00000+e00000])
2018-10-16 21:09:30.812603 7f37d26f2700 10 osd.240 pg_epoch: 620465 pg[29.223( v 620465'88140843 (
620465'88139287,620465'88140843] local-lis/les=620388/620389 n=3800 ec=45900/45900 lis/c 620388/62
0388 les/c/f 620389/620389/0 620386/620388/620388) [240,183,176] r=0 lpr=620388 crt=620465'8814084
3 lcod 620465'88140842 mlcod 620465'88140842 active+clean] do_osd_op 29:c45aebf9:::rbd_header.0904
ed238e1f29:head [omap-get-vals-by-keys]
2018-10-16 21:09:30.812618 7f37d26f2700 10 osd.240 pg_epoch: 620465 pg[29.223( v 620465'88140843 (
620465'88139287,620465'88140843] local-lis/les=620388/620389 n=3800 ec=45900/45900 lis/c 620388/62
0388 les/c/f 620389/620389/0 620386/620388/620388) [240,183,176] r=0 lpr=620388 crt=620465'8814084
3 lcod 620465'88140842 mlcod 620465'88140842 active+clean] do_osd_op  omap-get-vals-by-keys
2018-10-16 21:09:30.812623 7f37d26f2700 15 bluestore(/var/lib/ceph/osd/ceph-240) omap_get_values 2
9.223_head oid #29:c45aebf9:::rbd_header.0904ed238e1f29:head#
2018-10-16 21:09:30.812669 7f37d26f2700 10 bluefs _read_random h 0x55dd99b3f980 0x4000b42~16ea88 f
rom file(ino 201731 size 0x42c0c5f mtime 2018-10-16 20:41:03.911976 bdev 1 allocated 4300000 exten
ts [1:0xe0b0100000+4300000])
2018-10-16 21:09:30.816931 7f37d26f2700 10 bluefs _read_random h 0x55ddf0180d80 0x40005d5~42bac6 f
rom file(ino 201722 size 0x452f1f4 mtime 2018-10-16 20:02:42.995010 bdev 1 allocated 4600000 exten
ts [1:0xdff2100000+4600000])
2018-10-16 21:09:30.830286 7f37d26f2700 10 bluefs _read_random h 0x55ddf0180d80 0x442c440~102d7f f
rom file(ino 201722 size 0x452f1f4 mtime 2018-10-16 20:02:42.995010 bdev 1 allocated 4600000 exten
ts [1:0xdff2100000+4600000])
2018-10-16 21:09:30.833395 7f37d26f2700 10 bluefs _read_random h 0x55ddaca32680 0x400084c~a4673 fr
om file(ino 201593 size 0x41a812d mtime 2018-10-16 11:22:59.508274 bdev 1 allocated 4200000 extent
s [1:0xe048900000+4200000])
2018-10-16 21:09:30.835353 7f37d26f2700 10 bluefs _read_random h 0x55ddaca32680 0x40a5260~102e98 f
rom file(ino 201593 size 0x41a812d mtime 2018-10-16 11:22:59.508274 bdev 1 allocated 4200000 exten
ts [1:0xe048900000+4200000])
2018-10-16 21:09:30.838791 7f37d26f2700 10 bluefs _read_random h 0x55ddaca32680 0x22de541~fc1 from
```

```
 file(ino 201593 size 0x41a812d mtime 2018-10-16 11:22:59.508274 bdev 1 allocated 4200000 extents
[1:0xe048900000+4200000])
2018-10-16 21:09:30.839002 7f37d26f2700 10 bluestore(/var/lib/ceph/osd/ceph-240) omap_get_values 2
9.223_head oid #29:c45aebf9:::rbd_header.0904ed238e1f29:head# = 0
...
...
2018-10-16 21:09:30.839387 7f37d26f2700 10 osd.240 pg_epoch: 620465 pg[29.223( v 620465'88140843 (
620465'88139287,620465'88140843] local-lis/les=620388/620389 n=3800 ec=45900/45900 lis/c 620388/62
0388 les/c/f 620389/620389/0 620386/620388/620388) [240,183,176] r=0 lpr=620388 crt=620465'8814084
3 lcod 620465'88140842 mlcod 620465'88140842 active+clean]  dropping ondisk_read_lock
2018-10-16 21:09:30.839413 7f37d26f2700 10 osd.240 620465 dequeue_op 0x55de0be44540 finish
2018-10-16 21:09:30.839446 7f37d26f2700  1 heartbeat_map reset_timeout 'OSD::osd_op_tp thread 0x7f
37d26f2700' had timed out after 60
```

The OSD in this case seems to be reading it's whole database (20GB) and scanned through it all for finding these OMAP keys.

It could be triggered on this specific RBD header, other objects worked just fine on the same OSD.

We tested by stopping osd.240 and this caused osd.183 to become the primary for PG 29.223. It did NOT suffer from the same problem, it responded in just a few ms with the OMAP keys.

Trying to use ceph-objectstore-tool and do the same was also very slow:

