bluestore - Bug #25001

Crashing OSDs after going from 12.2.5 -> 12.2.6 -> 13.2.0

07/19/2018 03:25 PM - Troy Ablan

Status:	Can't reproduce	**% Done:**	0%
Priority:	High		
Assignee:			
Category:			
Target version:			
Source:	Community (user)	**Affected Versions:**	v13.2.0
Tags:		**ceph-qa-suite:**	
Backport:		**Pull request ID:**	
Regression:	No	**Crash signature (v1):**	
Severity:	2 - major	**Crash signature (v2):**	
Reviewed:			

Description

This bug has been opened following on from http://lists.ceph.com/pipermail/ceph-users-ceph.com/2018-July/028232.html

At some point in June, I updated my entire system to 12.2.5. Over the next few weeks I started noticing PGs would randomly go inconsistent. I would kick off the repair and it would come back clean.
Last weekend, I did a yum update, saw that there were new packages, so I updated the rest of the cluster.
I went to bed, woke up, and noticed things were in a really bad state. VMs had gotten IO errors and most mounted the disk in a read-only state. At this point, I looked for 12.2.6 release notes, found none on the website, and decided to switch to the Mimic repo as I could think of no other option at this point. As this didn't fix the problem I started looking around on the mailing list and only then did I understand that I should not have panicked and should have waited for 12.2.7.

I was mistaken when I mentioned on the ML that these were SSDs crashing. They're SATA drives. Not all of them are crashing, but the ones that do crash do so repeatedly.

At this point in time, most VMs cannot start due to read errors, and the ones that can start have long pauses because of the OSD churning.

I got a core and a log of an entire single invocation of ceph-osd with debug bluestore = 20 set.

Since the full log and core files are too large to attach, I've hosted them at https://mooinglemur.com/2018-07-ceph/

Just the trace is below:

```
ceph version 13.2.0 (79a10589f1f80dfe21e8f9794365ed98143071c4) mimic (stable)
 1: (()+0x8e1870) [0x559e3b6e9870]
 2: (()+0xf6d0) [0x7f55434b76d0]
 3: (gsignal()+0x37) [0x7f55424d8277]
 4: (abort()+0x148) [0x7f55424d9968]
 5: (BlueStore::_wctx_finish(BlueStore::TransContext*, boost::intrusive_ptr&lt;BlueStore::Collection&gt;&, boost::intrusive_ptr&lt;BlueStore::Onode&gt;, BlueStore::WriteContext*, std::set&lt;BlueStore::SharedBlob*, std
::less&lt;BlueStore::SharedBlob*&gt;, std::allocator&lt;BlueStore::SharedBlob*&gt; >
)+0xdea) [0x559e3b5d4cda]
 6: (BlueStore::_do_truncate(BlueStore::TransContext
, boost::intrusive_ptr&lt;BlueStore::Collection&gt;&, boost::intrusive_ptr&lt;BlueStore::Onode&gt;
, unsigned long, std::set&lt;BlueStore::SharedBlob*, std::less&lt;Blue
Store::SharedBlob*>, std::allocator&lt;BlueStore::SharedBlob*&gt; >)+0x13d) [0x559e3b5e3ead]
 7: (BlueStore::_do_remove(BlueStore::TransContext
, boost::intrusive_ptr&lt;BlueStore::Collection&gt;&, boost::intrusive_ptr&lt;BlueStore::Onode&gt;
)+0xbf) [0x559e3b5e467f]
 8: (BlueStore::_remove(BlueStore::TransContext*, boost::intrusive_ptr&lt;BlueStore::Collection&gt
;&, boost::intrusive_ptr&lt;BlueStore::Onode&gt;&)+0x60) [0x559e3b5e5e50]
 9: (BlueStore::_txc_add_transaction(BlueStore::TransContext*, ObjectStore::Transaction*)+0x105d)
```

```
[0x559e3b5f066d]
 10: (BlueStore::queue_transactions(boost::intrusive_ptr&lt;ObjectStore::CollectionImpl&gt;&, std:
:vector&lt;ObjectStore::Transaction, std::allocator&lt;ObjectStore::Transaction&gt; >&, boost::int
rusive_ptr&lt;TrackedOp&gt;,
 ThreadPool::TPHandle*)+0x519) [0x559e3b5f27b9]
 11: (ObjectStore::queue_transaction(boost::intrusive_ptr&lt;ObjectStore::CollectionImpl&gt;&, Obj
ectStore::Transaction&&, boost::intrusive_ptr&lt;TrackedOp&gt;, ThreadPool::TPHandle*)+0x80) [0x55
9e3b2038d0]
 12: (OSD::dispatch_context_transaction(PG::RecoveryCtx&, PG*, ThreadPool::TPHandle*)+0x58) [0x559
e3b19a788]
 13: (OSD::dequeue_peering_evt(OSDShard*, PG*, std::shared_ptr&lt;PGPeeringEvent&gt;, ThreadPool::
TPHandle&)+0xfe) [0x559e3b1c823e]
 14: (PGPeeringItem::run(OSD*, OSDShard*, boost::intrusive_ptr&lt;PG&gt;&, ThreadPool::TPHandle&)+
0x50) [0x559e3b41f820]
 15: (OSD::ShardedOpWQ::_process(unsigned int, ceph::heartbeat_handle_d*)+0x592) [0x559e3b1d2e02]
 16: (ShardedThreadPool::shardedthreadpool_worker(unsigned int)+0x3d3) [0x7f554695d333]
 17: (ShardedThreadPool::WorkThreadSharded::entry()+0x10) [0x7f554695df20]
 18: (()+0x7e25) [0x7f55434afe25]
 19: (clone()+0x6d) [0x7f55425a0bad]
 NOTE: a copy of the executable, or `objdump -rdS &lt;executable&gt;` is needed to interpret this.
```

Thanks!

History

#1 - 07/19/2018 09:24 PM - Brad Hubbard

- Project changed from Ceph to bluestore

- Source set to Community (user)

