Ceph - Bug #15298

rocksdb corrupt with bluefs

03/29/2016 07:22 AM - Xinxin Shu

Status:	Can't reproduce	**% Done:**	0%
Priority:	High	**Spent time:**	0.00 hour
Assignee:			
Category:			
Target version:			
Source:	other	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

2016-03-29 09:09:18.236160 7f821c7ff700 4 rocksdb: (Original Log Time 2016/03/29-09:09:18.236050) EVENT_LOG_v1 {"time_micros": 1459213758236023, "job": 11158, "event": "compaction_finished", "compaction_time_micros": 3262213, "output_level": 1, "num_output_files": 8, "total_output_size": 12833518, "num_input_records": 79130, "num_output_records": 68472, "num_subcompactions": 1, "lsm_state": [0, 8, 82, 520, 367, 0, 0]}
2016-03-29 09:09:18.236439 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236435, "job": 11158, "event": "table_file_deletion", "file_number": 72280}
2016-03-29 09:09:18.236464 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236462, "job": 11158, "event": "table_file_deletion", "file_number": 72025}
2016-03-29 09:09:18.236489 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236487, "job": 11158, "event": "table_file_deletion", "file_number": 72008}
2016-03-29 09:09:18.236506 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236504, "job": 11158, "event": "table_file_deletion", "file_number": 72006}
2016-03-29 09:09:18.236521 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236519, "job": 11158, "event": "table_file_deletion", "file_number": 72005}
2016-03-29 09:09:18.236536 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236535, "job": 11158, "event": "table_file_deletion", "file_number": 72004}
2016-03-29 09:09:18.236568 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236566, "job": 11158, "event": "table_file_deletion", "file_number": 72002}
2016-03-29 09:09:18.236584 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236583, "job": 11158, "event": "table_file_deletion", "file_number": 72001}
2016-03-29 09:09:18.236600 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236599, "job": 11158, "event": "table_file_deletion", "file_number": 71998}
2016-03-29 09:09:18.236617 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236615, "job": 11158, "event": "table_file_deletion", "file_number": 71996}
2016-03-29 09:09:18.236631 7f821c7ff700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1459213758236629, "job": 11158, "event": "table_file_deletion", "file_number": 71995}
2016-03-29 09:09:18.236716 7f821c7ff700 2 rocksdb: Waiting after background compaction error: NotFound: , Accumulated background error counts: 1

History

#1 - 03/29/2016 05:46 PM - Sage Weil

- Status changed from New to Need More Info

What version? Do you have a complete log (debug bluefs = 20)?

#2 - 03/30/2016 09:04 AM - Xinxin Shu

SHA1 is eeaab88e31d84c612fe16374c9b84e2cbd5072d6, the full log is several GB, i cannot upload, but it seems error about recycle log, i paste snippet

2016-03-30 10:57:24.375383 7fd111bff700 10 bluefs readdir db.wal
2016-03-30 10:57:24.375390 7fd111bff700 4 rocksdb: adding log 1071 to recycle list

2016-03-30 10:57:24.375392 7fd111bff700 4 rocksdb: adding log 1073 to recycle list

2016-03-30 10:57:24.375393 7fd111bff700 4 rocksdb: adding log 1074 to recycle list

2016-03-30 10:57:24.375400 7fd111bff700 4 rocksdb: (Original Log Time 2016/03/30-10:57:24.374758) [default] Level-0 commit table #1076 started
2016-03-30 10:57:24.375403 7fd111bff700 4 rocksdb: (Original Log Time 2016/03/30-10:57:24.375278) [default] Level-0 commit table #1076: memtable #1 done
2016-03-30 10:57:24.375406 7fd111bff700 4 rocksdb: (Original Log Time 2016/03/30-10:57:24.375279) [default] Level-0 commit table #1076: memtable #2 done
2016-03-30 10:57:24.375419 7fd111bff700 4 rocksdb: (Original Log Time 2016/03/30-10:57:24.375280) [default] Level-0 commit table #1076: memtable #3 done
2016-03-30 10:57:24.375421 7fd111bff700 4 rocksdb: (Original Log Time 2016/03/30-10:57:24.375287) EVENT_LOG_v1 {"time_micros": 1459306644375283, "job": 206, "event": "flush_finished", "lsm_state": [4, 6, 49, 0, 0, 0, 0]}
2016-03-30 10:57:24.375423 7fd111bff700 4 rocksdb: (Original Log Time 2016/03/30-10:57:24.375344) [default] Level summary: base level 1 max bytes base 10485760 files[4 6 49 0 0 0 0] max score 1.00