```
root@ceph37:~# time ceph-objectstore-tool --data-path /var/lib/ceph/osd/ceph-240 rbd_header.0904ed
238e1f29 list-omap
features
object_prefix
order
size
snap_seq
snapshot_0000000000000431

real    3m31.117s
user    0m19.016s
sys    0m8.524s
root@ceph37:~#
```

While doing so I could observe that the SSD was about 90% util and reading heavily.

Some information:

- Cluster deployed in September 2013 with Dumpling
- Upgraded to Luminous in March from Jewel
- OSDs migrated to BlueStore in March 2018
- Running mixture of 12.2.4, 12.2.5 and 12.2.8
- Ubuntu 16.04.5
- Samsumg PM863a 1.92TB SSDs

Offloading data from these OSDs to other OSDs made things better, but did not solve the problem.

I made a image with 'dd' from osd.240 in the state it was broken in. osd.240 is running again and is backfilling data to other OSDs (which is painfully slow), but that allows us to wipe the OSD at a later point and mkfs it again.

On request I have available:

- Logs of osd.240
- Image of osd.240 (1.92TB in size)

Related issues:

Related to bluestore - Bug #41213: BlueStore OSD taking more than 60 minutes ...	**Duplicate**	**08/12/2019**
Duplicated by bluestore - Bug #34526: OSD crash in KernelDevice::direct_read_...	**Duplicate**	**08/30/2018**

Copied to bluestore - Backport #40632: nautilus: High amount of Read I/O on B...	**Resolved**
Copied to bluestore - Backport #42428: mimic: High amount of Read I/O on Blue...	**Rejected**

History

#1 - 10/17/2018 12:22 PM - Wido den Hollander

To add to this, I am also to reproduce it on osd.246 in this cluster:

```
2018-10-17 14:06:55.926086 7f5470bd6700  1 heartbeat_map is_healthy 'OSD::osd_op_tp thread 0x7f54531aa700' had
 timed out after 60
2018-10-17 14:06:55.926192 7f54531aa700 10 bluefs _read_random h 0x55b695996b00 0x27ff7ca~fc9 from file(ino 21
1410 size 0x41dcdb0 mtime 2018-10-15 18:41:44.958802 bdev 1 allocated 4200000 extents [1:0xd726300000+4200000]
)
2018-10-17 14:06:55.926358 7f54531aa700 10 bluefs _read_random h 0x55b695996b00 0x2800793~f2e from file(ino 21
1410 size 0x41dcdb0 mtime 2018-10-15 18:41:44.958802 bdev 1 allocated 4200000 extents [1:0xd726300000+4200000]
)
2018-10-17 14:06:55.926509 7f54531aa700 10 bluefs _read_random h 0x55b695996b00 0x28016c1~f31 from file(ino 21
1410 size 0x41dcdb0 mtime 2018-10-15 18:41:44.958802 bdev 1 allocated 4200000 extents [1:0xd726300000+4200000]
)
2018-10-17 14:06:55.926658 7f54531aa700 10 bluefs _read_random h 0x55b695996b00 0x28025f2~f5a from file(ino 21
1410 size 0x41dcdb0 mtime 2018-10-15 18:41:44.958802 bdev 1 allocated 4200000 extents [1:0xd726300000+4200000]
)
2018-10-17 14:06:55.926817 7f54531aa700 10 bluefs _read_random h 0x55b725086600 0x2ac39e3~fdc from file(ino 21
5611 size 0x44777d4 mtime 2018-10-17 12:05:47.273988 bdev 1 allocated 4500000 extents [1:0xdcae300000+4500000]
)
2018-10-17 14:06:55.926979 7f54531aa700 10 bluefs _read_random h 0x55b695996b00 0x280354c~fb0 from file(ino 21
1410 size 0x41dcdb0 mtime 2018-10-15 18:41:44.958802 bdev 1 allocated 4200000 extents [1:0xd726300000+4200000]
)
2018-10-17 14:06:55.927139 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x30f9a28~f5b from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.927282 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x30fa983~f84 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.927442 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x30fb907~f45 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.927598 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x30fc84c~f45 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.927759 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x30fd791~f58 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.927918 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x30fe6e9~f17 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.928078 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x30ff600~f75 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.928236 7f54531aa700 10 bluefs _read_random h 0x55b725086600 0x2ac49bf~fcf from file(ino 21
5611 size 0x44777d4 mtime 2018-10-17 12:05:47.273988 bdev 1 allocated 4500000 extents [1:0xdcae300000+4500000]
)
2018-10-17 14:06:55.928392 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x3100575~f18 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.928557 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x310148d~f74 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.928715 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x3102401~f48 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.928876 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x3103349~f76 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.929034 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x31042bf~fae from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.929193 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x310526d~f31 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.929349 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x310619e~f45 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.929510 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x31070e3~f9b from file(ino 21
```