- Affected Versions v13.2.0 added

#2 - 07/19/2018 10:48 PM - Brad Hubbard

Looks like we are passing a bad bluestore_pextent_t into txc->released.insert.

```
(gdb) bt
#0  0x00007f55434b759b in raise (sig=sig@entry=6) at ../nptl/sysdeps/unix/sysv/linux/pt-raise.c:37
#1  0x0000559e3b6e9901 in reraise_fatal (signum=6) at /usr/src/debug/ceph-13.2.0/src/global/signal_handler.cc:
74
#2  handle_fatal_signal (signum=6) at /usr/src/debug/ceph-13.2.0/src/global/signal_handler.cc:138
#3  <signal handler called>
#4  0x00007f55424d8277 in __GI_raise (sig=sig@entry=6) at ../nptl/sysdeps/unix/sysv/linux/raise.c:56
#5  0x00007f55424d9968 in __GI_abort () at abort.c:90
#6  0x0000559e3b5d4cda in insert (plen=0x0, pstart=0x0, len=<optimized out>, start=<optimized out>, this=0x559
e524a8d80) at /usr/src/debug/ceph-13.2.0/src/include/interval_set.h:460
#7  BlueStore::_wctx_finish (this=this@entry=0x559e3e554000, txc=txc@entry=0x559e524a8c00, c=..., o=..., wctx=
wctx@entry=0x7f5523b366a0, maybe_unshared_blobs=maybe_unshared_blobs@entry=0x0)
    at /usr/src/debug/ceph-13.2.0/src/os/bluestore/BlueStore.cc:10757
#8  0x0000559e3b5e3ead in BlueStore::_do_truncate (this=this@entry=0x559e3e554000, txc=0x559e524a8c00, c=...,
o=..., offset=offset@entry=0, maybe_unshared_blobs=maybe_unshared_blobs@entry=0x0)
    at /usr/src/debug/ceph-13.2.0/src/os/bluestore/BlueStore.cc:11139
#9  0x0000559e3b5e467f in BlueStore::_do_remove (this=this@entry=0x559e3e554000, txc=<optimized out>, txc@entr
y=0x559e524a8c00, c=..., o=...) at /usr/src/debug/ceph-13.2.0/src/os/bluestore/BlueStore.cc:11185
#10 0x0000559e3b5e5e50 in BlueStore::_remove (this=this@entry=0x559e3e554000, txc=txc@entry=0x559e524a8c00, c=
..., o=...) at /usr/src/debug/ceph-13.2.0/src/os/bluestore/BlueStore.cc:11286
#11 0x0000559e3b5f066d in BlueStore::_txc_add_transaction (this=this@entry=0x559e3e554000, txc=txc@entry=0x559
e524a8c00, t=t@entry=0x559e530a9760) at /usr/src/debug/ceph-13.2.0/src/os/bluestore/BlueStore.cc:9707
#12 0x0000559e3b5f27b9 in BlueStore::queue_transactions (this=0x559e3e554000, ch=..., tls=std::vector of lengt
h 1, capacity 1 = {...}, op=..., handle=0x7f5523b37130) at /usr/src/debug/ceph-13.2.0/src/os/bluestore/BlueSto
re.cc:9479
#13 0x0000559e3b2038d0 in ObjectStore::queue_transaction(boost::intrusive_ptr<ObjectStore::CollectionImpl>&, O
bjectStore::Transaction&&, boost::intrusive_ptr<TrackedOp>, ThreadPool::TPHandle*) (this=0x559e3e554000, ch=..
.,
    t=<optimized out>, op=..., handle=0x7f5523b37130) at /usr/src/debug/ceph-13.2.0/src/os/ObjectStore.h:1435
#14 0x0000559e3b19a788 in OSD::dispatch_context_transaction (this=this@entry=0x559e3e6c2000, ctx=..., pg=pg@en
try=0x559e410fb000, handle=handle@entry=0x7f5523b37130) at /usr/src/debug/ceph-13.2.0/src/osd/OSD.cc:8222
#15 0x0000559e3b1c823e in OSD::dequeue_peering_evt (this=0x559e3e6c2000, sdata=0x559e3e6da280, pg=0x559e410fb0
00, evt=std::shared_ptr (count 2, weak 0) 0x559e5257be00, handle=...) at /usr/src/debug/ceph-13.2.0/src/osd/OS
```

```
D.cc:8921
#16 0x0000559e3b41f820 in PGPeeringItem::run (this=<optimized out>, osd=<optimized out>, sdata=<optimized out>
, pg=..., handle=...) at /usr/src/debug/ceph-13.2.0/src/osd/OpQueueItem.cc:34
#17 0x0000559e3b1d2e02 in run (handle=..., pg=..., sdata=<optimized out>, osd=<optimized out>, this=0x7f5523b3
7180) at /usr/src/debug/ceph-13.2.0/src/osd/OpQueueItem.h:134
#18 OSD::ShardedOpWQ::_process (this=0x559e3e6c3050, thread_index=<optimized out>, hb=<optimized out>) at /usr
/src/debug/ceph-13.2.0/src/osd/OSD.cc:9890
#19 0x00007f554695d333 in ShardedThreadPool::shardedthreadpool_worker (this=0x559e3e6c2938, thread_index=<opti
mized out>) at /usr/src/debug/ceph-13.2.0/src/common/WorkQueue.cc:339
#20 0x00007f554695df20 in ShardedThreadPool::WorkThreadSharded::entry (this=<optimized out>) at /usr/src/debug
/ceph-13.2.0/src/common/WorkQueue.h:690
#21 0x00007f55434afe25 in start_thread (arg=0x7f5523b3c700) at pthread_create.c:308
#22 0x00007f55425a0bad in clone () at ../sysdeps/unix/sysv/linux/x86_64/clone.S:113

(gdb) f 6
#6  0x0000559e3b5d4cda in insert (plen=0x0, pstart=0x0, len=<optimized out>, start=<optimized out>, this=0x559
e524a8d80) at /usr/src/debug/ceph-13.2.0/src/include/interval_set.h:460
460               ceph_abort();
(gdb) l
455         } else {
456           if (p->first < start) {
457
458             if (p->first + p->second != start) {
459               //cout << "p is " << p->first << "~" << p->second << ", start is " << start << ", len is " <
< len << endl;
460               ceph_abort();
461             }
462
463             p->second += len;               // append to end
464

(gdb) up
#7  BlueStore::_wctx_finish (this=this@entry=0x559e3e554000, txc=txc@entry=0x559e524a8c00, c=..., o=..., wctx=
wctx@entry=0x7f5523b366a0, maybe_unshared_blobs=maybe_unshared_blobs@entry=0x0)
    at /usr/src/debug/ceph-13.2.0/src/os/bluestore/BlueStore.cc:10757
10757         txc->released.insert(e.offset, e.length);
(gdb) p e
$6 = {static INVALID_OFFSET = 18446744073709551615, offset = 3411935821824, length = 65536}
```

#3 - 07/19/2018 11:02 PM - Brad Hubbard

- File gdb.txt.gz added

Attaching thread dump.

#4 - 07/22/2018 12:07 AM - Troy Ablan

It appears that I have access to all of the pools at this point now that one of the crashing OSDs is staying down, but there are some rbd images that return I/O errors when I try to do rbd export. I think I have backups of everything except for some of what's in CephFS, and that's what I'm most concerned about at this point. If possible, I would like to be able to get this data out, even if it's just to hack/limp it along to copy the files out and then recreate the cluster from scratch. Since I think the CephFS metadata got corrupted in 12.2.6, perhaps I could port some of the 12.2.7 fixes forward?

```
cluster:
    id:     b2873c9a-5539-4c76-ac4a-a6c9829bfed2
    health: HEALTH_ERR
            1 filesystem is degraded
            1 filesystem is offline
            1 mds daemon damaged
            noout,noscrub,nodeep-scrub flag(s) set
            1 osds down
            Degraded data redundancy: 3368046/75477204 objects degraded (4.462%), 100 pgs degraded


services:
    mon: 5 daemons, quorum a,b,c,d,e
    mgr: a(active), standbys: b, d, e, c
    mds: lcs-0/1/1 up , 2 up:standby, 1 damaged
    osd: 34 osds: 33 up, 34 in
         flags noout,noscrub,nodeep-scrub


data:
    pools:   15 pools, 640 pgs
    objects: 9.69 M objects, 13 TiB
    usage:   25 TiB used, 54 TiB / 79 TiB avail
    pgs:     3368046/75477204 objects degraded (4.462%)
             512 active+clean
             100 active+undersized+degraded
             28  active+undersized
```

#5 - 07/22/2018 09:57 PM - Brad Hubbard

- Priority changed from Normal to High

#6 - 07/23/2018 11:03 AM - Igor Fedotov

The root cause for the abort is duplicate pextent(0x31a67390000~10000) that is present at both
"6.3es4_head 4#6:7e10b97a:::rbd_data.5.5ba6874b0dc51.0000000000000132:head#"
and
"6.3es4_head 4#6:7e10b97a:::rbd_data.5.5ba6874b0dc51.0000000000000132:head#50816e"
objects.
It looks like the removal for both takes place within the single transaction context and hence adding the same extent twice asserts.

What caused the duplication is unclear though.

I suppose that fsck should detect such an issue and repair command from ceph-bluestore-tool will most probably fix it. Suggest to share fsck report here prior to repair since that's a pretty risky way and potentially it might be disruptive.