2016-03-30 10:57:24.375463 7fd111bff700 10 bluefs unlink db.wal/001074.log
2016-03-30 10:57:24.375468 7fd111bff700 20 bluefs _drop_link had refs 1 on file(ino 16 size 3875930 mtime 2016-03-30 10:57:24.114242 bdev 1 extents [1:5242880+3145728,1:12582912+1048576])
2016-03-30 10:57:24.375475 7fd111bff700 20 bluefs _drop_link destroying file(ino 16 size 3875930 mtime 2016-03-30 10:57:24.114242 bdev 1 extents [1:5242880+3145728,1:12582912+1048576])
2016-03-30 10:57:24.375495 7fd111bff700 10 bluefs unlink db.wal/001073.log
2016-03-30 10:57:24.375497 7fd111bff700 20 bluefs _drop_link had refs 1 on file(ino 20 size 3868510 mtime 2016-03-29 17:10:09.343188 bdev 1 extents [1:22020096+4194304])
2016-03-30 10:57:24.375504 7fd111bff700 20 bluefs _drop_link destroying file(ino 20 size 3868510 mtime 2016-03-29 17:10:09.343188 bdev 1 extents [1:22020096+4194304])
2016-03-30 10:57:24.375509 7fd111bff700 10 bluefs unlink db.wal/001071.log
2016-03-30 10:57:24.375506 7fd0e53ff700 3 rocksdb: ------- DUMPING STATS -------
2016-03-30 10:57:24.375511 7fd111bff700 20 bluefs _drop_link had refs 1 on file(ino 19 size 3868624 mtime 2016-03-29 17:10:01.185813 bdev 1 extents [1:17825792+4194304])
2016-03-30 10:57:24.375517 7fd111bff700 20 bluefs _drop_link destroying file(ino 19 size 3868624 mtime 2016-03-29 17:10:01.185813 bdev 1 extents [1:17825792+4194304])

in the above log, it seems than 1071 was added to recycle log, but the 001071.log file is deleted by rocksdb, then after that we reuse this log file, but it cannot be renamed since this log file has been deleted

2016-03-30 10:57:25.357404 7fd0e53ff700 20 bluefs _read_random read buffered 444585~4001 of 0:2086666240+2097152
2016-03-30 10:57:25.357405 7fd0e53ff700 5 bdev(/var/lib/ceph/mnt/osd-device-0-data/block) read_buffered 2087110825~4001
2016-03-30 10:57:25.357406 7fd1a07fa700 5 bdev(/var/lib/ceph/mnt/osd-device-0-data/block) flush in 0.000511
2016-03-30 10:57:25.357431 7fd1a07fa700 10 bluefs _flush 0x7fd059dcfd00 ignoring, length 3213 < min_flush_size 65536
2016-03-30 10:57:25.357463 7fd1a07fa700 10 bluefs _flush 0x7fd059dcfd00 ignoring, length 6405 < min_flush_size 65536
2016-03-30 10:57:25.357485 7fd1a07fa700 4 rocksdb: reusing log 1071 from recycle list

2016-03-30 10:57:25.357492 7fd1a07fa700 10 bluefs rename db.wal/001071.log -> db.wal/001079.log
2016-03-30 10:57:25.357494 7fd1a07fa700 20 bluefs rename dir db.wal (0x7fd1ae26fc90) file 001071.log not found

#3 - 04/15/2016 06:10 AM - Evgeniy Firsov

I hit the same error message:
"2016-04-14 22:42:17.904821 7ff84ebff700 2 rocksdb: Waiting after background compaction error: NotFound: , Accumulated background error counts: 1"
It happens every time 6 hours of 4k fio random write run finishes and another identical one starts.

Single OSD. Min alloc size = 4K
Revision: 99c6f30

#load
fio/fio --ioengine=rbd --clientname=admin --pool=rbd --rbdname=test --thread --output=out.fio.txt --direct=1 --rw=write -refill_buffers --randrepeat=0 --bs=256K --rwmixread=0 -iodepth=64 --group_reporting --name=4ktest --fill_device=1 --filesize=6T

#warmup
fio/fio --ioengine=rbd --clientname=admin --pool=rbd --rbdname=test --thread --output=out.fio.txt --direct=1 --rw=randrw -refill_buffers --randrepeat=0 --bs=4k --rwmixread=0 -iodepth=$IODEPTH --group_reporting --name=4ktest --thread --numjobs=$THREAD --runtime=21600 --filesize=6T --time_based --status-interval=10

<--- Here error appears in the log every time

#run
fio/fio --ioengine=rbd --clientname=admin --pool=rbd --rbdname=test --thread --output=out.fio.txt --direct=1 --rw=randrw -refill_buffers --randrepeat=0 --bs=4k --rwmixread=0 -iodepth=64 --group_reporting --name=4ktest --thread --numjobs=32 --runtime=900 --filesize=6T --time_based --status-interval=10

