# RADOS - Bug #21092

## OSD sporadically starts reading at 100% of ssd bandwidth

08/24/2017 10:10 AM - Aleksei Gutikov

| | | | |
|---|---|---|---|
| **Status:** | New | **Start date:** | 08/24/2017 |
| **Priority:** | Normal | **Due date:** | |
| **Assignee:** | | **% Done:** | 0% |
| **Category:** | | **Estimated time:** | 0.00 hour |
| **Target version:** | v12.2.0 | **Spent time:** | 0.00 hour |
| **Source:** | | **Affected Versions:** | v12.2.0 |
| **Tags:** | | **ceph-qa-suite:** | rados |
| **Backport:** | | **Component(RADOS):** | OSD |
| **Regression:** | No | **Pull request ID:** | |
| **Severity:** | 2 - major | **Crash signature:** | |
| **Reviewed:** | | | |

**Description**

luminous v12.1.4
bluestore

Periodically (10 mins) some osd starts reading ssd disk at maximum available speed (450-480Mb/sec).
This continuous for 1-3 minutes.
Then after some delay other osd starts same.
Obviously that leads to stuck of pgs on this osd.
The load of other osds not changed during this glitch.
The client io in 'ceph -s' not shows increasing of client traffic.

Stracing of osd at the time of glitch shows over 100 calls pread64 per second with same size and same offset.

```
3393366 09:15:10.679105 pread64(24, "\267\5\0\0\4\0\0\0\361\36\221\203\2662/f\211\305\344\30\253
\324\17\251\310\0328\206\316\t\243"..., 2940928, 96121688064) = 2940928
```

pread64 called with next sizes:
- 4096
- 8192
- 385024
- 1048576
- 2940928

This size and offset (2940928, 96121688064) observed only in combination at same time.
Total 339 calls per second, total 363651072 bytes/sec.
reads of 96121688064 offset 335265792 bytes/sec.

All threads performing those pread64 calls have names "tp_osd_tp".

### History

#### #1 - 08/24/2017 04:13 PM - Aleksei Gutikov

*- File osd-27.debug_bluefs-20.log.gz added*

#### #2 - 08/24/2017 05:23 PM - Aleksei Gutikov

*- File 59.log.gz added*

59.log more obviously shows the issue with repeating part:

```
Aug 24 17:16:29 P20B-SR4-R1-CEPH-DB1 ceph-osd[3559297]: 2017-08-24 17:16:29.322645 7f750d78d700 10 bluefs _rea
d_random h 0x558aa14662d0 0x3f5c993~337c1c from file(ino 2603 size 0x43e7dd9 mtime 2017-08-24 17:13:08.329203
bdev 1 extents [1:0x1671b00000+400000,1:0x1672200000+400000,1:0x1672900000+400000,1:0x1673000000+400000,1:0x16
73600000+400000,1:0x1673d00000+400000,1:0x1674400000+400000,1:0x1674a00000+400000,1:0x1675100000+400000,1:0x16
75800000+400000,1:0x1676000000+400000,1:0x1676600000+400000,1:0x1676d00000+400000,1:0x1677400000+400000,1:0x16
77b00000+400000,1:0x1678200000+200000,1:0x167dc00000+600000])
Aug 24 17:16:29 P20B-SR4-R1-CEPH-DB1 ceph-osd[3559297]: 2017-08-24 17:16:29.322666 7f750d78d700 20 bluefs _rea
d_random read buffered 0x15c993~337c1c of 1:0x167dc00000+600000
Aug 24 17:16:29 P20B-SR4-R1-CEPH-DB1 ceph-osd[3559297]: 2017-08-24 17:16:29.329342 7f750b789700 20 bluefs _rea
d_random got 3374108
```