```
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.929668 7f54531aa700 10 bluefs _read_random h 0x55b725086600 0x2ac598e~fcf from file(ino 21
5611 size 0x44777d4 mtime 2018-10-17 12:05:47.273988 bdev 1 allocated 4500000 extents [1:0xdcae300000+4500000]
)
2018-10-17 14:06:55.929828 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x310807e~fc5 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.929990 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x3109043~f58 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.930140 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x3109f9b~f70 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.930297 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x310af0b~f6f from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.930456 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x310be7a~f73 from file(ino 21
4920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000]
)
2018-10-17 14:06:55.930612 7f54531aa700 10 bluefs _read_random h 0x55b695996880 0x310cded~1000 from file(ino 2
14920 size 0x443816f mtime 2018-10-16 01:55:02.893256 bdev 1 allocated 4500000 extents [1:0xe3f9000000+4500000
])
2018-10-17 14:06:55.930771 7f54531aa700 10 bluefs _read_random h 0x55b725086600 0x2ac695d~f73 from file(ino 21
5611 size 0x44777d4 mtime 2018-10-17 12:05:47.273988 bdev 1 allocated 4500000 extents [1:0xdcae300000+4500000]
)
2018-10-17 14:06:55.943271 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op 29:d19cf8fc:::rbd_header.778d1b2ae8944a:head [omap-get-vals-by-keys]
2018-10-17 14:06:55.943292 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op  omap-get-vals-by-keys
2018-10-17 14:06:55.943379 7f54531aa700 10 bluefs _read_random h 0x55b6c5899380 0x37d3ee7~fcf from file(ino 21
5356 size 0x44bff2d mtime 2018-10-16 12:29:43.142071 bdev 1 allocated 4500000 extents [1:0xd732c00000+4500000]
)
2018-10-17 14:06:55.943594 7f54531aa700 10 bluestore(/var/lib/ceph/osd/ceph-246) omap_get_values 29.18b_head o
id #29:d19cf8fc:::rbd_header.778d1b2ae8944a:head# = 0
2018-10-17 14:06:55.943605 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] method called response length=20
2018-10-17 14:06:55.943626 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op  call rbd.get_parent
2018-10-17 14:06:55.943649 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] call method rbd.get_parent
2018-10-17 14:06:55.943656 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op 29:d19cf8fc:::rbd_header.778d1b2ae8944a:head [stat]
2018-10-17 14:06:55.943672 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op  stat
2018-10-17 14:06:55.943682 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] stat oi has 0 2018-10-17 14:01:35.033620
2018-10-17 14:06:55.943696 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op 29:d19cf8fc:::rbd_header.778d1b2ae8944a:head [omap-get-vals-by-keys]
2018-10-17 14:06:55.943714 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op  omap-get-vals-by-keys
2018-10-17 14:06:55.943758 7f54531aa700 10 bluestore(/var/lib/ceph/osd/ceph-246) omap_get_values 29.18b_head o
id #29:d19cf8fc:::rbd_header.778d1b2ae8944a:head# = 0
2018-10-17 14:06:55.943775 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
```