#7 - 07/23/2018 05:49 PM - Troy Ablan

Here's the fsck commandline and its output. Log file has been uploaded to https://mooinglemur.com/2018-07-ceph/

Beware: the log file is 10+ GB uncompressed.

```
-[/scratch:#]- ceph-bluestore-tool fsck -l /scratch/fsck-18.log --log-level=20 --path /var/lib/ceph/osd/ceph-1
8
2018-07-23 17:08:49.943 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: #-1:9e39ad72:::osdmap
.27033:0# extent 0x31a66ec0000~10000 or a subset is already allocated (misreferenced)
2018-07-23 17:09:02.098 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 0#6:4598ef96:::rbd_da
ta.5.5ba6874b0dc51.00000000000001ad:head# extent 0x30553600000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:10:57.442 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 1#6:2dcb2b51:::rbd_da
ta.5.7d86d74b0dc51.000000000000010e:head# extent 0x31a67060000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:12:34.234 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 1#14:df7dbab0:::10000
769367.00000000:head# extent 0x161a42d0000~10000 or a subset is already allocated (misreferenced)
2018-07-23 17:18:41.565 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 4#6:7c87b036:::rbd_da
ta.5.7d86d74b0dc51.0000000000000223:head# extent 0xec9f1a0000~10000 or a subset is already allocated (misrefer
enced)
2018-07-23 17:18:43.828 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 4#6:7e10b97a:::rbd_da
ta.5.5ba6874b0dc51.0000000000000132:head#50816e extent 0x31a67350000~40000 or a subset is already allocated (m
isreferenced)
2018-07-23 17:18:43.828 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 4#6:7e10b97a:::rbd_da
ta.5.5ba6874b0dc51.0000000000000132:head#50816e extent 0x31a672d0000~80000 or a subset is already allocated (m
isreferenced)
2018-07-23 17:18:43.828 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 4#6:7e10b97a:::rbd_da
ta.5.5ba6874b0dc51.0000000000000132:head# nid 68694335 already in use
2018-07-23 17:20:50.230 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 5#6:85ab30db:::rbd_da
ta.5.5ba6874b0dc51.00000000000001b8:head# extent 0x30a620c0000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:21:29.437 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 6#6:66e4eeaf:::rbd_da
ta.5.806ab643c9869.0000000000000113:head# extent 0x3113f4f0000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:21:35.816 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 6#6:6e9fa0d8:::rbd_da
ta.5.7d86d74b0dc51.0000000000000270:head# extent 0x1128a210000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:21:39.450 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 6#6:797d6888:::rbd_da
ta.5.806ab643c9869.0000000000000c88:1cce# extent 0x31f0dbd0000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:21:54.152 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 6#6:c22113bd:::rbd_da
ta.5.806ab643c9869.00000000000001dc:head# extent 0x31e80e40000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:22:03.105 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 6#6:e799b675:::rbd_da
ta.5.7d86d74b0dc51.0000000000000203:head# extent 0x31e5bb30000~10000 or a subset is already allocated (misrefe
```

```
renced)
2018-07-23 17:23:12.283 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 7#6:044edce5:::rbd_da
ta.5.7d86d74b0dc51.0000000000000173:head# extent 0x31a56b70000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:23:13.056 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 7#6:0502dbb7:::rbd_da
ta.5.806ab643c9869.0000000000000143:head# extent 0x16675b50000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:23:20.975 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 7#6:12e32e88:::rbd_da
ta.5.7d86d74b0dc51.00000000000001a6:head# extent 0x30c67a80000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:23:21.963 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 7#6:13842744:::rbd_da
ta.5.5ba6874b0dc51.0000000000000140:head# extent 0x31823500000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:23:24.915 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 7#6:c974bf19:::rbd_da
ta.5.7d86d74b0dc51.0000000000000137:head# extent 0x31e2def0000~10000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:26:10.638 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 4#6:4886eb24:::rbd_da
ta.5.5ba6374b0dc51.0000000000001280:1cb9# extent 0x313acd60000~80000 or a subset is already allocated (misrefe
renced)
2018-07-23 17:26:41.755 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: actual store_statfs(0
x22383f3f000/0x3223b1d1000, stored 0xe3b582983a/0xfd76130000, compress 0xa2a780f6/0x1433c0000/0x287040000) !=
expected store_statfs(0x22383f3f000/0x3223b1d1000, stored 0xe3b6dc883a/0xfd77720000, compress 0xa2a780f6/0x143
3c0000/0x287040000)
2018-07-23 17:27:10.595 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: free extent 0xab97450
000~60000 intersects allocated blocks
2018-07-23 17:27:31.270 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: free extent 0x3140c78
0000~630000 intersects allocated blocks
2018-07-23 17:27:31.502 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: free extent 0x31ae17a
0000~bc0000 intersects allocated blocks
2018-07-23 17:27:31.820 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0xab973
90000~c0000
2018-07-23 17:27:31.822 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0xec9f0
e0000~c0000
2018-07-23 17:27:31.825 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x1128a
150000~c0000
2018-07-23 17:27:31.826 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x161a4
210000~c0000
2018-07-23 17:27:31.826 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x16675
a90000~c0000
2018-07-23 17:27:31.842 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x1757b
1b0000~c0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x30553
540000~c0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x30a62
000000~c0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x30c67
9c0000~c0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x3113f
430000~c0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x313ac
ca0000~c0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x313be
1f0000~d0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x3140c
6c0000~c0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31823
440000~c0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x319f5
830000~d0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x319f7
bb0000~d0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31a56
ab0000~c0000
2018-07-23 17:27:31.843 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31a66
e00000~c0000
2018-07-23 17:27:31.844 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31a66
ed0000~190000
2018-07-23 17:27:31.844 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31ae1
6e0000~c0000
2018-07-23 17:27:31.844 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31e2d
e30000~c0000
2018-07-23 17:27:31.844 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31e5a
900000~d0000
2018-07-23 17:27:31.844 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31e5b
a70000~c0000
```

```
2018-07-23 17:27:31.844 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31e5cf20000~d0000
2018-07-23 17:27:31.844 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31e80d80000~c0000
2018-07-23 17:27:31.844 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31f0db10000~c0000
2018-07-23 17:27:31.844 7f8e17f66a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31f0f040000~d0000
fsck success
```

