Ceph - Bug #12673

cache agent is idle although one object is left in the cache

08/12/2015 08:25 AM - Loic Dachary

Status:	Resolved	**Start date:**	08/12/2015
Priority:	High	**Due date:**	
Assignee:	Loic Dachary	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:		**Severity:**	
Tags:		**Reviewed:**	
Backport:	firefly,hammer	**Affected Versions:**	
Regression:	No	**ceph-qa-suite:**	

Description

Steps to reproduce:

```
    ./stop.sh
    rm -fr out dev ; MON=1 OSD=3 ./vstart.sh -X -d -n -l mon osd

    ceph osd pool create slow  1 1
    ceph osd pool create fast  1 1
    ceph osd tier add slow fast
    ceph osd tier cache-mode fast writeback
    ceph osd tier set-overlay slow fast
    ceph osd pool set fast hit_set_type bloom
    rados -p slow put obj3 /etc/group
    ceph osd pool set fast target_max_objects  1
    ceph osd pool set fast hit_set_count 1
    ceph osd pool set fast hit_set_period 5

    sleep 30
    ceph df

    rados -p slow get obj3 /tmp/obj3
    ceph df

    sleep 30

    ceph df
    ceph health detail
```

The obj3 is not evicted from the fast pool.

```
$ ceph df
*** DEVELOPER MODE: setting PATH, PYTHONPATH and LD_LIBRARY_PATH ***
GLOBAL:
    SIZE     AVAIL      RAW USED     %RAW USED
    547G     53735M          494G         90.40
POOLS:
    NAME     ID     USED     %USED     MAX AVAIL     OBJECTS
    rbd      0         0         0        17909M           0
    slow     1      3834         0        17909M           3
    fast     2      1363         0        17909M           2
$ rados -p fast ls
obj3
```

Related issues:		
Copied to Ceph - Backport #12882: cache agent is idle although one object is ...	**Resolved**	**08/12/2015**
Copied to Ceph - Backport #12883: cache agent is idle although one object is ...	**Resolved**	**08/12/2015**

Associated revisions

Revision e1f58feb - 08/14/2015 12:31 PM - Loic Dachary

osd: trigger the cache agent after a promotion

When a proxy read happens, the object promotion is done in parallel. The agent_choose_mode function must be called to reconsider the situation to protect against the following scenario:

- proxy read
- agent_choose_mode finds no object exists and the agent goes idle
- object promotion happens
- the agent does not reconsider and eviction does not happen although it should

http://tracker.ceph.com/issues/12673 Fixes: #12673

Signed-off-by: Loic Dachary <ldachary@redhat.com>

Revision 79242319 - 08/14/2015 12:31 PM - Loic Dachary

tests: tiering agent and proxy read

Verify that an object promoted to a cache tier because of a proxy read is evicted as expected.

http://tracker.ceph.com/issues/12673 Refs: #12673

Signed-off-by: Loic Dachary <ldachary@redhat.com>

Revision aa911767 - 09/04/2015 03:13 PM - Loic Dachary

osd: trigger the cache agent after a promotion

When a proxy read happens, the object promotion is done in parallel. The agent_choose_mode function must be called to reconsider the situation to protect against the following scenario:

- proxy read
- agent_choose_mode finds no object exists and the agent goes idle
- object promotion happens
- the agent does not reconsider and eviction does not happen although it should

http://tracker.ceph.com/issues/12673 Fixes: #12673

Signed-off-by: Loic Dachary <ldachary@redhat.com>
(cherry picked from commit e1f58feb9b1d20b72f2eb2eefdea5982e0cddccd)

Revision 2c0d7fee - 09/04/2015 03:17 PM - Loic Dachary

tests: tiering agent and proxy read

Verify that an object promoted to a cache tier because of a proxy read
is evicted as expected.

http://tracker.ceph.com/issues/12673 Refs: #12673

Signed-off-by: Loic Dachary <ldachary@redhat.com>
(cherry picked from commit 7924231930732bd297d3bd034c8295e96cb81088)

Conflicts:
qa/workunits/cephtool/test.sh
resolved by manually adding the new test to TESTS

Revision 5656eec0 - 09/06/2015 09:28 PM - Loic Dachary

osd: trigger the cache agent after a promotion

When a proxy read happens, the object promotion is done in parallel. The
agent_choose_mode function must be called to reconsider the situation
to protect against the following scenario:

- proxy read
- agent_choose_mode finds no object exists and the agent goes idle
- object promotion happens
- the agent does not reconsider and eviction does not happen although it should

http://tracker.ceph.com/issues/12673 Fixes: #12673

Signed-off-by: Loic Dachary <ldachary@redhat.com>
(cherry picked from commit e1f58feb9b1d20b72f2eb2eefdea5982e0cddccd)

Revision 9f696601 - 09/06/2015 09:28 PM - Loic Dachary

tests: tiering agent and proxy read

Verify that an object promoted to a cache tier because of a proxy read
is evicted as expected.

http://tracker.ceph.com/issues/12673 Refs: #12673

Signed-off-by: Loic Dachary <ldachary@redhat.com>
(cherry picked from commit 7924231930732bd297d3bd034c8295e96cb81088)

History

#1 - 08/12/2015 08:25 AM - Loic Dachary

- Assignee set to Loic Dachary

- Affected Versions v0.94.2 added

#2 - 08/12/2015 02:24 PM - Loic Dachary

Confirmed on master as well.

#3 - 08/12/2015 04:13 PM - Loic Dachary

The cache agent does not wake up when a proxy read happens and promotes the object in ReplicatedPG::maybe_handle_cache. The last manifestation of the agent is when OSDService::agent_entry found an empty queue and started waiting on agent_cond.

```
2015-08-12 18:03:26.195127 7f2e7f7ae700 20 osd.2 17 agent_entry empty queue
```

After the rados get, agent_choose_mode is called

```
2015-08-12 18:03:52.175635 7f2e959ab700 10 osd.2 pg_epoch: 17 pg[2.0( v 17'5 (0'0,17'5] local-les=11 n=1 ec=10
 les/c 11/11 10/10/10) [2,1,0] r=0 lpr=10 luod=17'4 lua=17'4 crt=16'3 lcod 16'3 mlcod 16'3 active+clean] agent
_choose_mode flush_mode: idle evict_mode: idle num_objects: 1 num_bytes: 83 num_objects_dirty: 0 num_objects_o
map: 0 num_dirty: 0 num_user_objects: 0 num_user_bytes: 0 pool.info.target_max_bytes: 0 pool.info.target_max_o
bjects: 1
2015-08-12 18:03:52.175641 7f2e959ab700 20 osd.2 pg_epoch: 17 pg[2.0( v 17'5 (0'0,17'5] local-les=11 n=1 ec=10
 les/c 11/11 10/10/10) [2,1,0] r=0 lpr=10 luod=17'4 lua=17'4 crt=16'3 lcod 16'3 mlcod 16'3 active+clean] agent
_choose_mode dirty 0 full 0
```

and it could call agent_enable_pg or agent_adjust_pg which could call _enqueue which could signal agent_cond. But that apparently does not happen.

#4 - 08/12/2015 04:14 PM - Loic Dachary

- Description updated

#5 - 08/12/2015 05:20 PM - Sage Weil

The hit_set_period value of 5 is super low.. you probably want 600 at a minimum (10 minutes). But I'm not sure that's your problem.

Another possibility is that there are so few user objects in the pool vs the hit_set objects (1 + 1?), and there's an off-by-one or rounding error in the code that prevents the agent from busy looping trying to flush/evict. Or that a histogram of 1 object isn't behaving (e.g., showing that 1 object as 100th instead of 0th percentile vs the configured threshold).

Try putting more objects in the pool and see if the effect is still there? And increase the hit_set_period?

#6 - 08/12/2015 08:59 PM - Loic Dachary

- Status changed from Verified to Won't Fix

When using 50 objects instead of just one as in the tracker description, the cache eviction happens as expected.

```
    ./stop.sh
    rm -fr out dev ; MON=1 OSD=3 ./vstart.sh -X -d -n -l mon osd

    ceph osd pool create slow  1 1
    ceph osd pool create fast  1 1
    ceph osd tier add slow fast
    ceph osd tier cache-mode fast writeback
    ceph osd tier set-overlay slow fast
    ceph osd pool set fast hit_set_type bloom
    for i in $(seq 1 50) ; do
        rados -p slow put obj$i /etc/group
    done
    ceph osd pool set fast target_max_objects  1
    ceph osd pool set fast hit_set_count 1
    ceph osd pool set fast hit_set_period 5

    sleep 30
    ceph df

    for i in $(seq 1 5) ; do
        rados -p slow get obj$i /tmp/obj$i
    done
    ceph df

    sleep 30

    ceph df
    ceph health detail
```

The cache eviction mode determined by ReplicatedPG::agent_choose_mode does not count the objects, it relies on ratios. As Sage explains, that can lead to rounding / off by one errors when there are few objects. In other words, a single object may be stuck in the cache pool because the cache agent is not running.