#### #3 - 08/25/2017 09:20 AM - Aleksei Gutikov

Stacktrace of thread performing reads of 2445312 bytes from offset 96117329920

```
Thread 46 (Thread 0x7f2d851ea700 (LWP 3561173)):
#0  0x00007f2da0d7cd43 in pread64 () at ../sysdeps/unix/syscall-template.S:84
#1  0x0000557bcd058911 in pread64 (__offset=96117329920, __nbytes=2445312, __buf=<optimized out>, __fd=<optimi
zed out>) at /usr/include/x86_64-linux-gnu/bits/unistd.h:99
#2  KernelDevice::direct_read_unaligned (this=this@entry=0x557bd66bcfc0, off=off@entry=96117331901, len=len@en
try=2440717, buf=buf@entry=0x557bf76f6000 "") at /docker/ceph/src/os/bluestore/KernelDevice.cc:724
#3  0x0000557bcd059393 in KernelDevice::read_random (this=0x557bd66bcfc0, off=96117331901, len=2440717, buf=0x
557bf76f6000 "", buffered=<optimized out>) at /docker/ceph/src/os/bluestore/KernelDevice.cc:758
#4  0x0000557bcd02924a in BlueFS::_read_random (this=0x557bd66b6700, h=0x557be5c36480, off=46798781, len=24407
17, out=out@entry=0x557bf76f6000 "") at /docker/ceph/src/os/bluestore/BlueFS.cc:906
#5  0x0000557bcd053760 in BlueFS::read_random (out=0x557bf76f6000 "", len=<optimized out>, offset=<optimized o
ut>, h=<optimized out>, this=<optimized out>) at /docker/ceph/src/os/bluestore/BlueFS.h:423
#6  BlueRocksRandomAccessFile::Read (this=<optimized out>, offset=<optimized out>, n=<optimized out>, result=0
x7f2d851e3a00, scratch=0x557bf76f6000 "") at /docker/ceph/src/os/bluestore/BlueRocksEnv.cc:94
#7  0x0000557bcd4834cf in rocksdb::RandomAccessFileReader::Read(unsigned long, unsigned long, rocksdb::Slice*,
 char*) const ()
#8  0x0000557bcd453d83 in rocksdb::ReadBlockContents(rocksdb::RandomAccessFileReader*, rocksdb::Footer const&,
 rocksdb::ReadOptions const&, rocksdb::BlockHandle const&, rocksdb::BlockContents*, rocksdb::ImmutableCFOption
s const&, bool, rocksdb::Slice const&, rocksdb::PersistentCacheOptions const&) ()
#9  0x0000557bcd444edf in rocksdb::BlockBasedTable::ReadFilter(rocksdb::BlockHandle const&, bool) const ()
#10 0x0000557bcd44547f in rocksdb::BlockBasedTable::GetFilter(rocksdb::BlockHandle const&, bool, bool) const (
)
#11 0x0000557bcd4456cb in rocksdb::BlockBasedTable::GetFilter(bool) const ()
#12 0x0000557bcd44885f in rocksdb::BlockBasedTable::Get(rocksdb::ReadOptions const&, rocksdb::Slice const&, ro
cksdb::GetContext*, bool) ()
#13 0x0000557bcd52cf35 in rocksdb::TableCache::Get(rocksdb::ReadOptions const&, rocksdb::InternalKeyComparator
 const&, rocksdb::FileDescriptor const&, rocksdb::Slice const&, rocksdb::GetContext*, rocksdb::HistogramImpl*,
 bool, int) ()
#14 0x0000557bcd4085cb in rocksdb::Version::Get(rocksdb::ReadOptions const&, rocksdb::LookupKey const&, rocksd
b::PinnableSlice*, rocksdb::Status*, rocksdb::MergeContext*, rocksdb::RangeDelAggregator*, bool*, bool*, unsig
ned long*) ()
#15 0x0000557bcd4d1fce in rocksdb::DBImpl::GetImpl(rocksdb::ReadOptions const&, rocksdb::ColumnFamilyHandle*,
rocksdb::Slice const&, rocksdb::PinnableSlice*, bool*) ()
---Type <return> to continue, or q <return> to quit---
#16 0x0000557bcd4d2432 in rocksdb::DBImpl::Get(rocksdb::ReadOptions const&, rocksdb::ColumnFamilyHandle*, rock
sdb::Slice const&, rocksdb::PinnableSlice*) ()
#17 0x0000557bccfc3561 in rocksdb::DB::Get (value=0x7f2d851e5c30, key=..., column_family=0x557bd6b20f00, optio
ns=..., this=0x557bd6810000) at /docker/ceph/src/rocksdb/include/rocksdb/db.h:289
#18 rocksdb::DB::Get (this=0x557bd6810000, options=..., key=..., value=value@entry=0x7f2d851e5c30) at /docker/
ceph/src/rocksdb/include/rocksdb/db.h:299
#19 0x0000557bccfbacb4 in RocksDBStore::get (this=0x557bd68bb180, prefix=..., key=..., out=0x7f2d851e5fe0) at
/docker/ceph/src/kv/RocksDBStore.cc:768
#20 0x0000557bccefb631 in BlueStore::ExtentMap::<lambda(const string&)>::operator()(const std::__cxx11::string