#8 - 07/27/2018 09:53 PM - Troy Ablan

I have since updated to 13.2.1 and the cluster appears to be allowing me to get at least some of my data out. However, I then tried fsck followed by repair and I'm ending up with a similar fsck output and a failing repair. Logs and core can be found at https://mooinglemur.com/2018-07-ceph/

```
-[/var/log/ceph:#]- ceph-bluestore-tool fsck -l /scratch/2018-07-27-fsck-18.log --log-level=20 --path /var/lib/ceph/osd/ceph-18
2018-07-27 20:54:45.093 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: #-1:9e39ad72:::osdmap.27033:0# extent 0x31a66ec0000~10000 or a subset is already allocated (misreferenced)
2018-07-27 20:54:59.670 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 0#6:4598ef96:::rbd_data.5.5ba6874b0dc51.00000000000001ad:head# extent 0x30553600000~10000 or a subset is already allocated (misreferenced)
2018-07-27 20:57:00.065 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 1#6:2dcb2b51:::rbd_data.5.7d86d74b0dc51.000000000000010e:head# extent 0x31a67060000~10000 or a subset is already allocated (misreferenced)
2018-07-27 20:58:37.864 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 1#14:df7dbab0:::10000769367.00000000:head# extent 0x161a42d0000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:04:43.035 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 4#6:7c87b036:::rbd_data.5.7d86d74b0dc51.0000000000000223:head# extent 0xec9f1a0000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:04:45.404 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 4#6:7e10b97a:::rbd_data.5.5ba6874b0dc51.0000000000000132:head#50816e extent 0x31a67350000~40000 or a subset is already allocated (misreferenced)
2018-07-27 21:04:45.404 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 4#6:7e10b97a:::rbd_data.5.5ba6874b0dc51.0000000000000132:head#50816e extent 0x31a672d0000~80000 or a subset is already allocated (misreferenced)
2018-07-27 21:04:45.404 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 4#6:7e10b97a:::rbd_data.5.5ba6874b0dc51.0000000000000132:head# nid 68694335 already in use
2018-07-27 21:06:50.509 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 5#6:85ab30db:::rbd_data.5.5ba6874b0dc51.00000000000001b8:head# extent 0x30a620c0000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:07:30.040 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 6#6:66e4eeaf:::rbd_data.5.806ab643c9869.0000000000000113:head# extent 0x3113f4f0000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:07:36.536 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 6#6:6e9fa0d8:::rbd_data.5.7d86d74b0dc51.0000000000000270:head# extent 0x1128a210000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:07:39.985 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 6#6:797d6888:::rbd_data.5.806ab643c9869.0000000000000c88:1cce# extent 0x31f0dbd0000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:07:52.126 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 6#6:c22113bd:::rbd_data.5.806ab643c9869.00000000000001dc:head# extent 0x31e80e40000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:08:00.818 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 6#6:e799b675:::rbd_data.5.7d86d74b0dc51.0000000000000203:head# extent 0x31e5bb30000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:09:17.408 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 7#6:044edce5:::rbd_data.5.7d86d74b0dc51.0000000000000173:head# extent 0x31a56b70000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:09:18.142 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 7#6:0502dbb7:::rbd_data.5.806ab643c9869.0000000000000143:head# extent 0x16675b50000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:09:26.503 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 7#6:12e32e88:::rbd_data.5.7d86d74b0dc51.00000000000001a6:head# extent 0x30c67a80000~10000 or a subset is already allocated (misreferenced)
```

```
2018-07-27 21:09:27.299 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 7#6:13842744:::rbd_da
ta.5.5ba6874b0dc51.0000000000000140:head# extent 0x31823500000~10000 or a subset is already allocated (misrefe
renced)
2018-07-27 21:09:30.357 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 7#6:c974bf19:::rbd_da
ta.5.7d86d74b0dc51.0000000000000137:head# extent 0x31e2def0000~10000 or a subset is already allocated (misrefe
renced)
2018-07-27 21:12:11.569 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 4#6:4886eb24:::rbd_da
ta.5.5ba6374b0dc51.0000000000001280:1cb9# extent 0x313acd60000~80000 or a subset is already allocated (misrefe
renced)
2018-07-27 21:12:43.145 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: actual store_statfs(0
x223864af000/0x3223b1d1000, stored 0xe3b5e52141/0xfd745c0000, compress 0xa2a780f6/0x1433c0000/0x287040000) !=
expected store_statfs(0x223864af000/0x3223b1d1000, stored 0xe3b73f1141/0xfd75bb0000, compress 0xa2a780f6/0x143
3c0000/0x287040000)
2018-07-27 21:13:14.890 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: free extent 0xab97450
000~60000 intersects allocated blocks
2018-07-27 21:13:38.657 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: free extent 0x3140c78
0000~630000 intersects allocated blocks
2018-07-27 21:13:39.002 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: free extent 0x31ae17a
0000~bc0000 intersects allocated blocks
2018-07-27 21:13:39.318 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0xab973
90000~c0000
2018-07-27 21:13:39.320 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0xec9f0
e0000~c0000
2018-07-27 21:13:39.323 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x1128a
150000~c0000
2018-07-27 21:13:39.323 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x161a4
210000~c0000
2018-07-27 21:13:39.324 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x16675
a90000~c0000
2018-07-27 21:13:39.341 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x1757b
1b0000~c0000
2018-07-27 21:13:39.341 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x30553
540000~c0000
2018-07-27 21:13:39.341 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x30a62
000000~c0000
2018-07-27 21:13:39.341 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x30c67
9c0000~c0000
2018-07-27 21:13:39.341 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x3113f
430000~c0000
2018-07-27 21:13:39.341 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x313ac
ca0000~c0000
2018-07-27 21:13:39.341 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x313be
1f0000~d0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x3140c
6c0000~c0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31823
440000~c0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x319f5
830000~d0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x319f7
bb0000~d0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31a56
ab0000~c0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31a66
e00000~c0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31a66
ed0000~190000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31ae1
6e0000~c0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31e2d
e30000~c0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31e5a
900000~d0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31e5b
a70000~c0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31e5c
f20000~d0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31e80
d80000~c0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31f0d