```
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op 29:d19cf8fc:::rbd_header.778d1b2ae8944a:head [omap-get-vals-by-keys]
2018-10-17 14:06:55.943784 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op  omap-get-vals-by-keys
2018-10-17 14:06:55.943824 7f54531aa700 10 bluestore(/var/lib/ceph/osd/ceph-246) omap_get_values 29.18b_head o
id #29:d19cf8fc:::rbd_header.778d1b2ae8944a:head# = 0
2018-10-17 14:06:55.943829 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] method called response length=28
2018-10-17 14:06:55.943848 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op  call lock.get_info
2018-10-17 14:06:55.943868 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] call method lock.get_info
2018-10-17 14:06:55.943879 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op 29:d19cf8fc:::rbd_header.778d1b2ae8944a:head [getxattr lock.rbd_lock]
2018-10-17 14:06:55.943887 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] do_osd_op  getxattr lock.rbd_lock
2018-10-17 14:06:55.943907 7f54531aa700 10 bluestore(/var/lib/ceph/osd/ceph-246) getattr 29.18b_head #29:d19cf
8fc:::rbd_header.778d1b2ae8944a:head# _lock.rbd_lock = -61
2018-10-17 14:06:55.943923 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean] method called response length=15
2018-10-17 14:06:55.943932 7f54531aa700 10 osd.246 pg_epoch: 623992 pg[29.18b( v 623992'105484940 (623992'1054
83363,623992'105484940] local-lis/les=620277/620278 n=3671 ec=45900/45900 lis/c 620277/620277 les/c/f 620278/6
20295/0 620277/620277/620277) [246,110,238] r=0 lpr=620277 crt=623992'105484940 lcod 623992'105484939 mlcod 62
3992'105484939 active+clean]  dropping ondisk_read_lock
2018-10-17 14:06:55.943996 7f54531aa700 10 osd.246 623992 dequeue_op 0x55b6d7531880 finish
2018-10-17 14:06:55.944023 7f54531aa700  1 heartbeat_map reset_timeout 'OSD::osd_op_tp thread 0x7f54531aa700'
had timed out after 60
```

I reversed looked-up **rbd_header.778d1b2ae8944a** and I tried to get the information from the RBD image:

```
root@mon01:~# time rbd -p rbd-ssd-c03-p02 info 476ebf43-e121-41da-a6f9-a966a5310170
rbd image '476ebf43-e121-41da-a6f9-a966a5310170':
    size 20GiB in 5120 objects
    order 22 (4MiB objects)
    block_name_prefix: rbd_data.778d1b2ae8944a
    format: 2
    features: layering, striping
    flags:
    stripe unit: 4MiB
    stripe count: 1

real    5m14.127s
user    0m0.128s
sys    0m0.044s
root@mon01:~#
```

As you can see 5 minutes.

In this time the osd is super busy with reading from it's SSD which appears to be RocksDB reads when looking at bluefs.

Again, this OSD is unusable and causes slow requests.

The **nodown** flag is set at the moment, because otherwise the OSDs keep marking each other as down.

#2 - 10/17/2018 12:29 PM - Igor Fedotov

- Status changed from New to 12

#3 - 10/17/2018 12:31 PM - Igor Fedotov

Just one thing to add - reads from BlueFS are performed in a sequential manner using pretty ineffective block sizes (mostly 0xf??).

Manual DB compaction didn't help.

#4 - 10/17/2018 12:33 PM - Igor Fedotov

and it's BlueStore::get_omap_iterator() and/or its subsequent usage which triggered these long massive reads.

#5 - 10/17/2018 01:43 PM - Wido den Hollander

One thing to add is that a few ago, at 15-10-2018 at 18:13 multiple OSDs in this cluster were showing these messages:

