Linux kernel client - Bug #42757

deadlock on lock_rwsem: rbd_quiesce_lock() vs watch errors

11/11/2019 09:21 PM - Ilya Dryomov

Status:	Resolved	**% Done:**	0%
Priority:	Urgent	**Spent time:**	0.00 hour
Assignee:	Ilya Dryomov		
Category:	rbd		
Target version:			
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Crash signature (v1):**	
Severity:	3 - minor	**Crash signature (v2):**	

Description

http://pulpito.front.sepia.ceph.com/dis-2019-11-09_22:31:53-krbd-master-distro-basic-smithi/4492205

```
INFO: task kworker/2:1:73 blocked for more than 120 seconds.
     Not tainted 4.18.0-ceph-g2c9e340c576d #2
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
kworker/2:1     D    0    73      2 0x80004000
Workqueue: ceph-msgr ceph_con_workfn [libceph]
Call Trace:
? __schedule+0x2a9/0x650
schedule+0x39/0xa0
rwsem_down_read_slowpath+0x20d/0x490
rbd_img_handle_request+0x33/0x250 [rbd]
rbd_obj_handle_request+0x2c/0x40 [rbd]
__complete_request+0x26/0x80 [libceph]
dispatch+0x354/0xbc0 [libceph]
? read_partial_message+0x229/0x810 [libceph]
? ceph_tcp_recvmsg+0x6c/0xa0 [libceph]
try_read+0x8b9/0x11c0 [libceph]
? finish_task_switch+0x7d/0x2b0
ceph_con_workfn+0xd1/0x600 [libceph]
process_one_work+0x171/0x380
worker_thread+0x49/0x3f0
kthread+0xf8/0x130
? max_active_store+0x80/0x80
? kthread_bind+0x10/0x10
ret_from_fork+0x35/0x40

INFO: task kworker/u16:6:3030 blocked for more than 120 seconds.
     Not tainted 4.18.0-ceph-g2c9e340c576d #2
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
kworker/u16:6   D    0  3030      2 0x80004000
Workqueue: rbd0-tasks rbd_reregister_watch [rbd]
Call Trace:
? __schedule+0x2a9/0x650
schedule+0x39/0xa0
schedule_timeout+0x1c8/0x290
wait_for_completion+0x123/0x190
? wake_up_q+0x70/0x70
rbd_quiesce_lock+0x9a/0x120 [rbd]
rbd_reregister_watch+0x183/0x240 [rbd]
process_one_work+0x171/0x380
worker_thread+0x49/0x3f0
kthread+0xf8/0x130
? max_active_store+0x80/0x80
```

```
? kthread_bind+0x10/0x10
ret_from_fork+0x35/0x40

INFO: task kworker/u16:9:3583 blocked for more than 120 seconds.
     Not tainted 4.18.0-ceph-g2c9e340c576d #2
"echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
kworker/u16:9   D    0  3583      2 0x80004000
Workqueue: ceph-watch-notify do_watch_error [libceph]
Call Trace:
? __schedule+0x2a9/0x650
? wake_up_klogd+0x30/0x40
schedule+0x39/0xa0
rwsem_down_write_slowpath+0x2b8/0x478
? __switch_to_asm+0x41/0x70
rbd_watch_errcb+0x27/0xa0 [rbd]
do_watch_error+0x7e/0xf0 [libceph]
process_one_work+0x171/0x380
worker_thread+0x49/0x3f0
kthread+0xf8/0x130
? max_active_store+0x80/0x80
? kthread_bind+0x10/0x10
ret_from_fork+0x35/0x40
```

rbd_quiesce_lock() is holding lock_rwsem for read, waiting for in-flight image requests to complete. In order to complete, each image request needs to take lock_rwsem, also for read. If a writer sneaks in onto the semaphore queue, we get a deadlock because no further readers are granted the lock. Here that writer is rbd_watch_errcb(), attempting to grab the semaphore for write to update owner_cid.

Related issues:

Duplicated by Linux kernel client - Bug #51136: Random hanging issues with rb...	**Duplicate**

History

#1 - 04/27/2021 01:20 PM - Ilya Dryomov

- Priority changed from Normal to Urgent

This is being hit in production and causing some folks to downgrade to pre-5.3 kernels.

#2 - 06/15/2021 11:31 AM - Ilya Dryomov

- Status changed from New to In Progress

#3 - 07/06/2021 05:14 PM - Ilya Dryomov

- Status changed from In Progress to Fix Under Review

[PATCH] rbd: always kick acquire on "acquired" and "released" notifications
[PATCH] rbd: don't hold lock_rwsem while running_list is being drained

#4 - 07/06/2021 05:15 PM - Ilya Dryomov

- Duplicated by Bug #51136: Random hanging issues with rbd after network issues added

#5 - 07/26/2021 09:32 AM - Ilya Dryomov

- Status changed from Fix Under Review to Pending Backport

In 5.14-rc3:

https://git.kernel.org/pub/scm/linux/kernel/git/torvalds/linux.git/commit/?id=8798d070d416d18a75770fc19787e96705073f43
https://git.kernel.org/pub/scm/linux/kernel/git/torvalds/linux.git/commit/?id=ed9eb71085ecb7ded9a5118cec2ab70667cc7350

#6 - 08/10/2021 10:36 PM - Ilya Dryomov

- Status changed from Pending Backport to Resolved

In 5.4.136, 5.10.54 and 5.13.6.