b10000~c0000
2018-07-27 21:13:39.342 7f57b4c1fa00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: leaked extent 0x31f0f
040000~d0000
fsck success
```

```
-[/scratch/ceph:#]- ceph-bluestore-tool repair -l /scratch/2018-07-27-repair-18.log --log-level=20 --path /var
/lib/ceph/osd/ceph-18
2018-07-27 21:30:03.771 7f261bea7a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: #-1:9e39ad72:::osdmap
.27033:0# extent 0x31a66ec0000~10000 or a subset is already allocated (misreferenced)
2018-07-27 21:30:16.784 7f261bea7a00 -1 bluestore(/var/lib/ceph/osd/ceph-18) fsck error: 0#6:4598ef96:::rbd_da
ta.5.5ba6874b0dc51.00000000000001ad:head# extent 0x30553600000~10000 or a subset is already allocated (misrefe
renced)
/home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/c
entos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/common/bloom_filter.hpp: In function 'vo
id bloom_filter::insert(uint32_t)' thread 7f261bea7a00 time 2018-07-27 21:31:34.163644
/home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/c
entos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/common/bloom_filter.hpp: 159: FAILED ass
ert(bit_table_)
 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0xff) [0x7f2612239e1f]
 2: (()+0x284fe7) [0x7f2612239fe7]
 3: (BlueStore::_fsck_check_extents(coll_t const&, ghobject_t const&, std::vector<bluestore_pextent_t, mempool
::pool_allocator<(mempool::pool_index_t)4, bluestore_pextent_t> > const&, bool, boost::dynamic_bitset<unsigned
 long, mempool::pool_allocator<(mempool::pool_index_t)5, unsigned long> >&, unsigned long, BlueStoreRepairer*,
 store_statfs_t&)+0x4aa) [0x55f43ab265fa]
 4: (BlueStore::_fsck(bool, bool)+0x36b6) [0x55f43ab91f66]
 5: (main()+0xea6) [0x55f43aa184f6]
 6: (__libc_start_main()+0xf5) [0x7f260ee51445]
 7: (()+0x2365ff) [0x55f43aaea5ff]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.
2018-07-27 21:31:34.164 7f261bea7a00 -1 /home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_A
RCH/x86_64/AVAILABLE_DIST/centos7/DIST/centos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/
common/bloom_filter.hpp: In function 'void bloom_filter::insert(uint32_t)' thread 7f261bea7a00 time 2018-07-27
 21:31:34.163644
/home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/c
entos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/common/bloom_filter.hpp: 159: FAILED ass
ert(bit_table_)

 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0xff) [0x7f2612239e1f]
 2: (()+0x284fe7) [0x7f2612239fe7]
 3: (BlueStore::_fsck_check_extents(coll_t const&, ghobject_t const&, std::vector<bluestore_pextent_t, mempool
::pool_allocator<(mempool::pool_index_t)4, bluestore_pextent_t> > const&, bool, boost::dynamic_bitset<unsigned
 long, mempool::pool_allocator<(mempool::pool_index_t)5, unsigned long> >&, unsigned long, BlueStoreRepairer*,
 store_statfs_t&)+0x4aa) [0x55f43ab265fa]
 4: (BlueStore::_fsck(bool, bool)+0x36b6) [0x55f43ab91f66]
 5: (main()+0xea6) [0x55f43aa184f6]
 6: (__libc_start_main()+0xf5) [0x7f260ee51445]
 7: (()+0x2365ff) [0x55f43aaea5ff]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.

     0> 2018-07-27 21:31:34.164 7f261bea7a00 -1 /home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVA
ILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/centos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.
2.1/src/common/bloom_filter.hpp: In function 'void bloom_filter::insert(uint32_t)' thread 7f261bea7a00 time 20
18-07-27 21:31:34.163644
/home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/c
entos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/common/bloom_filter.hpp: 159: FAILED ass
ert(bit_table_)

 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0xff) [0x7f2612239e1f]
 2: (()+0x284fe7) [0x7f2612239fe7]
 3: (BlueStore::_fsck_check_extents(coll_t const&, ghobject_t const&, std::vector<bluestore_pextent_t, mempool
::pool_allocator<(mempool::pool_index_t)4, bluestore_pextent_t> > const&, bool, boost::dynamic_bitset<unsigned
 long, mempool::pool_allocator<(mempool::pool_index_t)5, unsigned long> >&, unsigned long, BlueStoreRepairer*,
 store_statfs_t&)+0x4aa) [0x55f43ab265fa]
 4: (BlueStore::_fsck(bool, bool)+0x36b6) [0x55f43ab91f66]
 5: (main()+0xea6) [0x55f43aa184f6]
 6: (__libc_start_main()+0xf5) [0x7f260ee51445]
 7: (()+0x2365ff) [0x55f43aaea5ff]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.

*** Caught signal (Aborted) **
 in thread 7f261bea7a00 thread_name:ceph-bluestore-
 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (()+0x466e90) [0x55f43ad1ae90]
```

```
 2: (()+0xf6d0) [0x7f261047a6d0]
 3: (gsignal()+0x37) [0x7f260ee65277]
 4: (abort()+0x148) [0x7f260ee66968]
 5: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0x242) [0x7f2612239f62]
 6: (()+0x284fe7) [0x7f2612239fe7]
 7: (BlueStore::_fsck_check_extents(coll_t const&, ghobject_t const&, std::vector<bluestore_pextent_t, mempool
::pool_allocator<(mempool::pool_index_t)4, bluestore_pextent_t> > const&, bool, boost::dynamic_bitset<unsigned
 long, mempool::pool_allocator<(mempool::pool_index_t)5, unsigned long> >&, unsigned long, BlueStoreRepairer*,
 store_statfs_t&)+0x4aa) [0x55f43ab265fa]
 8: (BlueStore::_fsck(bool, bool)+0x36b6) [0x55f43ab91f66]
 9: (main()+0xea6) [0x55f43aa184f6]
 10: (__libc_start_main()+0xf5) [0x7f260ee51445]
 11: (()+0x2365ff) [0x55f43aaea5ff]
2018-07-27 21:31:34.199 7f261bea7a00 -1 *** Caught signal (Aborted) **
 in thread 7f261bea7a00 thread_name:ceph-bluestore-

 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (()+0x466e90) [0x55f43ad1ae90]
 2: (()+0xf6d0) [0x7f261047a6d0]
 3: (gsignal()+0x37) [0x7f260ee65277]
 4: (abort()+0x148) [0x7f260ee66968]
 5: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0x242) [0x7f2612239f62]
 6: (()+0x284fe7) [0x7f2612239fe7]
 7: (BlueStore::_fsck_check_extents(coll_t const&, ghobject_t const&, std::vector<bluestore_pextent_t, mempool
::pool_allocator<(mempool::pool_index_t)4, bluestore_pextent_t> > const&, bool, boost::dynamic_bitset<unsigned
 long, mempool::pool_allocator<(mempool::pool_index_t)5, unsigned long> >&, unsigned long, BlueStoreRepairer*,
 store_statfs_t&)+0x4aa) [0x55f43ab265fa]
 8: (BlueStore::_fsck(bool, bool)+0x36b6) [0x55f43ab91f66]
 9: (main()+0xea6) [0x55f43aa184f6]
 10: (__libc_start_main()+0xf5) [0x7f260ee51445]
 11: (()+0x2365ff) [0x55f43aaea5ff]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.

     0> 2018-07-27 21:31:34.199 7f261bea7a00 -1 *** Caught signal (Aborted) **
 in thread 7f261bea7a00 thread_name:ceph-bluestore-

 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (()+0x466e90) [0x55f43ad1ae90]
 2: (()+0xf6d0) [0x7f261047a6d0]
 3: (gsignal()+0x37) [0x7f260ee65277]
 4: (abort()+0x148) [0x7f260ee66968]
 5: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0x242) [0x7f2612239f62]
 6: (()+0x284fe7) [0x7f2612239fe7]
 7: (BlueStore::_fsck_check_extents(coll_t const&, ghobject_t const&, std::vector<bluestore_pextent_t, mempool
::pool_allocator<(mempool::pool_index_t)4, bluestore_pextent_t> > const&, bool, boost::dynamic_bitset<unsigned
 long, mempool::pool_allocator<(mempool::pool_index_t)5, unsigned long> >&, unsigned long, BlueStoreRepairer*,
 store_statfs_t&)+0x4aa) [0x55f43ab265fa]
 8: (BlueStore::_fsck(bool, bool)+0x36b6) [0x55f43ab91f66]
 9: (main()+0xea6) [0x55f43aa184f6]
 10: (__libc_start_main()+0xf5) [0x7f260ee51445]
 11: (()+0x2365ff) [0x55f43aaea5ff]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.

Aborted (core dumped)
```