Although these border cases are common when engineering a test case, they do not matter when there is a real workload because the agent will act properly if

- there are a few more objects
- a read or a write happen

If a test case needs to validate the eviction behavior, adding more objects as above is a workaround to avoid this rounding error.

Although it would be possible to fix this behavior so that it always do the right thing regardless, it probably is not worth the effort. It would slightly help someone crafting tests for the eviction logic. But this same someone would need to know about the eviction logic and understand why this situation should be avoided.

If someone feels differently, please re-open this issue.

#7 - 08/12/2015 09:00 PM - Loic Dachary

- Subject changed from object is not evicted to cache agent is idle although one object is left in the cache

#8 - 08/12/2015 09:01 PM - Loic Dachary

- Affected Versions deleted (v0.94.2)

#9 - 08/13/2015 05:44 AM - Kefu Chai

Loïc, just want to rectify some points here.

> agent_choose_mode flush_mode: idle evict_mode: idle **num_objects: 1 num_bytes: 83** num_objects_dirty: 0 num_objects_omap: 0 num_dirty: 0 **num_user_objects: 0 num_user_bytes: 0** pool.info.target_max_bytes: 0 pool.info.target_max_objects: 1

so there was **1** hit_set_archive object which took 83 bytes. and this sort of objects is not taken into the consideration of cache eviction. as what we care about is only the number/size of user objects. that's the root cause why we failed to evict this non-user object in your test.

> The cache eviction mode determined by ReplicatedPG::agent_choose_mode does not count the objects, it relies on ratios.

yes, so we need to take following things into consideration

- the target_max_bytes/target_max_objects are per-pool settings, but we are doing the evict at the PG level. so we normalize the number using pg_num. but your pool's pg_num is 1. so we are fine.
- even if the number of **user** object is just 1, we can still have it evicted. as long as it is a genuine object.

probably we should update the section of "ceph df" section in manpage of "ceph" to clarify that there could be some objects are non-user object, for example, are used by hit_set archiving and/or objects-omap.

#10 - 08/13/2015 06:50 AM - Loic Dachary

- Tracker changed from Bug to Documentation

- Status changed from Won't Fix to Verified

- Assignee deleted (Loic Dachary)

Even if this is not fixed, it should be documented in a place where the developer / tester / user facing this situation will find it when looking for an explanation.

#11 - 08/13/2015 10:25 AM - Loic Dachary

- Tracker changed from Documentation to Bug

- Assignee set to Loic Dachary

- Regression set to No

#12 - 08/13/2015 10:33 AM - Loic Dachary

Running the script in the description, here is what I find in the logs: when the agent_choose_mode function runs at 2015-08-13 11:47:33.129187 it finds there are **num_user_objects: 0**. Later on at 2015-08-13 11:47:33.131668 **finish_promote 2/6cf8deff/obj1/head** happens and if agent_choose_mode was to be run again, it would find that there is one user_objects. But it does not run and does not get a chance to reconsider.

#13 - 08/13/2015 10:34 AM - Loic Dachary

- File osd.2.log.gz added

#14 - 08/13/2015 11:55 AM - Loic Dachary

- Status changed from Verified to Need Review

https://github.com/ceph/ceph/pull/5570

#15 - 08/13/2015 01:42 PM - Kefu Chai

- Backport set to firefly,hammer

#16 - 08/19/2015 02:47 PM - Loic Dachary

- Priority changed from Normal to High

#17 - 08/24/2015 10:52 AM - Kefu Chai

- Status changed from Need Review to Testing

#18 - 08/31/2015 04:05 AM - Kefu Chai

- Status changed from Testing to Pending Backport

#19 - 10/20/2015 07:40 PM - Loic Dachary

- Status changed from Pending Backport to Resolved

Files

osd.2.log.gz	727 KB	08/13/2015	Loic Dachary