 &) const (__closure=0x557be4e34910, final_key=...) at /docker/ceph/src/os/bluestore/BlueStore.cc:2646
#21 0x0000557bccefb045 in std::function<void (std::__cxx11::basic_string<char, std::char_traits<char>, std::al
locator<char> > const&)>::operator()(std::__cxx11::basic_string<char, std::char_traits<char>, std::allocator<c
har> > const&) const (__args#0=..., this=0x7f2d851e6060) at /usr/include/c++/5/functional:2267
#22 generate_extent_shard_key_and_apply<std::__cxx11::basic_string<char, std::char_traits<char>, mempool::pool
_allocator<(mempool::pool_index_t)4, char> > >(const std::__cxx11::basic_string<char, std::char_traits<char>,
```

```
mempool::pool_allocator<(mempool::pool_index_t)4, char> > &, uint32_t, std::__cxx11::string *, std::function<void(const std::__cxx11::basic_string<char, std::char_traits<char>, std::allocator<char> >&)>) (onode_key=..., offset=<optimized out>,
    key=0x7f2d851e6040, apply=...) at /docker/ceph/src/os/bluestore/BlueStore.cc:477
#23 0x0000557bccf2b4d8 in BlueStore::ExtentMap::fault_range (this=0x557bed2d1590, db=0x557bd68bb180, offset=offset@entry=0, length=length@entry=4194304) at /docker/ceph/src/os/bluestore/BlueStore.cc:2654
#24 0x0000557bccf5ec88 in BlueStore::_do_truncate (this=this@entry=0x557bd6808000, txc=0x557bd83c6f00, c=..., o=..., offset=offset@entry=0, maybe_unshared_blobs=maybe_unshared_blobs@entry=0x0)
    at /docker/ceph/src/os/bluestore/BlueStore.cc:10417
#25 0x0000557bccf5f6a5 in BlueStore::_do_remove (this=this@entry=0x557bd6808000, txc=txc@entry=0x557bd83c6f00, c=..., o=...) at /docker/ceph/src/os/bluestore/BlueStore.cc:10466
#26 0x0000557bccf60f6b in BlueStore::_remove (this=this@entry=0x557bd6808000, txc=txc@entry=0x557bd83c6f00, c=..., o=...) at /docker/ceph/src/os/bluestore/BlueStore.cc:10561
#27 0x0000557bccf7769c in BlueStore::_txc_add_transaction (this=this@entry=0x557bd6808000, txc=txc@entry=0x557bd83c6f00, t=t@entry=0x557bea54b520) at /docker/ceph/src/os/bluestore/BlueStore.cc:9022
#28 0x0000557bccf7879e in BlueStore::queue_transactions (this=0x557bd6808000, posr=<optimized out>, tls=..., op=..., handle=0x0) at /docker/ceph/src/os/bluestore/BlueStore.cc:8807
#29 0x0000557bccb28a2c in ObjectStore::queue_transactions (handle=0x0, op=..., onreadable_sync=0x0, ondisk=0x0, onreadable=<optimized out>, tls=..., osr=0x557bdd4daf80, this=0x557bd6808000) at /docker/ceph/src/os/ObjectStore.h:1485
#30 ObjectStore::queue_transaction(ObjectStore::Sequencer*, ObjectStore::Transaction&&, Context*, Context*, Context*, boost::intrusive_ptr<TrackedOp>, ThreadPool::TPHandle*) (this=0x557bd6808000, osr=0x557bdd4daf80, t=<optimized out>,
    onreadable=onreadable@entry=0x0, ondisk=ondisk@entry=0x0, onreadable_sync=onreadable_sync@entry=0x0, op=..., handle=0x0) at /docker/ceph/src/os/ObjectStore.h:1473
#31 0x0000557bcccb22e8 in PrimaryLogPG::queue_transaction(ObjectStore::Transaction&&, boost::intrusive_ptr<OpRequest>) (this=<optimized out>, t=<optimized out>, op=...) at /docker/ceph/src/osd/PrimaryLogPG.h:292
#32 0x0000557bccdcc8aa in ReplicatedBackend::_do_pull_response (this=this@entry=0x557bdcc24580, op=...) at /docker/ceph/src/osd/ReplicatedBackend.cc:925
#33 0x0000557bccdd00fc in ReplicatedBackend::do_push (op=..., this=0x557bdcc24580) at /docker/ceph/src/osd/ReplicatedBackend.h:231
#34 ReplicatedBackend::_handle_message (this=0x557bdcc24580, op=...) at /docker/ceph/src/osd/ReplicatedBackend.cc:202
#35 0x0000557bccce06e0 in PGBackend::handle_message (this=<optimized out>, op=...) at /docker/ceph/src/osd/PGBackend.cc:114
#36 0x0000557bccc453cd in PrimaryLogPG::do_request (this=0x557bdac07000, op=..., handle=...) at /docker/ceph/src/osd/PrimaryLogPG.cc:1726
#37 0x0000557bccac9fd9 in OSD::dequeue_op (this=0x557bd6b0e000, pg=..., op=..., handle=...) at /docker/ceph/src/osd/OSD.cc:9518