```
Oct 15 18:13:00 ceph24 ceph-osd[1491508]: 2018-10-15 18:13:00.752036 7f469068b700 -1 bluestore(/var/lib/ceph/o
sd/ceph-150) _balance_bluefs_freespace allocate failed on 0x80000000 min_alloc_size 0x4000
Oct 15 18:13:01 ceph24 CRON[1517085]: (root) CMD (/usr/local/sbin/ceph-osd-stats >/dev/null 2>&1)
Oct 15 18:13:09 ceph24 ceph-osd[1491508]: /build/ceph-12.2.4/src/os/bluestore/BlueStore.cc: In function 'int B
lueStore::_balance_bluefs_freespace(PExtentVector*)' thread 7f469068b700 time 2018-10-15 18:13:09.778
655
Oct 15 18:13:09 ceph24 ceph-osd[1491508]: /build/ceph-12.2.4/src/os/bluestore/BlueStore.cc: 4950: FAILED asser
t(0 == "allocate failed, wtf")
Oct 15 18:13:09 ceph24 ceph-osd[1491508]:  ceph version 12.2.4 (52085d5249a80c5f5121a76d6288429f35e4e77b) lumi
nous (stable)
Oct 15 18:13:09 ceph24 ceph-osd[1491508]:  1: (ceph::__ceph_assert_fail(char const*, char const*, int, char co
nst*)+0x102) [0x556648b55872]
Oct 15 18:13:09 ceph24 ceph-osd[1491508]:  2: (BlueStore::_balance_bluefs_freespace(std::vector<bluestore_pext
ent_t, mempool::pool_allocator<(mempool::pool_index_t)4, bluestore_pextent_t> >*)+0x1b21) [0x5566489e61d1]
Oct 15 18:13:09 ceph24 ceph-osd[1491508]:  3: (BlueStore::_kv_sync_thread()+0x1ac0) [0x5566489e8c50]
Oct 15 18:13:09 ceph24 ceph-osd[1491508]:  4: (BlueStore::KVSyncThread::entry()+0xd) [0x556648a2d08d]
Oct 15 18:13:09 ceph24 ceph-osd[1491508]:  5: (()+0x76ba) [0x7f46a140b6ba]
Oct 15 18:13:09 ceph24 ceph-osd[1491508]:  6: (clone()+0x6d) [0x7f46a048241d]
Oct 15 18:13:09 ceph24 ceph-osd[1491508]:  NOTE: a copy of the executable, or `objdump -rdS <executable>` is n
eeded to interpret this.
Oct 15 18:13:09 ceph24 ceph-osd[1491508]: 2018-10-15 18:13:09.784205 7f469068b700 -1 /build/ceph-12.2.4/src/os
/bluestore/BlueStore.cc: In function 'int BlueStore::_balance_bluefs_freespace(PExtentVector*)' thread 7f46906
8b700 time 2018-10-15 18:13:09.778655
Oct 15 18:13:09 ceph24 ceph-osd[1491508]: /build/ceph-12.2.4/src/os/bluestore/BlueStore.cc: 4950: FAILED asser
t(0 == "allocate failed, wtf")
```

These OSDs were still running 12.2.4 indeed. The OSDs are no longer suffering from these issues, but it might be related that all these OSDs had these messages in their logs?

On other OSDs I also saw these messages show up earlier:

```
2018-10-10 18:40:40.022988 7fa0ea0b7700 -1 bluestore(/var/lib/ceph/osd/ceph-242) _balance_bluefs_freespace all
ocate failed on 0x80000000 min_alloc_size 0x4000
2018-10-10 18:40:40.023008 7fa0ea0b7700  0 stupidalloc 0x0x55dbd1ae3ab0 dump free bin 0: 0 extents
2018-10-10 18:40:40.023011 7fa0ea0b7700  0 stupidalloc 0x0x55dbd1ae3ab0 dump free bin 1: 0 extents
2018-10-10 18:40:40.023012 7fa0ea0b7700  0 stupidalloc 0x0x55dbd1ae3ab0 dump free bin 2: 0 extents
2018-10-10 18:40:40.023013 7fa0ea0b7700  0 stupidalloc 0x0x55dbd1ae3ab0 dump free bin 3: 3527582 extents
2018-10-10 18:40:40.023014 7fa0ea0b7700  0 stupidalloc 0x0x55dbd1ae3ab0 dump  0x30000~4000
2018-10-10 18:40:40.023015 7fa0ea0b7700  0 stupidalloc 0x0x55dbd1ae3ab0 dump  0x38000~4000
2018-10-10 18:40:40.023015 7fa0ea0b7700  0 stupidalloc 0x0x55dbd1ae3ab0 dump  0x48000~4000
```

See: https://tracker.ceph.com/issues/23063

Not sure if it is related.

#6 - 10/17/2018 07:21 PM - Wido den Hollander

There might be a work-around/fix for this: compacting the database

I did this:

```
ceph daemon osd.240 compact
```

That initially didn't work. It compacted the database, but it stayed slow. Then I tried a restart of the OSD and that worked!

Boot times of this OSD were about 5 minutes, but now it booted in <30 seconds, like it should.

The compact should have happened internally and automatically, but it resolved the situation for now.

All the 96 OSDs in this cluster have been compacted and restarted and ever since we haven't seen slow requests. The last slow request was 3 hours ago. They were happening every 5 minutes.

#7 - 10/17/2018 07:47 PM - Igor Fedotov

- Project changed from Ceph to bluestore

- Category deleted (OSD)

#8 - 10/17/2018 07:47 PM - Igor Fedotov

- Subject changed from bluestore: High amount of Read I/O on BlueFS/DB when listing omap keys to High amount of Read I/O on BlueFS/DB when listing omap keys

#9 - 02/28/2019 02:19 PM - Igor Fedotov

I think this is the same issue:
https://marc.info/?l=ceph-devel&m=155134206210976&w=2

#10 - 02/28/2019 02:41 PM - Igor Fedotov

We've got another occurrence for this issue too.
Omap listing for specific onode consistently takes ~2 mins while doing intensive reads (primarily from HDD where data has been spilled over to).

