rgw - Bug #10295

rgw deleting S3 objects leaves __shadow_ objects behind

12/11/2014 05:55 AM - Richard Chan

<table>
<thead>
<tr>
<th>Status:</th>
<th>Resolved</th>
<th>Start date:</th>
<th>12/11/2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority:</td>
<td>Urgent</td>
<td>Due date:</td>
<td></td>
</tr>
<tr>
<td>Assignee:</td>
<td>Orit Wasserman</td>
<td>% Done:</td>
<td>0%</td>
</tr>
<tr>
<td>Category:</td>
<td></td>
<td>Estimated time:</td>
<td>0.00 hour</td>
</tr>
<tr>
<td>Target version:</td>
<td></td>
<td>Spent time:</td>
<td>0.00 hour</td>
</tr>
<tr>
<td>Source:</td>
<td>other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tags:</td>
<td></td>
<td>Reviewed:</td>
<td></td>
</tr>
<tr>
<td>Backport:</td>
<td>hammer, firefly</td>
<td>Affected Versions:</td>
<td>ceph-qa-suite:</td>
</tr>
<tr>
<td>Regression:</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Severity:</td>
<td>2 - major</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Description

Using 0.80.7, observed also on .87,.89

1. Create 100 objects 10MB ea in rgw.
2. Observe that .rgw.buckets occupies 1GB data (as expected).
3. Delete all 100 objects.

4. rgw is now empty

   1. radosgw-admin bucket list --bucket=test

      []

5. Space occupied is unexpectedly high.

rados df: .rgw.buckets 11 959M 0 7447G 303

There are 303 objects with names like default.4209.1__shadow__.cC7M_LMTe8oNgW0GZOINCIJco9cEEOI_1

ceph df: the object store is basically empty but 3GB RAW is used

GLOBAL:

<table>
<thead>
<tr>
<th>SIZE</th>
<th>AVAIL</th>
<th>RAW USED</th>
<th>%RAW USED</th>
</tr>
</thead>
<tbody>
<tr>
<td>22345G</td>
<td>22342G</td>
<td>3126M</td>
<td>0.01</td>
</tr>
</tbody>
</table>

POOLS:

<table>
<thead>
<tr>
<th>NAME</th>
<th>ID</th>
<th>USED</th>
<th>%USED</th>
<th>MAX</th>
<th>AVAIL</th>
<th>OBJECTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>data</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7447G</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>metadata</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7447G</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>rbd</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>7447G</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>.rgw.root</td>
<td>3</td>
<td>840</td>
<td>0</td>
<td>7447G</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>.rgw.control</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>7447G</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>.rgw</td>
<td>5</td>
<td>331</td>
<td>0</td>
<td>7447G</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>.rgw.gc</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>7447G</td>
<td>521</td>
<td>4</td>
</tr>
<tr>
<td>.users.uid</td>
<td>7</td>
<td>1341</td>
<td>0</td>
<td>7447G</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>.users.email</td>
<td>8</td>
<td>43</td>
<td>0</td>
<td>7447G</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>.users</td>
<td>9</td>
<td>43</td>
<td>0</td>
<td>7447G</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>.rgw.buckets.index</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>7447G</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>.rgw.buckets</td>
<td>11</td>
<td>959M</td>
<td>0</td>
<td>7447G</td>
<td>303</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7447G</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Related issues:

Copied to Backport #11719: rgw deleting S3 objects leaves __shadow_ objects b... Resolved 12/11/2014

Copied to Backport #11720: rgw deleting S3 objects leaves __shadow_ objects b... Resolved 12/11/2014
Need to wait for gc to run its course. Not all entries are going to show in the gc listing immediately. You need to run `radosgw-admin gc list --include-all`.

Tested again: reduced rgw_gc settings to 300s.

On 0.80.7, the shadow object were purged.

On 0.87, 0.89 the garbage collection does not happen and remain behind.

```bash
1. ceph --admin-daemon /var/run/ceph/ceph-mon.b.asok config show | grep gc
   "rgw_gc_max_objs": "521",
   "rgw_gc_obj_min_wait": "1200",
   "rgw_gc_processor_max_time": "600",
   "rgw_gc_processor_period": "600",
```

On 0.89:

```bash
[root@radosgw ~]# ceph df
GLOBAL:
SIZE    AVAIL    RAW USED    %RAW USED
596G     588G     8402M      1.38
.rgw.buckets  16    3962M    0.65    294G    14885
```

3G is correct "current" blobs.
1G of test data was created and rm'ed.