#38 0x0000557bccd613f7 in PGQueueable::RunVis::operator() (this=this@entry=0x7f2d851e7f00, op=...) at /docker/ceph/src/osd/PGQueueable.cc:22
#39 0x0000557bccaf157e in boost::detail::variant::invoke_visitor<PGQueueable::RunVis>::internal_visit<boost::intrusive_ptr<OpRequest> > (operand=..., this=<synthetic pointer>)
    at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/variant.hpp:1046
#40 boost::detail::variant::visitation_impl_invoke_impl<boost::detail::variant::invoke_visitor<PGQueueable::RunVis>, void*, boost::intrusive_ptr<OpRequest> > (storage=0x7f2d851e7f50, visitor=<synthetic pointer>)
    at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/detail/visitation_impl.hpp:114
#41 boost::detail::variant::visitation_impl_invoke<boost::detail::variant::invoke_visitor<PGQueueable::RunVis>, void*, boost::intrusive_ptr<OpRequest>, boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub, PGRecovery>::has_fallback_type_> (t=0x0, storage=0x7f2d851e7f50, visitor=<synthetic pointer>, internal_which=<optimized out>) at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/detail/visitation_impl.hpp:157
#42 boost::detail::variant::visitation_impl<mpl_::int_<0>, boost::detail::variant::visitation_impl_step<boost::mpl::l_iter<boost::mpl::l_item<mpl_::long_<4l>, boost::intrusive_ptr<OpRequest>, boost::mpl::l_item<mpl_::long_<3l>, PGSnapTrim, boost::mpl::l_item<mpl_::long_<2l>, PGScrub, boost::mpl::l_item<mpl_::long_<1l>, PGRecovery, boost::mpl::l_end> > > > >, boost::mpl::l_iter<boost::mpl::l_end> >, boost::detail::variant::invoke_visitor<PGQueueable::RunVis>, void*, boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub, PGRecovery>::has_fallback_type_> (no_backup_flag=..., storage=0x7f2d851e7f50, visitor=<synthetic pointer>, logical_which=<optimized out>, internal_which=<optimized out>)
    at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/detail/visitation_impl.hpp:238
#43 boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub, PGRecovery>::internal_apply_visitor_impl<boost::detail::variant::invoke_visitor<PGQueueable::RunVis>, void*> (storage=0x7f2d851e7f50, visitor=<synthetic pointer>,
    logical_which=<optimized out>, internal_which=<optimized out>) at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/variant.hpp:2389
#44 boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub, PGRecovery>::internal_apply_visitor<boost::detail::variant::invoke_visitor<PGQueueable::RunVis> > (visitor=<synthetic pointer>, this=0x7f2d851e7f48)
    at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/variant.hpp:2400
#45 boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub, PGRecovery>::apply_visitor<PGQueueable::RunVis> (visitor=..., this=0x7f2d851e7f48) at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/variant.hpp:2423
#46 boost::apply_visitor<PGQueueable::RunVis, boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub, PGRecovery> > (visitable=..., visitor=...)
    at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/detail/apply_visitor_unary.hpp:70
```

```
#47 PGQueueable::run (handle=..., pg=..., osd=<optimized out>, this=0x7f2d851e7f48) at /docker/ceph/src/osd/PG
Queueable.h:140
#48 OSD::ShardedOpWQ::_process (this=0x557bd6b0f560, thread_index=<optimized out>, hb=0x557be10742d0) at /dock
er/ceph/src/osd/OSD.cc:10301
#49 0x0000557bcd0bfe34 in ShardedThreadPool::shardedthreadpool_worker (this=0x557bd6b0eb88, thread_index=7) at
 /docker/ceph/src/common/WorkQueue.cc:339
#50 0x0000557bcd0c2e70 in ShardedThreadPool::WorkThreadSharded::entry (this=<optimized out>) at /docker/ceph/s
rc/common/WorkQueue.h:689
#51 0x00007f2da0d736ba in start_thread (arg=0x7f2d851ea700) at pthread_create.c:333
#52 0x00007f2d9fdea82d in clone () at ../sysdeps/unix/sysv/linux/x86_64/clone.S:109
```

