bluestore - Bug #54547

Deferred writes might cause "rocksdb: Corruption: Bad table magic number"

03/14/2022 12:24 PM - Igor Fedotov

Status:	Resolved	**% Done:**	0%
Priority:	Urgent		
Assignee:	Adam Kupczyk		
Category:			
Target version:			
Source:		**Affected Versions:**	v16.2.7
Tags:	backport_processed	**ceph-qa-suite:**	
Backport:	pacific, quincy	**Pull request ID:**	46890
Regression:	No	**Crash signature (v1):**	
Severity:	3 - minor	**Crash signature (v2):**	
Reviewed:			

Description

Looks like under some circumstances deferred write op might persist in DB longer then allocated extents it's targeted to.
Hence this could cause data corruption if those extents are reallocated and overwritten by e.g. RocksDB/BlueFS on OSD startup.
Following deferred write commit would invalidate new data in that case...

Related issues:

Duplicated by bluestore - Bug #54409: OSD fails to start up with "rocksdb: Co...	**Duplicate**
Copied to bluestore - Backport #56668: quincy: Deferred writes might cause "r...	**Resolved**
Copied to bluestore - Backport #56669: pacific: Deferred writes might cause "...	**Resolved**

History

#1 - 03/14/2022 12:55 PM - Igor Fedotov

- File squeezed_log_for_ticket.txt added

#2 - 03/14/2022 12:56 PM - Igor Fedotov

- Duplicates Bug #54409: OSD fails to start up with "rocksdb: Corruption: Bad table magic number" error added

#3 - 03/14/2022 12:57 PM - Igor Fedotov

- Status changed from New to Duplicate

#4 - 03/14/2022 12:57 PM - Igor Fedotov

- Duplicates deleted (Bug #54409: OSD fails to start up with "rocksdb: Corruption: Bad table magic number" error)

#5 - 03/14/2022 12:58 PM - Igor Fedotov

- Status changed from Duplicate to New

#6 - 03/14/2022 12:58 PM - Igor Fedotov

- Status changed from New to Triaged

#7 - 03/14/2022 12:59 PM - Igor Fedotov

- Subject changed from Deferred writes might cause data corruption to Deferred writes might cause "rocksdb: Corruption: Bad table magic number"

#8 - 03/14/2022 12:59 PM - Igor Fedotov

- Duplicated by Bug #54409: OSD fails to start up with "rocksdb: Corruption: Bad table magic number" error added

#9 - 03/14/2022 04:18 PM - Igor Fedotov

I managed to reproduce (to a major degree) the bug with vstart-ed cluster:
- osd_fast_shutdown = true
- rbd above ec 2+1 pool
- dd if=/dev/random of=/dev/rbd0 count=5 bs=4096
dd if=/dev/random of=/dev/rbd0 count=5 bs=4096
dd if=/dev/random of=/dev/rbd count=5 bs=4096
kill <ceph-osd-pid>

as a result on a subsequent ceph-osd start one can see the following output (with debug-bluestore = 20)
2022-03-14T18:45:58.127+0300 7fc1cf4e1e40 10 freelist enumerate_next 0x42000~3000
2022-03-14T18:45:58.127+0300 7fc1cf4e1e40 10 AvlAllocator init_add_free offset 0x42000 length 0x3000
2022-03-14T18:45:58.127+0300 7fc1cf4e1e40 10 freelist enumerate_next 0x51000~6000
2022-03-14T18:45:58.127+0300 7fc1cf4e1e40 10 AvlAllocator init_add_free offset 0x51000 length 0x6000
2022-03-14T18:45:58.127+0300 7fc1cf4e1e40 10 freelist enumerate_next 0x5a000~18fffa6000
2022-03-14T18:45:58.127+0300 7fc1cf4e1e40 10 AvlAllocator init_add_free offset 0x5a000 length 0x18fffa6000
2022-03-14T18:45:58.127+0300 7fc1cf4e1e40 10 freelist enumerate_next end

022-03-14T18:45:58.663+0300 7fc1cf4e1e40 20 bluestore(/home/if/pacific/build/dev/osd0) _deferred_submit_unlock seq 4 0x51000~3000
2022-03-14T18:45:58.663+0300 7fc1cf4e1e40 20 bluestore(/home/if/pacific/build/dev/osd0) _deferred_submit_unlock seq 5 0x54000~3000
2022-03-14T18:45:58.663+0300 7fc1cf4e1e40 20 bluestore(/home/if/pacific/build/dev/osd0) _deferred_submit_unlock seq 6 0x57000~3000

Please note extents 0x51000~3000 and 0x54000~3000 are marked as free in freelist while still being accessed by deferred write procedure.
If bluefs allocates these extents on startup - original DB corruption would occur. Which is rather a seldom case though...

#10 - 03/14/2022 04:56 PM - Igor Fedotov

a couple addendums to the previous comment:
- vstart cluster above should use spinning drives or benefit from setting bluestore debug enforce settings = hdd to get deferred writes
- it's pacific release which I've been using. Quincy one generally suffers from the same but NCB stuff might hide the allocation conflicts

#11 - 06/28/2022 06:57 AM - Adam Kupczyk

https://github.com/ceph/ceph/pull/46856 is a consistent replicator for deferred writes corrupting RocksDB.

#12 - 06/29/2022 08:44 AM - Adam Kupczyk

- Backport set to octopus, pacific, quincy

- Pull request ID set to 46890

#13 - 07/21/2022 10:48 PM - Yuri Weinstein

https://github.com/ceph/ceph/pull/46890 merged

#14 - 07/21/2022 10:54 PM - Neha Ojha

- Status changed from Triaged to Pending Backport

- Backport changed from octopus, pacific, quincy to pacific, quincy

#15 - 07/21/2022 10:55 PM - Backport Bot

- Copied to Backport #56668: quincy: Deferred writes might cause "rocksdb: Corruption: Bad table magic number" added

#16 - 07/21/2022 10:55 PM - Backport Bot

- Copied to Backport #56669: pacific: Deferred writes might cause "rocksdb: Corruption: Bad table magic number" added

#17 - 07/28/2022 02:51 PM - Igor Fedotov

- Assignee set to Adam Kupczyk

#18 - 08/08/2022 04:28 PM - Backport Bot

- Tags set to backport_processed

#19 - 08/18/2022 03:14 PM - Igor Fedotov

- Status changed from Pending Backport to Resolved

Files

squeezed_log_for_ticket.txt	9.86 KB	03/14/2022	Igor Fedotov