#9 - 07/27/2018 10:01 PM - Troy Ablan

repair log and core respectively

ceph-post-file: 189d2970-662a-49c0-ac6c-6ee6d124d523
ceph-post-file: 8675da71-1bfd-44cd-901e-0e0899754683

#10 - 07/28/2018 08:07 PM - Troy Ablan

It looks like I have quite a few more than this one OSD that's crashing. They all fsck successfully, but repair will core. Here is an example of another one with a different assert message.

Uploaded the repair log and core file respectively.

ceph-post-file: 4853e3e8-9191-42d9-be5a-2af6a4c76c0b
ceph-post-file: 711eadf8-8257-4670-a596-4b16da766c12

```
-[/scratch:#]- ceph-bluestore-tool repair -l /scratch/2018-07-28-repair-8.log --log-level=20 --path /var/lib/c
eph/osd/ceph-8


2018-07-28 19:55:02.968 7f6d5cc56a00 -1 bluestore(/var/lib/ceph/osd/ceph-8) fsck error: 2#6:2d56c8b8:::rbd_dat
a.5.5ba6874b0dc51.000000000000020d:head# extent 0xd0cac20000~10000 or a subset is already allocated (misrefere
nced)
2018-07-28 19:57:53.308 7f6d5cc56a00 -1 bluestore(/var/lib/ceph/osd/ceph-8) fsck error: 2#6:ef6f1c45:::rbd_dat
a.5.2a9052ae8944a.0000000000000cb0:1f73# extent 0x237c3b30000~80000 or a subset is already allocated (misrefer
enced)
2018-07-28 19:57:53.315 7f6d5cc56a00 -1 bluestore(/var/lib/ceph/osd/ceph-8) fsck error: 2#6:ef841d62:::rbd_dat
a.5.36b822ae8944a.0000000000000880:1923# extent 0x9a28b10000~10000 or a subset is already allocated (misrefere
nced)
/home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/c
entos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/os/bluestore/BlueStore.cc: In function '
size_t BlueStoreRepairer::StoreSpaceTracker::filter_out(const interval_set<long unsigned int>&)' thread 7f6d5c
c56a00 time 2018-07-28 19:5
7:53.556030
/home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/c
entos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/os/bluestore/BlueStore.cc: 12252: FAILED
 assert(collections_bfs[pos].element_count() == objects_bfs[pos].element_count())
 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0xff) [0x7f6d52febe1f]
 2: (()+0x284fe7) [0x7f6d52febfe7]
 3: (BlueStoreRepairer::StoreSpaceTracker::filter_out(interval_set<unsigned long, std::map<unsigned long, unsi
gned long, std::less<unsigned long>, std::allocator<std::pair<unsigned long const, unsigned long> > > > const&
)+0x4be) [0x55f561965d4e]
 4: (BlueStoreRepairer::preprocess_misreference(KeyValueDB*)+0x56) [0x55f561965e06]
 5: (BlueStore::_fsck(bool, bool)+0x77d9) [0x55f5619a1089]
 6: (main()+0xea6) [0x55f5618234f6]
 7: (__libc_start_main()+0xf5) [0x7f6d4fc03445]
 8: (()+0x2365ff) [0x55f5618f55ff]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.
2018-07-28 19:57:53.583 7f6d5cc56a00 -1 /home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_A
RCH/x86_64/AVAILABLE_DIST/centos7/DIST/centos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/
os/bluestore/BlueStore.cc: In function 'size_t BlueStoreRepairer::StoreSpaceTracker::filter_out(const interval
_set<long unsigned int>&)'
thread 7f6d5cc56a00 time 2018-07-28 19:57:53.556030
/home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/c
entos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/os/bluestore/BlueStore.cc: 12252: FAILED
 assert(collections_bfs[pos].element_count() == objects_bfs[pos].element_count())

 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0xff) [0x7f6d52febe1f]
 2: (()+0x284fe7) [0x7f6d52febfe7]
 3: (BlueStoreRepairer::StoreSpaceTracker::filter_out(interval_set<unsigned long, std::map<unsigned long, unsi
gned long, std::less<unsigned long>, std::allocator<std::pair<unsigned long const, unsigned long> > > > const&
)+0x4be) [0x55f561965d4e]
 4: (BlueStoreRepairer::preprocess_misreference(KeyValueDB*)+0x56) [0x55f561965e06]
 5: (BlueStore::_fsck(bool, bool)+0x77d9) [0x55f5619a1089]
 6: (main()+0xea6) [0x55f5618234f6]
 7: (__libc_start_main()+0xf5) [0x7f6d4fc03445]
 8: (()+0x2365ff) [0x55f5618f55ff]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.
```

```
     0> 2018-07-28 19:57:53.583 7f6d5cc56a00 -1 /home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVA
ILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/centos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.
2.1/src/os/bluestore/BlueStore.cc: In function 'size_t BlueStoreRepairer::StoreSpaceTracker::filter_out(const
 interval_set<long unsigned
int>&)' thread 7f6d5cc56a00 time 2018-07-28 19:57:53.556030
/home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/c
entos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/os/bluestore/BlueStore.cc: 12252: FAILED
 assert(collections_bfs[pos].element_count() == objects_bfs[pos].element_count())

 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0xff) [0x7f6d52febe1f]
 2: (()+0x284fe7) [0x7f6d52febfe7]
 3: (BlueStoreRepairer::StoreSpaceTracker::filter_out(interval_set<unsigned long, std::map<unsigned long, unsi
gned long, std::less<unsigned long>, std::allocator<std::pair<unsigned long const, unsigned long> > > > const&
)+0x4be) [0x55f561965d4e]
 4: (BlueStoreRepairer::preprocess_misreference(KeyValueDB*)+0x56) [0x55f561965e06]
 5: (BlueStore::_fsck(bool, bool)+0x77d9) [0x55f5619a1089]
 6: (main()+0xea6) [0x55f5618234f6]
 7: (__libc_start_main()+0xf5) [0x7f6d4fc03445]
 8: (()+0x2365ff) [0x55f5618f55ff]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.

*** Caught signal (Aborted) **
 in thread 7f6d5cc56a00 thread_name:ceph-bluestore-
 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (()+0x466e90) [0x55f561b25e90]
 2: (()+0xf6d0) [0x7f6d5122c6d0]
 3: (gsignal()+0x37) [0x7f6d4fc17277]
 4: (abort()+0x148) [0x7f6d4fc18968]
 5: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0x242) [0x7f6d52febf62]
 6: (()+0x284fe7) [0x7f6d52febfe7]
 7: (BlueStoreRepairer::StoreSpaceTracker::filter_out(interval_set<unsigned long, std::map<unsigned long, unsi
gned long, std::less<unsigned long>, std::allocator<std::pair<unsigned long const, unsigned long> > > > const&
)+0x4be) [0x55f561965d4e]
 8: (BlueStoreRepairer::preprocess_misreference(KeyValueDB*)+0x56) [0x55f561965e06]
 9: (BlueStore::_fsck(bool, bool)+0x77d9) [0x55f5619a1089]
 10: (main()+0xea6) [0x55f5618234f6]
 11: (__libc_start_main()+0xf5) [0x7f6d4fc03445]
 12: (()+0x2365ff) [0x55f5618f55ff]
2018-07-28 19:57:53.603 7f6d5cc56a00 -1 *** Caught signal (Aborted) **
 in thread 7f6d5cc56a00 thread_name:ceph-bluestore-

 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (()+0x466e90) [0x55f561b25e90]
 2: (()+0xf6d0) [0x7f6d5122c6d0]
 3: (gsignal()+0x37) [0x7f6d4fc17277]
 4: (abort()+0x148) [0x7f6d4fc18968]
 5: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0x242) [0x7f6d52febf62]
 6: (()+0x284fe7) [0x7f6d52febfe7]
 7: (BlueStoreRepairer::StoreSpaceTracker::filter_out(interval_set<unsigned long, std::map<unsigned long, unsi
gned long, std::less<unsigned long>, std::allocator<std::pair<unsigned long const, unsigned long> > > > const&
)+0x4be) [0x55f561965d4e]
 8: (BlueStoreRepairer::preprocess_misreference(KeyValueDB*)+0x56) [0x55f561965e06]
 9: (BlueStore::_fsck(bool, bool)+0x77d9) [0x55f5619a1089]
 10: (main()+0xea6) [0x55f5618234f6]
 11: (__libc_start_main()+0xf5) [0x7f6d4fc03445]
 12: (()+0x2365ff) [0x55f5618f55ff]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.

     0> 2018-07-28 19:57:53.603 7f6d5cc56a00 -1 *** Caught signal (Aborted) **
 in thread 7f6d5cc56a00 thread_name:ceph-bluestore-

 ceph version 13.2.1 (5533ecdc0fda920179d7ad84e0aa65a127b20d77) mimic (stable)
 1: (()+0x466e90) [0x55f561b25e90]
 2: (()+0xf6d0) [0x7f6d5122c6d0]
 3: (gsignal()+0x37) [0x7f6d4fc17277]
 4: (abort()+0x148) [0x7f6d4fc18968]
 5: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0x242) [0x7f6d52febf62]
 6: (()+0x284fe7) [0x7f6d52febfe7]
 7: (BlueStoreRepairer::StoreSpaceTracker::filter_out(interval_set<unsigned long, std::map<unsigned long, unsi
gned long, std::less<unsigned long>, std::allocator<std::pair<unsigned long const, unsigned long> > > > const&
)+0x4be) [0x55f561965d4e]
 8: (BlueStoreRepairer::preprocess_misreference(KeyValueDB*)+0x56) [0x55f561965e06]
 9: (BlueStore::_fsck(bool, bool)+0x77d9) [0x55f5619a1089]
 10: (main()+0xea6) [0x55f5618234f6]
```

```
 11: (__libc_start_main()+0xf5) [0x7f6d4fc03445]
 12: (()+0x2365ff) [0x55f5618f55ff]
 NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.

Aborted (core dumped)
```

#11 - 08/06/2018 03:15 AM - Troy Ablan

Wanted to update that I

1. Have gotten 100% of the data out of the cluster, and as far as I can tell, everything is intact.
2. Am currently paving over all of the OSDs to reinitialize the cluster.

Before reinitializing everything, I removed all pools and all OSDs came up stably. At this point, I let everything peer for a while and let the OSDs do what they do.
Before I took the crashy ones down, and ran a bluestore fsck. fsck came back with similar errors to before despite having no PGs.

Priority can certainly be lowered on this issue at this point.

Thanks for the attention so far. If I can help with finding root cause with additional configuration details, I'll certainly try.

#12 - 08/06/2018 03:16 AM - Troy Ablan

Troy Ablan wrote:

> Before I took the crashy ones down, and ran a bluestore fsck. fsck came back with similar errors to before despite having no PGs.

This should read

> Then I took the crashy ones down, and ran a bluestore fsck. fsck came back with similar errors to before despite having no PGs.