#### #4 - 08/25/2017 10:28 AM - Aleksei Gutikov

Another stack trace that leads to pread same size and same offset:

```
Thread 46 (Thread 0x7f2d851ea700 (LWP 3561173)):
#0  0x00007f2da0d7cd43 in pread64 () at ../sysdeps/unix/syscall-template.S:84
#1  0x0000557bcd058911 in pread64 (__offset=96117329920, __nbytes=2445312, __buf=<optimized out>, __fd=<optimi
zed out>) at /usr/include/x86_64-linux-gnu/bits/unistd.h:99
#2  KernelDevice::direct_read_unaligned (this=this@entry=0x557bd66bcfc0, off=off@entry=96117331901, len=len@en
try=2440717, buf=buf@entry=0x557bf8ab8000 "") at /docker/ceph/src/os/bluestore/KernelDevice.cc:724
#3  0x0000557bcd059393 in KernelDevice::read_random (this=0x557bd66bcfc0, off=96117331901, len=2440717, buf=0x
557bf8ab8000 "", buffered=<optimized out>) at /docker/ceph/src/os/bluestore/KernelDevice.cc:758
#4  0x0000557bcd02924a in BlueFS::_read_random (this=0x557bd66b6700, h=0x557be5c36480, off=46798781, len=24407
17, out=out@entry=0x557bf8ab8000 "") at /docker/ceph/src/os/bluestore/BlueFS.cc:906
#5  0x0000557bcd053760 in BlueFS::read_random (out=0x557bf8ab8000 "", len=<optimized out>, offset=<optimized o
ut>, h=<optimized out>, this=<optimized out>) at /docker/ceph/src/os/bluestore/BlueFS.h:423
#6  BlueRocksRandomAccessFile::Read (this=<optimized out>, offset=<optimized out>, n=<optimized out>, result=0
x7f2d851e2e40, scratch=0x557bf8ab8000 "") at /docker/ceph/src/os/bluestore/BlueRocksEnv.cc:94
#7  0x0000557bcd4834cf in rocksdb::RandomAccessFileReader::Read(unsigned long, unsigned long, rocksdb::Slice*,
 char*) const ()
#8  0x0000557bcd453d83 in rocksdb::ReadBlockContents(rocksdb::RandomAccessFileReader*, rocksdb::Footer const&,
 rocksdb::ReadOptions const&, rocksdb::BlockHandle const&, rocksdb::BlockContents*, rocksdb::ImmutableCFOption
s const&, bool, rocksdb::Slice const&, rocksdb::PersistentCacheOptions const&) ()
#9  0x0000557bcd444edf in rocksdb::BlockBasedTable::ReadFilter(rocksdb::BlockHandle const&, bool) const ()
#10 0x0000557bcd44547f in rocksdb::BlockBasedTable::GetFilter(rocksdb::BlockHandle const&, bool, bool) const (
)
#11 0x0000557bcd4456cb in rocksdb::BlockBasedTable::GetFilter(bool) const ()
#12 0x0000557bcd44885f in rocksdb::BlockBasedTable::Get(rocksdb::ReadOptions const&, rocksdb::Slice const&, ro
cksdb::GetContext*, bool) ()
#13 0x0000557bcd52cf35 in rocksdb::TableCache::Get(rocksdb::ReadOptions const&, rocksdb::InternalKeyComparator
 const&, rocksdb::FileDescriptor const&, rocksdb::Slice const&, rocksdb::GetContext*, rocksdb::HistogramImpl*,
 bool, int) ()
#14 0x0000557bcd4085cb in rocksdb::Version::Get(rocksdb::ReadOptions const&, rocksdb::LookupKey const&, rocksd
b::PinnableSlice*, rocksdb::Status*, rocksdb::MergeContext*, rocksdb::RangeDelAggregator*, bool*, bool*, unsig
ned long*) ()