```
time ceph-objectstore-tool --data-path /var/lib/ceph/osd/ceph-20 --pgid 7.24 10000609512.00000000 list-omap
Error getting attr on : 7.24_head,#-9:24000000:::scrub_7.24:head#, (61) No data
available
lane.html_head
laneBarcode.html_head

real    2m14.627s
user    0m14.714s
sys     0m10.941s
```

At another host the same listing takes ~10 secons:

```
time ceph-objectstore-tool --data-path /var/lib/ceph/osd/ceph-1 --pgid 7.24 10000609512.00000000 list-omap
```

```
Error getting attr on : 7.24_head,#-9:24000000:::scrub_7.24:head#, (61) No data
available
lane.html_head
laneBarcode.html_head

real    0m9.600s
user    0m3.387s
sys     0m0.339s
```

Here is the full log for OSD doing the scrub for this specific PG with increased bluestore and bluefs debug levels:
ceph-post-file: 095fd9da-a9db-4846-a83c-bbe4e9cca25d

Looks like there are massive sequential BlueFS reads using <4K data chunks from both SSD and HDD drives, starting at:

```
2019-02-27 12:19:57.479738 7ff38ba07700 15 bluestore(/var/lib/ceph/osd/ceph-20) omap_get_header 7.24_head oid
#7:2463dac2:::10000609512.
00000000:head#
```

e.g. log snippet for reading from just one ino:

```
2019-02-27 12:20:39.818490 7ff38ba07700 10 bluefs _read_random h 0x55717ac0f700 0x409330c~17b10b from file(ino
 51310 size 0x420e44c mtime 2019-02-12 12:31:32.904562 bdev 2 allocated 4300000 extents [2:0x3a3da600000+43000
00])
2019-02-27 12:20:39.818498 7ff38ba07700 20 bluefs _read_random read buffered 0x409330c~17b10b of 2:0x3a3da6000
00+4300000
2019-02-27 12:20:39.833662 7ff38ba07700 20 bluefs _read_random got 1552651
2019-02-27 12:20:39.833933 7ff38ba07700 10 bluefs _read_random h 0x55717ac0f700 0x0~f99 from file(ino 51310 si
ze 0x420e44c mtime 2019-02-12 12:31:32.904562 bdev 2 allocated 4300000 extents [2:0x3a3da600000+4300000])
2019-02-27 12:20:39.833938 7ff38ba07700 20 bluefs _read_random read buffered 0x0~f99 of 2:0x3a3da600000+430000
0
2019-02-27 12:20:39.834039 7ff38ba07700 20 bluefs _read_random got 3993
2019-02-27 12:20:39.834054 7ff38ba07700 10 bluefs _read_random h 0x55717ac0f700 0xf99~e71 from file(ino 51310
 size 0x420e44c mtime 2019-02-12 12:31:32.904562 bdev 2 allocated 4300000 extents [2:0x3a3da600000+4300000])
2019-02-27 12:20:39.834058 7ff38ba07700 20 bluefs _read_random read buffered 0xf99~e71 of 2:0x3a3da600000+4300
000
2019-02-27 12:20:39.834195 7ff38ba07700 20 bluefs _read_random got 3697
...
2019-02-27 12:20:42.783982 7ff38ba07700 10 bluefs _read_random h 0x55717ac0f700 0x3ffc7fb~ea7 from file(ino 51
310 size 0x420e44c mtime 2019-02-12 12:31:32.904562 bdev 2 allocated 4300000 extents [2:0x3a3da600000+43
00000])
2019-02-27 12:20:42.783988 7ff38ba07700 20 bluefs _read_random read buffered 0x3ffc7fb~ea7 of 2:0x3a3da600000+
4300000
2019-02-27 12:20:42.784061 7ff38ba07700 20 bluefs _read_random got 3751
2019-02-27 12:20:42.784073 7ff38ba07700 10 bluefs _read_random h 0x55717ac0f700 0x3ffd6a2~eb1 from file(ino 51
310 size 0x420e44c mtime 2019-02-12 12:31:32.904562 bdev 2 allocated 4300000 extents [2:0x3a3da600000+43
00000])
2019-02-27 12:20:42.784078 7ff38ba07700 20 bluefs _read_random read buffered 0x3ffd6a2~eb1 of 2:0x3a3da600000+
4300000
2019-02-27 12:20:42.784150 7ff38ba07700 20 bluefs _read_random got 3761
2019-02-27 12:20:42.784161 7ff38ba07700 10 bluefs _read_random h 0x55717ac0f700 0x3ffe553~e9a from file(ino 51
310 size 0x420e44c mtime 2019-02-12 12:31:32.904562 bdev 2 allocated 4300000 extents [2:0x3a3da600000+43
00000])
2019-02-27 12:20:42.784166 7ff38ba07700 20 bluefs _read_random read buffered 0x3ffe553~e9a of 2:0x3a3da600000+
4300000
2019-02-27 12:20:42.784238 7ff38ba07700 20 bluefs _read_random got 3738
2019-02-27 12:20:42.784267 7ff38ba07700 10 bluefs _read_random h 0x55717ac0f700 0x3fff3ed~ecb from file(ino 51
310 size 0x420e44c mtime 2019-02-12 12:31:32.904562 bdev 2 allocated 4300000 extents [2:0x3a3da600000+43
00000])
2019-02-27 12:20:42.784274 7ff38ba07700 20 bluefs _read_random read buffered 0x3fff3ed~ecb of 2:0x3a3da600000+
4300000
2019-02-27 12:20:42.784348 7ff38ba07700 20 bluefs _read_random got 3787
2019-02-27 12:20:42.784364 7ff38ba07700 10 bluefs _read_random h 0x55717a654b00 0x1d374c5~f99 from file(ino 56
018 size 0x4517dd9 mtime 2019-02-27 12:18:24.566347 bdev 1 allocated 4600000 extents [1:0xad500000+46000
00])
2019-02-27 12:20:42.784371 7ff38ba07700 20 bluefs _read_random read buffered 0x1d374c5~f99 of 1:0xad500000+460
0000
2019-02-27 12:20:42.784455 7ff38ba07700 20 bluefs _read_random got 3993
2019-02-27 12:20:42.784473 7ff38ba07700 10 bluefs _read_random h 0x55717ac0f700 0x40002b8~1c1 from file(ino 51
```