#4 - 04/18/2016 03:23 PM - Haodong Tang

When I use fio to test overlay write, I met the same error,

##Key configuration:
bluestore_overlay_max = 512
bluestore_overlay_max_length = 65536
bluestore block db create = true
bluestore block wal create = true
bluefs = true or false

##use fio
[global]
direct=1
time_based
[rand_write_test]
rw=randwrite
bs=4k
iodepth=64
ramp_time=100
runtime=400
size=10g
filename=/dev/vdb
ioengine=libaio
iodepth_batch_submit=1
iodepth_batch_complete=1
norandommap

randrepeat=0

##debug
2016-04-18 21:52:41.108687 7fb37b4ee700 2 rocksdb: Waiting after background compaction error: Corruption: block checksum mismatch, Accumulated background error counts: 1
2016-04-18 21:52:42.108928 7fb37b4ee700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1460987562108920, "job": 5, "event": "table_file_deletion", "file_number": 20}
2016-04-18 22:15:46.394469 7fb391cec700 -1 ** **Caught signal (Aborted) ***
in thread 7fb391cec700 thread_name:tp_osd_tp
2016-04-18 21:52:41.080552 7fb37b4ee700 4 rocksdb: [default] [JOB 5] Compacting 4@0 files to L1, score 1.00
2016-04-18 21:52:41.080563 7fb37b4ee700 4 rocksdb: [default] Compaction start summary: Base version 4 Base level 0, inputs: [19(2210KB) 15(3090KB) 7(1223B) 4(644B)]

2016-04-18 21:52:41.108647 7fb37b4ee700 3 rocksdb: Compaction error: Corruption: block checksum mismatch

2016-04-18 21:52:41.108687 7fb37b4ee700 2 rocksdb: Waiting after background compaction error: Corruption: block checksum mismatch, Accumulated background error counts: 1
2016-04-18 21:52:42.108928 7fb37b4ee700 4 rocksdb: EVENT_LOG_v1 {"time_micros": 1460987562108920, "job": 5, "event": "table_file_deletion", "file_number": 20}
2016-04-18 22:15:46.394469 7fb391cec700 ~~1 ** Caught signal (Aborted) *~~
~~in thread 7fb391cec700 thread_name:tp_osd_tp~~
– begin dump of recent events ---
~~9980> 2016-04-18 22:15:44.751990 7fb383365700 5 -~~ op tracker -- seq: 1330338, time: 2016-04-18 22:15:44.751989, event: queued_for_pg, op: osd_repop(client.4589.0:327229 1.11c)
~~9979> 2016-04-18 22:15:44.752018 7fb392cec700 5 -~~ op tracker -- seq: 1330338, time: 2016-04-18 22:15:44.752017, event: reached_pg, op: osd_repop(client.4589.0:327229 1.11c)
20: (clone()+0x6d) [0x7fb3aed2647d]
NOTE: a copy of the executable, or `objdump -rdS <executable>` is needed to interpret this.

#5 - 04/19/2016 05:58 PM - Evgeniy Firsov

If it helps, for me it usually neighboring with freelist assert:
-2> 2016-04-19 05:32:14.837394 7f7978ffc700 2 rocksdb: Waiting after background compaction error: NotFound: , Accumulated background error counts: 1
-1> 2016-04-19 05:48:37.737309 7f79693fa700 -1 freelist release bad release 6903857618944~4096 overlaps with 6903857618944~4096
0> 2016-04-19 05:48:38.668714 7f79693fa700 -1 os/bluestore/FreelistManager.cc: In function 'int FreelistManager::release(uint64_t, uint64_t, KeyValueDB::T
ransaction)' thread 7f79693fa700 time 2016-04-19 05:48:37.810680
os/bluestore/FreelistManager.cc: 237: FAILED assert(0 == "bad release overlap")

```
ceph version 10.1.0-284-g9a7b051 (9a7b051eefc7eeb8d4ffe8cbb32c375dcd3f1981)
 1: (ceph::__ceph_assert_fail(char const*, char const*, int, char const*)+0x8b) [0x7f7988d556db]
 2
```

#6 - 05/04/2016 01:39 AM - Xinxin Shu

- Status changed from Need More Info to New

#7 - 05/17/2016 09:25 PM - Sage Weil

The overlay code is broken--do not use it.

I think we need to rebase rocksdb on current master and retest. I think Somnath offered to do this?

#8 - 05/19/2016 08:04 PM - Evgeniy Firsov

I just tried with latest RocksDB master and with bluefs disabled, still hit the problem.

How can I disable overlay? Is it enabled by default?

#9 - 05/19/2016 08:10 PM - Evgeniy Firsov

Sorry, I hit "bad release overlap", not the original ticket problem.

#10 - 04/19/2017 03:21 PM - Sage Weil

- Status changed from New to Can't reproduce

#11 - 04/19/2017 03:21 PM - Sage Weil

overlay is long gone. pls open a new ticket if you see problems with kraken or master!