#15 0x0000557bcd4d1fce in rocksdb::DBImpl::GetImpl(rocksdb::ReadOptions const&, rocksdb::ColumnFamilyHandle*,
 rocksdb::Slice const&, rocksdb::PinnableSlice*, bool*) ()
#16 0x0000557bcd4d2432 in rocksdb::DBImpl::Get(rocksdb::ReadOptions const&, rocksdb::ColumnFamilyHandle*, rock
sdb::Slice const&, rocksdb::PinnableSlice*) ()
#17 0x0000557bccfc3561 in rocksdb::DB::Get (value=0x7f2d851e5070, key=..., column_family=0x557bd6b20f00, optio
ns=..., this=0x557bd6810000) at /docker/ceph/src/rocksdb/include/rocksdb/db.h:289
#18 rocksdb::DB::Get (this=0x557bd6810000, options=..., key=..., value=value@entry=0x7f2d851e5070) at /docker/
ceph/src/rocksdb/include/rocksdb/db.h:299
#19 0x0000557bccfbacb4 in RocksDBStore::get (this=0x557bd68bb180, prefix=..., key=..., out=0x7f2d851e5420) at
 /docker/ceph/src/kv/RocksDBStore.cc:768
#20 0x0000557bccefb631 in BlueStore::ExtentMap::<lambda(const string&)>::operator()(const std::__cxx11::string
 &) const (__closure=0x557be12f3990, final_key=...) at /docker/ceph/src/os/bluestore/BlueStore.cc:2646
#21 0x0000557bccefb045 in std::function<void (std::__cxx11::basic_string<char, std::char_traits<char>, std::al
locator<char> > const&)>::operator()(std::__cxx11::basic_string<char, std::char_traits<char>, std::allocator<c
har> > const&) const (__args#0=..., this=0x7f2d851e54a0) at /usr/include/c++/5/functional:2267
#22 generate_extent_shard_key_and_apply<std::__cxx11::basic_string<char, std::char_traits<char>, mempool::pool
```

```
_allocator<(mempool::pool_index_t)4, char> > >(const std::__cxx11::basic_string<char, std::char_traits<char>,
mempool::pool_allocator<(mempool::pool_index_t)4, char> > &, uint32_t, std::__cxx11::string *, std::function<v
oid(const std::__cxx11::basic_string<char, std::char_traits<char>, std::allocator<char> >&)>) (onode_key=...,
offset=<optimized out>,
    key=0x7f2d851e5480, apply=...) at /docker/ceph/src/os/bluestore/BlueStore.cc:477
#23 0x0000557bccf2b4d8 in BlueStore::ExtentMap::fault_range (this=this@entry=0x557beb8f3850, db=0x557bd68bb180
, offset=offset@entry=1687552, length=length@entry=131072) at /docker/ceph/src/os/bluestore/BlueStore.cc:2654
#24 0x0000557bccf67df9 in BlueStore::_do_read (this=this@entry=0x557bd6808000, c=c@entry=0x557bd6b41a00, o=...
, offset=offset@entry=1687552, length=length@entry=131072, bl=..., op_flags=0)
    at /docker/ceph/src/os/bluestore/BlueStore.cc:6386
#25 0x0000557bccf6b3a7 in BlueStore::read (this=0x557bd6808000, c_=..., oid=..., offset=1687552, length=131072
, bl=..., op_flags=0) at /docker/ceph/src/os/bluestore/BlueStore.cc:6299
#26 0x0000557bccdbab95 in ReplicatedBackend::objects_read_sync (this=<optimized out>, hoid=..., off=1687552, l
en=131072, op_flags=0, bl=0x557be2af8270) at /docker/ceph/src/osd/ReplicatedBackend.cc:279
#27 0x0000557bccc2a4d1 in PrimaryLogPG::do_read (this=this@entry=0x557bdac07000, ctx=0x557bf0a29000, osd_op=..