```
310 size 0x420e44c mtime 2019-02-12 12:31:32.904562 bdev 2 allocated 4300000 extents [2:0x3a3da600000+43
00000])
2019-02-27 12:20:42.784480 7ff38ba07700 20 bluefs _read_random read buffered 0x40002b8~1c1 of 2:0x3a3da600000+
4300000
2019-02-27 12:20:42.784541 7ff38ba07700 20 bluefs _read_random got 449
```

#11 - 02/28/2019 02:43 PM - Igor Fedotov

- Priority changed from Normal to High

#12 - 02/28/2019 03:14 PM - Sage Weil

- it looks like implementing readahead in bluefs would help
- we think newer rocksdb does its own readahead

#13 - 02/28/2019 03:31 PM - Sage Weil

- Duplicated by Bug #34526: OSD crash in KernelDevice::direct_read_unaligned while scrubbing added

#14 - 03/01/2019 04:44 PM - Wido den Hollander

FYI, I think I hit another case with this with this in the last two weeks.

A RGW only case where if you would list certain RGW buckets some OSDs would eat up 100% of the disk I/O.

Compacting the RocksDB database would help for a while, but in the end it would come back. We worked around it by moving the RGW indexes to NVMe, but that only masks the problem.

#15 - 04/25/2019 02:44 PM - Igor Fedotov

- Status changed from 12 to In Progress

#16 - 05/03/2019 12:49 PM - Igor Fedotov

Some notes on the similar issue I'm investigating.
I've got Luminous OSD image that contains at least one object with pretty slow omap enumeration.
Following command:

ceph-objectstore-tool --data-path /mnt/ceph --pgid 7.24 10000609512.00000000 list-omap

takes 99 seconds.
Using Octopus code base reduces this time to 55 seconds.
And enabling RocksDB/BlueFS read-ahead (https://github.com/ceph/ceph/pull/24861 or https://github.com/ceph/ceph/pull/27782) reduces it to 30-33 seconds.

Also https://github.com/ceph/ceph/pull/27627 mentions pretty relevant issue that might be the root cause - need to read through tons of deleted omaps to reach the list end.
I've made an experiment (against Octopus code base) to learn how omap records are retrieved in my case and where most of the time is spent.
Here is the log snippet with some additional logging inserted into BlueStore::OmapIteratorImpl::next() and BlueStore::OmapIteratorImpl::key():