#13 - 09/04/2018 11:38 PM - Radoslaw Zarzynski

I'm afraid the corruption can be caused by the racy SharedBlob::put() (fixed since 12.2.6). There are scenarios where a corrupted SharedBlob instance is used for a while and can reach RocksDB. In the spot of this particular bug I'm focusing on potential impact on unsharing in BlueStore::_do_remove and the BlueStore::_wctx_finish.

```
void BlueStore::_wctx_finish(
  TransContext *txc,
```

```
  CollectionRef& c,
  OnodeRef o,
  WriteContext *wctx,
  set<SharedBlob*> *maybe_unshared_blobs)
{
  auto oep = wctx->old_extents.begin();
  while (oep != wctx->old_extents.end()) {
    auto &lo = *oep;
    oep = wctx->old_extents.erase(oep);
    dout(20) << __func__ << " lex_old " << lo.e << dendl;
    BlobRef b = lo.e.blob;
    const bluestore_blob_t& blob = b->get_blob();

    // ...
    auto& r = lo.r;
    // ...
    if (!r.empty()) {
      dout(20) << __func__ << "  blob release " << r << dendl;
      if (blob.is_shared()) {
        // ...
        r.clear();
        r.swap(final);
      }
    }

    // ...
    for (auto e : r) {
      dout(20) << __func__ << "  release " << e << dendl;
      txc->released.insert(e.offset, e.length);
    }
  }
}

// ...

int BlueStore::_do_remove(
  TransContext *txc,
  CollectionRef& c,
  OnodeRef o)
{
  set<SharedBlob*> maybe_unshared_blobs;
  bool is_gen = !o->oid.is_no_gen();
  _do_truncate(txc, c, o, 0, is_gen ? &maybe_unshared_blobs : nullptr);

  // ...

  // see if we can unshare blobs still referenced by the head
  dout(10) << __func__ << " gen and maybe_unshared_blobs "
           << maybe_unshared_blobs << dendl;
  ghobject_t nogen = o->oid;
  nogen.generation = ghobject_t::NO_GEN;
  OnodeRef h = c->onode_map.lookup(nogen);

  if (!h || !h->exists) {
    return 0;
  }

  dout(20) << __func__ << " checking for unshareable blobs on " << h
           << " " << h->oid << dendl;
  map<SharedBlob*,bluestore_extent_ref_map_t> expect;
  for (auto& e : h->extent_map.extent_map) {
    const bluestore_blob_t& b = e.blob->get_blob();
    SharedBlob *sb = e.blob->shared_blob.get();
    if (b.is_shared() &&
        sb->loaded &&
        maybe_unshared_blobs.count(sb)) {
      if (b.is_compressed()) {
        expect[sb].get(0, b.get_ondisk_length());
      } else {
        b.map(e.blob_offset, e.length, [&](uint64_t off, uint64_t len) {
            expect[sb].get(off, len);
            return 0;
          });
      }
    }
```

```
  }

  vector<SharedBlob*> unshared_blobs;
  unshared_blobs.reserve(maybe_unshared_blobs.size());
  for (auto& p : expect) {
    dout(20) << " ? " << *p.first << " vs " << p.second << dendl;
    if (p.first->persistent->ref_map == p.second) {
      SharedBlob *sb = p.first;
      dout(20) << __func__ << "  unsharing " << *sb << dendl;
      unshared_blobs.push_back(sb);
      txc->unshare_blob(sb);
      uint64_t sbid = c->make_blob_unshared(sb);
      string key;
      get_shared_blob_key(sbid, &key);
      txc->t->rmkey(PREFIX_SHARED_BLOB, key);
    }
  }

  if (unshared_blobs.empty()) {
    return 0;
  }

  for (auto& e : h->extent_map.extent_map) {
    const bluestore_blob_t& b = e.blob->get_blob();
    SharedBlob *sb = e.blob->shared_blob.get();
    if (b.is_shared() &&
        std::find(unshared_blobs.begin(), unshared_blobs.end(),
                  sb) != unshared_blobs.end()) {
      dout(20) << __func__ << "  unsharing " << e << dendl;
      bluestore_blob_t& blob = e.blob->dirty_blob();
      blob.clear_flag(bluestore_blob_t::FLAG_SHARED);
      h->extent_map.dirty_range(e.logical_offset, 1);
    }
  }
  txc->write_onode(h);

  return 0;
}
```

If so, we would need to verify/improve fsck.

#14 - 09/10/2018 05:35 PM - Stefan Priebe

+1 i have the same crashes running 12.2.7. Is there anything i can do now?

#15 - 09/10/2018 05:53 PM - Radoslaw Zarzynski

Stefan, has the crashing OSD seen anything older than 12.2.6?

fsck is supposed to help in such cases. Take a look on comment #6 from Igor. As it has been said, it's good idea to provide a log (debug_bluestore=20) before repairing. I would expect a few misreferenced/leaked SharedBlobs as a result of the race in <12.2.6. An alternative is to nuke the OSD.

#16 - 09/10/2018 06:00 PM - Stefan Priebe

Yes 12.2.2 and 12.2.5. So i sould just run fsck on the affected osd? Do i need 12.2.8 before?

#17 - 09/10/2018 06:05 PM - Radoslaw Zarzynski

Let's start from the plain fsck on the affected OSD and 12.2.7.

#18 - 09/10/2018 07:09 PM - Stefan Priebe

1. ceph-bluestore-tool fsck --path /var/lib/ceph/osd/ceph-36
   fsck success

there was no more output?

#19 - 09/11/2018 05:22 PM - Radoslaw Zarzynski

fsck has found nothing:

```
    if (action == "fsck") {
      r = bluestore.fsck(fsck_deep);
    } else {
      r = bluestore.repair(fsck_deep);
    }
    if (r < 0) {
      cerr << "error from fsck: " << cpp_strerror(r) << std::endl;
      exit(EXIT_FAILURE);
    }
    cout << action << " success" << std::endl;
  }
```

Interesting! Could you please provide log with debug_bluestore=20 from a corresponding OSD crash?

#20 - 09/11/2018 05:31 PM - Stefan Priebe

i don't know how to crash the OSD it happens once out of nothing... and i think i can't run debug_bluestore for a few days...

#21 - 09/11/2018 06:13 PM - Radoslaw Zarzynski

I see. What about in-memory logging (debug_bluestore=0/20)? Only predefined number of recent debugs is stored and they are written to log file when a crash happens. Still, performance would be impacted.

#22 - 09/13/2018 11:09 AM - Igor Fedotov

Stefan, just in case - would you mind to share existing crash logs, please? Given that fsck detects no error it might be a different case...

#23 - 09/13/2018 01:31 PM - Stefan Priebe

Sure. Currently i don't get a new segault.

Here we go:
Sep 07 18:46:53 cloud1-1468 ceph-osd[11765]: **** Caught signal (Segmentation fault) ***
Sep 07 18:46:53 cloud1-1468 ceph-osd[11765]: in thread 7f7391ffe700 thread_name:tp_osd_disk

and another one:

Sep 07 16:38:07 cloud1-1468 ceph-osd[2404]: 2018-09-07 16:38:07.700240 7ff32d58ce00 -1 osd.19 1186786 log_to_monitors {default=true}
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: **** Caught signal (Segmentation fault) ***
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: in thread 7ff2c7fff700 thread_name:tp_osd_disk
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: ceph version 12.2.7-11-gc4753e57fe (c4753e57feb768e9584a7580b67ac072e82a052e) luminous
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 1: (()+0xa38a94) [0x557c33cf6a94]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 2: (()+0x110c0) [0x7ff32abcb0c0]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 3: (BlueStore::_txc_write_nodes(BlueStore::TransContext, std::shared_ptr<KeyValueDB::T
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 4: (BlueStore::queue_transactions(ObjectStore::Sequencer*, std::vector<ObjectStore::Tra
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 5: (ObjectStore::queue_transaction(ObjectStore::Sequencer*, ObjectStore::Transaction&&,
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 6: (remove_dir(CephContext*, ObjectStore*, SnapMapper*, OSDriver*, ObjectStore::Sequenc
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 7: (OSD::RemoveWQ::_process(std::pair<boost::intrusive_ptr<PG>, std::shared_ptr<Deletin
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 8: (ThreadPool::WorkQueueVal<std::pair<boost::intrusive_ptr<PG>, std::shared_ptr<Deleti
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 9: (ThreadPool::worker(ThreadPool::WorkThread*)+0xeb8) [0x557c33d459b8]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 10: (ThreadPool::WorkThread::entry()+0x10) [0x557c33d46b50]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 11: (()+0x7494) [0x7ff32abc1494]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 12: (clone()+0x3f) [0x7ff329c48acf]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 2018-09-07 19:14:35.020256 7ff2c7fff700 -1 Caught signal (Segmentation fault)
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: in thread 7ff2c7fff700 thread_name:tp_osd_disk
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: ceph version 12.2.7-11-gc4753e57fe (c4753e57feb768e9584a7580b67ac072e82a052e) luminous
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 1: (()+0xa38a94) [0x557c33cf6a94]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 2: (()+0x110c0) [0x7ff32abcb0c0]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 3: (BlueStore::_txc_write_nodes(BlueStore::TransContext, std::shared_ptr<KeyValueDB::T
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 4: (BlueStore::queue_transactions(ObjectStore::Sequencer*, std::vector<ObjectStore::Tra
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 5: (ObjectStore::queue_transaction(ObjectStore::Sequencer*, ObjectStore::Transaction&&,
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 6: (remove_dir(CephContext*, ObjectStore*, SnapMapper*, OSDriver*, ObjectStore::Sequenc
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 7: (OSD::RemoveWQ::_process(std::pair<boost::intrusive_ptr<PG>, std::shared_ptr<Deletin
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 8: (ThreadPool::WorkQueueVal<std::pair<boost::intrusive_ptr<PG>, std::shared_ptr<Deleti
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 9: (ThreadPool::worker(ThreadPool::WorkThread*)+0xeb8) [0x557c33d459b8]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 10: (ThreadPool::WorkThread::entry()+0x10) [0x557c33d46b50]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 11: (()+0x7494) [0x7ff32abc1494]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 12: (clone()+0x3f) [0x7ff329c48acf]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret t
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: -43> 2018-09-07 16:38:07.700240 7ff32d58ce00 -1 osd.19 1186786 log_to_monitors {defau
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 0> 2018-09-07 19:14:35.020256 7ff2c7fff700 -1 Caught signal (Segmentation fault
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: in thread 7ff2c7fff700 thread_name:tp_osd_disk
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: ceph version 12.2.7-11-gc4753e57fe (c4753e57feb768e9584a7580b67ac072e82a052e) luminous
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 1: (()+0xa38a94) [0x557c33cf6a94]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 2: (()+0x110c0) [0x7ff32abcb0c0]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 3: (BlueStore::_txc_write_nodes(BlueStore::TransContext, std::shared_ptr<KeyValueDB::T
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 4: (BlueStore::queue_transactions(ObjectStore::Sequencer*, std::vector<ObjectStore::Tra
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 5: (ObjectStore::queue_transaction(ObjectStore::Sequencer*, ObjectStore::Transaction&&,
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 6: (remove_dir(CephContext*, ObjectStore*, SnapMapper*, OSDriver*, ObjectStore::Sequenc
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 7: (OSD::RemoveWQ::_process(std::pair<boost::intrusive_ptr<PG>, std::shared_ptr<Deletin
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 8: (ThreadPool::WorkQueueVal<std::pair<boost::intrusive_ptr<PG>, std::shared_ptr<Deleti
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 9: (ThreadPool::worker(ThreadPool::WorkThread*)+0xeb8) [0x557c33d459b8]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 10: (ThreadPool::WorkThread::entry()+0x10) [0x557c33d46b50]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 11: (()+0x7494) [0x7ff32abc1494]
Sep 07 19:14:35 cloud1-1468 ceph-osd[2404]: 12: (clone()+0x3f) [0x7ff329c48acf]

#24 - 09/13/2018 01:39 PM - Igor Fedotov

Thanks, Stefan.
But it looks like call stack is a bit different in you case:
BlueStore::_txc_write_nodes function vs. BlueStore::_wctx_finish()

Also general assert message, like the following would be highly appreciated (it's usually at the very beginning of all that stuff produced on crash):
2018-07-27 21:31:34.164 7f261bea7a00 -1
/home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/centos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/common/bloom_filter.hpp: In function 'void bloom_filter::insert(uint32_t)' thread 7f261bea7a00 time 2018-07-27 21:31:34.163644
/home/jenkins-build/build/workspace/ceph-build/ARCH/x86_64/AVAILABLE_ARCH/x86_64/AVAILABLE_DIST/centos7/DIST/centos7/MACHINE_SIZE/huge/release/13.2.1/rpm/el7/BUILD/ceph-13.2.1/src/common/bloom_filter.hpp: 159: FAILED assert(bit_table_)

It contains actual source file and code line along with the assert message.

#25 - 09/13/2018 01:41 PM - Igor Fedotov

I've just submitted a PR to fix a bug in BlueStore repairer which Troy faced a while ago (This isn't this ticket fix though!):

https://github.com/ceph/ceph/pull/24076

#26 - 09/14/2018 06:12 PM - Stefan Priebe

I'm sorry but no the segfaults do not contain more lines or informations.

i started to monitor all segfaults now. Today i got this one:
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: *** Caught signal (Segmentation fault) ***
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: in thread 7fd82bbff700 thread_name:fn_jrn_objstore
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: ceph version 12.2.7-11-gc4753e57fe (c4753e57feb768e9584a7580b67ac072e82a052e) luminous (stable)
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 1: (()+0xa38a94) [0x56095facaa94]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 2: (()+0x110c0) [0x7fd8582670c0]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 3: (()+0xd6a0) [0x7fd859cff6a0]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 4: (()+0x2973f) [0x7fd859d1b73f]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 5: (malloc()+0x2a6) [0x7fd859cf80d6]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 6: (operator new(unsigned long)+0x18) [0x7fd857b457a8]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 7: (void std::vector<Context, std::allocator<Context*> >::_M_emplace_back_aux<Context*>(Context*&&)+0x56) [0x56095f557756]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 8: (FileStore::_journaled_ahead(FileStore::OpSequencer*, FileStore::Op*, Context*)+0x82b) [0x56095f88e96b]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 9: (Context::complete(int)+0x9) [0x56095f537469]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 10: (Finisher::finisher_thread_entry()+0x4c0) [0x56095fb10d50]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 11: (()+0x7494) [0x7fd85825d494]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 12: (clone()+0x3f) [0x7fd8572e4acf]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 2018-09-14 18:22:17.269375 7fd82bbff700 -1 Caught signal (Segmentation fault) *
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: in thread 7fd82bbff700 thread_name:fn_jrn_objstore
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: ceph version 12.2.7-11-gc4753e57fe (c4753e57feb768e9584a7580b67ac072e82a052e) luminous (stable)
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 1: (()+0xa38a94) [0x56095facaa94]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 2: (()+0x110c0) [0x7fd8582670c0]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 3: (()+0xd6a0) [0x7fd859cff6a0]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 4: (()+0x2973f) [0x7fd859d1b73f]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 5: (malloc()+0x2a6) [0x7fd859cf80d6]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 6: (operator new(unsigned long)+0x18) [0x7fd857b457a8]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 7: (void std::vector<Context*, std::allocator<Context*> >::_M_emplace_back_aux<Context*>(Context*&&)+0x56) [0x56095f557756]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 8: (FileStore::_journaled_ahead(FileStore::OpSequencer*, FileStore::Op*, Context*)+0x82b) [0x56095f88e96b]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 9: (Context::complete(int)+0x9) [0x56095f537469]

Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 10: (Finisher::finisher_thread_entry()+0x4c0) [0x56095fb10d50]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 11: (()+0x7494) [0x7fd85825d494]
Sep 14 18:22:17 cloud2-1345 ceph-osd[3288]: 12: (clone()+0x3f) [0x7fd8572e4acf]

#27 - 11/29/2018 03:11 PM - Sage Weil

- Status changed from New to Can't reproduce

I believe this is related to the SharedBLob refcounting bugs. See 7031addfe6fcd070df8c4c7b175f374bda77a671 and ff8833d0a0d97ae3ce249f10df40efc4195fabc2 for example; both came after 12.2.5

Files

gdb.txt.gz	6.38 KB	07/19/2018	Brad Hubbard