.) at /docker/ceph/src/osd/PrimaryLogPG.cc:4827
#28 0x0000557bccc79e99 in PrimaryLogPG::do_osd_ops (this=this@entry=0x557bdac07000, ctx=ctx@entry=0x557bf0a290
00, ops=...) at /docker/ceph/src/osd/PrimaryLogPG.cc:5116
#29 0x0000557bccc8692f in PrimaryLogPG::prepare_transaction (this=this@entry=0x557bdac07000, ctx=ctx@entry=0x5
57bf0a29000) at /docker/ceph/src/osd/PrimaryLogPG.cc:7437
#30 0x0000557bccc871ab in PrimaryLogPG::execute_ctx (this=this@entry=0x557bdac07000, ctx=ctx@entry=0x557bf0a29
000) at /docker/ceph/src/osd/PrimaryLogPG.cc:3227
#31 0x0000557bccc8bc8b in PrimaryLogPG::do_op (this=0x557bdac07000, op=...) at /docker/ceph/src/osd/PrimaryLog
PG.cc:2327
#32 0x0000557bccc45d03 in PrimaryLogPG::do_request (this=0x557bdac07000, op=..., handle=...) at /docker/ceph/s
rc/osd/PrimaryLogPG.cc:1746
#33 0x0000557bccac9fd9 in OSD::dequeue_op (this=0x557bd6b0e000, pg=..., op=..., handle=...) at /docker/ceph/sr
c/osd/OSD.cc:9518
#34 0x0000557bccd613f7 in PGQueueable::RunVis::operator() (this=this@entry=0x7f2d851e7f00, op=...) at /docker/
ceph/src/osd/PGQueueable.cc:22
#35 0x0000557bccaf157e in boost::detail::variant::invoke_visitor<PGQueueable::RunVis>::internal_visit<boost::i
ntrusive_ptr<OpRequest> > (operand=..., this=<synthetic pointer>)
    at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/variant.hpp:1046
#36 boost::detail::variant::visitation_impl_invoke_impl<boost::detail::variant::invoke_visitor<PGQueueable::Ru
nVis>, void*, boost::intrusive_ptr<OpRequest> > (storage=0x7f2d851e7f50, visitor=<synthetic pointer>)
    at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/detail/visitation_impl.hpp:114
---Type <return> to continue, or q <return> to quit---
#37 boost::detail::variant::visitation_impl_invoke<boost::detail::variant::invoke_visitor<PGQueueable::RunVis>
, void*, boost::intrusive_ptr<OpRequest>, boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub,
 PGRecovery>::has_fallback_type_> (t=0x0, storage=0x7f2d851e7f50, visitor=<synthetic pointer>, internal_which=
<optimized out>) at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/detail/visitation_impl.hpp:1
57
#38 boost::detail::variant::visitation_impl<mpl_::int_<0>, boost::detail::variant::visitation_impl_step<boost:
:mpl::l_iter<boost::mpl::l_item<mpl_::long_<4l>, boost::intrusive_ptr<OpRequest>, boost::mpl::l_item<mpl_::lon
g_<3l>, PGSnapTrim, boost::mpl::l_item<mpl_::long_<2l>, PGScrub, boost::mpl::l_item<mpl_::long_<1l>, PGRecover
y, boost::mpl::l_end> > > > >, boost::mpl::l_iter<boost::mpl::l_end> >, boost::detail::variant::invoke_visitor
<PGQueueable::RunVis>, void*, boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub, PGRecovery>
::has_fallback_type_> (no_backup_flag=..., storage=0x7f2d851e7f50, visitor=<synthetic pointer>, logical_which=
<optimized out>, internal_which=<optimized out>)