```
2019-05-03 12:27:07.393 7ff895ced740 10 bluestore(/mnt/ceph) get_omap_iterator has_omap = 1
2019-05-03 12:27:07.421 7ff895ced740  2 bluestore.OmapIteratorImpl(0x55cced692af0) valid is at 0x0000000000bc6089'.lane.html_head'
2019-05-03 12:27:07.421 7ff895ced740  2 bluestore.OmapIteratorImpl(0x55cced692af0) key = lane.html_head
2019-05-03 12:27:07.421 7ff895ced740  2 bluestore.OmapIteratorImpl(0x55cced692af0) next
2019-05-03 12:27:07.421 7ff895ced740  2 bluestore.OmapIteratorImpl(0x55cced692af0) valid is at 0x0000000000bc6089'.laneBarcode.html_head'
2019-05-03 12:27:07.421 7ff895ced740  2 bluestore.OmapIteratorImpl(0x55cced692af0) key = laneBarcode.html_head
2019-05-03 12:27:07.421 7ff895ced740  2 bluestore.OmapIteratorImpl(0x55cced692af0) next
....
2019-05-03 12:27:53.478 7ff895ced740  0 bluestore(/mnt/ceph) log_latency_fn slow operation observed for next, latency = 46.0528s, lat = 46s
omap_iterator(cid = 7.24_head, oid = #7:2463dac2:::10000609512.00000000:head#)
```

2019-05-03 12:27:53.478 7ff895ced740 2 bluestore.OmapIteratorImpl(0x55cced692af0) valid is at 0x0000000000c3ef1d'-'

So one can see that actual records are retrieved pretty fast and most of time is spent in the last next() call looking (probably) for the next valid record. Which confirms potential benefit from OMAP tail as per https://github.com/ceph/ceph/pull/27627

I'm not sure that OMAP tail presence resolves slow omap enumeration totally (e.g. we can probably face cases when there is a large span between valid keys of the same object) but it definitely might make things better. Surely along with DB read-ahead support.

#17 - 05/16/2019 09:56 AM - Igor Fedotov

Just to mention - find two more onodes at customer's cluster with slow omap listing - both spend all the time while seeking for the tail (as suggested in https://github.com/ceph/ceph/pull/27627)

#18 - 07/02/2019 11:16 AM - Igor Fedotov

- Status changed from In Progress to Pending Backport

- Backport set to nautilus

- Pull request ID set to 27627

#19 - 07/02/2019 08:33 PM - Nathan Cutler

- Copied to Backport #40632: nautilus: High amount of Read I/O on BlueFS/DB when listing omap keys added

#20 - 07/31/2019 01:53 PM - Igor Fedotov

- Status changed from Pending Backport to Resolved

#21 - 08/01/2019 05:33 PM - Robin Johnson

Nathan/Igor: any chance of a Luminious backport for v12.2.13, and Mimic as well?

#22 - 08/05/2019 12:11 PM - Igor Fedotov

Robin,
not sure about the next Mimic/Luminous releases, but may be later. Let this patch bake for a bit in Nautilus and master...

#23 - 08/13/2019 05:30 PM - Vikhyat Umrao

- Related to Bug #41213: BlueStore OSD taking more than 60 minutes to boot added

#24 - 10/21/2019 12:18 PM - Gerben Meijer

How long would this need to "bake"? Running into this frequently (several times per day). Is this going to make it into 12.2.13?

#25 - 10/21/2019 06:26 PM - Nathan Cutler

Gerben Meijer wrote:

> How long would this need to "bake"? Running into this frequently (several times per day). Is this going to make it into 12.2.13?

Luminous is nearly End Of Life (EOL). Can you upgrade to mimic or nautilus?

#26 - 10/21/2019 06:27 PM - Nathan Cutler

- Status changed from Resolved to Pending Backport

- Backport changed from nautilus to nautilus, mimic

#27 - 10/21/2019 07:07 PM - Gerben Meijer

There's several clusters on luminous that can't be upgraded just yet, but will upgrade what we can. I'm just trying to determine if it'll be backported or not. Based on the tags I presume not?

#28 - 10/22/2019 10:08 AM - Nathan Cutler

Upon discussion with @Igor, the backports of this issue will require

- https://github.com/ceph/ceph/pull/27627
- https://github.com/ceph/ceph/pull/27782
- RocksDB upgrade

and will be non-trivial.

#29 - 10/22/2019 10:36 AM - Nathan Cutler

- Copied to Backport #42428: mimic: High amount of Read I/O on BlueFS/DB when listing omap keys added

#30 - 10/22/2019 03:00 PM - Sage Weil

- Status changed from Pending Backport to Resolved

- Backport changed from nautilus, mimic to nautilus

After discussing in the rados team standup, we've decided not to backport to mimic/luminous at this time. Nautilus fixes the issue and has been available for a while now.