    at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/detail/visitation_impl.hpp:238
#39 boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub, PGRecovery>::internal_apply_visitor_i
mpl<boost::detail::variant::invoke_visitor<PGQueueable::RunVis>, void*> (storage=0x7f2d851e7f50, visitor=<synt
hetic pointer>,
    logical_which=<optimized out>, internal_which=<optimized out>) at /docker/ceph/obj-x86_64-linux-gnu/boost/
include/boost/variant/variant.hpp:2389
#40 boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub, PGRecovery>::internal_apply_visitor<b
oost::detail::variant::invoke_visitor<PGQueueable::RunVis> > (visitor=<synthetic pointer>, this=0x7f2d851e7f48
)
    at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/variant.hpp:2400
#41 boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGScrub, PGRecovery>::apply_visitor<PGQueueabl
e::RunVis> (visitor=..., this=0x7f2d851e7f48) at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant
/variant.hpp:2423
#42 boost::apply_visitor<PGQueueable::RunVis, boost::variant<boost::intrusive_ptr<OpRequest>, PGSnapTrim, PGSc
rub, PGRecovery> > (visitable=..., visitor=...)
    at /docker/ceph/obj-x86_64-linux-gnu/boost/include/boost/variant/detail/apply_visitor_unary.hpp:70
#43 PGQueueable::run (handle=..., pg=..., osd=<optimized out>, this=0x7f2d851e7f48) at /docker/ceph/src/osd/PG
Queueable.h:140
#44 OSD::ShardedOpWQ::_process (this=0x557bd6b0f560, thread_index=<optimized out>, hb=0x557be10742d0) at /dock
er/ceph/src/osd/OSD.cc:10301
#45 0x0000557bcd0bfe34 in ShardedThreadPool::shardedthreadpool_worker (this=0x557bd6b0eb88, thread_index=7) at
 /docker/ceph/src/common/WorkQueue.cc:339
#46 0x0000557bcd0c2e70 in ShardedThreadPool::WorkThreadSharded::entry (this=<optimized out>) at /docker/ceph/s
rc/common/WorkQueue.h:689
#47 0x00007f2da0d736ba in start_thread (arg=0x7f2d851ea700) at pthread_create.c:333
#48 0x00007f2d9fdea82d in clone () at ../sysdeps/unix/sysv/linux/x86_64/clone.S:109
```

Both operations similar from level of BlueStore::ExtentMap::fault_range
Is it possible that range size=2440717 containing bluestore metadata laying in offset=96117331901
and for some reason this range was not cached or was displaced and it cause read of this range on every osd op?

#### #5 - 08/29/2017 03:08 PM - Aleksei Gutikov

Seems that is side effect of too small value for bluestore_cache_size.
We set it to 50M to reduce osd memory consumption and reduce frequency of OOM kills of osds.
Possibly that leads to this behaviour if bluefs metadata not suits into bluestore cache.
There were no such effects at least till 12.1.1

### Files

| | | | |
|---|---|---|---|
| osd-36.starce.1.txt.gz | 616 KB | 08/24/2017 | Aleksei Gutikov |
| osd-27.debug_bluefs-20.log.gz | 153 KB | 08/24/2017 | Aleksei Gutikov |
| 59.log.gz | 161 KB | 08/24/2017 | Aleksei Gutikov |