The "extra" 962M is not purged.
2472 default.95874.2_shadow_* objects

Reproduced this on a fresh 0.89.

Created 400 objects, deleted 300.

Found 3 objects in gc list --include-all.

After running gc process I have 697 objects in .rgw.buckets.
1. 100 objects are "real objects"
2. 597 objects are "shadow objects" (600-3; not sure why only 3 objects appeared in gc list --include-all)
3. Now gc list --include-all is empty
4. This pool has size = 2; on the 3 OSD /var/lib/ceph/osd/ceph-XX there are 1194 files with names shadow

find /var/lib/ceph | grep shadow | wc -l across all OSDs show
1194 files.

#4 - 12/15/2014 08:44 AM - Yehuda Sadeh
Do you by any chance have non default zone configuration?

#5 - 12/15/2014 09:02 AM - Richard Chan
- File config_show.txt added

No - I am following the exemplary simple configuration
with default zones; only changes are custom cluster name
and civetweb port=80

1. rados --cluster saturn lspools
   rbd
   .rgw.root
   .rgw.control
   .rgw
   .rgw.gc
   .users.uid
   .users.email
   .users
   .rgw.buckets.index
   .rgw.buckets

Every time an object is deleted the radosgw log file produces:

```
2014-12-15 23:57:02.572230 7fa4effb7700  1 ====== starting new request req=0x7fa52c062f40 ======
2014-12-15 23:57:02.579443 7fa4effb7700  0 RGWObjManifest::operator++(): result: ofs=4718592 stripe_ofs=4718592 part_ofs=0 rule->part_size=0
2014-12-15 23:57:02.579449 7fa4effb7700  0 RGWObjManifest::operator++(): result: ofs=8912896 stripe_ofs=8912896 part_ofs=0 rule->part_size=0
2014-12-15 23:57:02.579453 7fa4effb7700  0 RGWObjManifest::operator++(): result: ofs=10485760 stripe_ofs=10485760 part_ofs=0 rule->part_size=0
2014-12-15 23:57:02.579556 7fa4effb7700  1 ====== req done req=0x7fa52c062f40 http_status=204 ======
2014-12-15 23:57:02:579943 7fa4effb7700 0 RGWObjManifest::operator++(): result: ofs=4718592 stripe_ofs=4718592 part_ofs=0 rule->part_size=0
2014-12-15 23:57:02.579949 7fa4effb7700 0 RGWObjManifest::operator++(): result: ofs=8912896 stripe_ofs=8912896 part_ofs=0 rule->part_size=0
2014-12-15 23:57:02.579953 7fa4effb7700 0 RGWObjManifest::operator++(): result: ofs=10485760 stripe_ofs=10485760 part_ofs=0 rule->part_size=0
2014-12-15 23:57:02.579956 7fa4effb7700 1 ====== req done req=0x7fa52c062f40 http_status=204 ======
2014-12-15 23:57:02.579962 7fa4effb7700 0 RGWObjManifest::operator++(): result: ofs=4718592 stripe_ofs=4718592 part_ofs=0 rule->part_size=0
2014-12-15 23:57:02.580003 7fa4effb7700 1     civetweb: 0x7fa52c0139d0: 192.168.122.1 - - [15/Dec/2014:23:57:02 +0800] "DELETE /testdata/data292 HTTP/1.1" -1 0 - -
```
It is a civetweb bug.

I added 100 objects and rm'ed - 10M ea, 300 shadow objects created. All 300 shadow objects appear in gc list --include-all and are purged.

Should have written:

It is a civetweb bug.

With apache/fastcgi: I added 100 objects and rm'ed - 10M ea, 300 shadow objects created. All 300 shadow objects appear in gc list --include-all and are purged.

Summary of bug:

civetweb 0.89 is not correctly preparing the garbage collection list:
1. create 100 rgw objects (10M size) so 400 rados objects are created in .rgw.buckets
2. rm 100 rgw objects, 300 rados objects left in .rgw.bucket -- all shadow objects
3. gc list --include-all only shows 3 oids

With apache/fastcgi
gc list --include-all will show 300 oids

Not sure how civetweb could affect gc work. I do see that your rgw_gc_max_objs is non standard. Did you put this config on the global section, or only under the section for the specific rgw user?

- File saturn-conf.txt added

The rgw parameters are in the global section.

civetweb:

when you delete a second object, it does not appear in gc list --include-all

How to reproduce:
1. add two objects (ea 10M)
2. rm one object, observe 3 oids in gc list --include-all
2. rm 2nd object, observe only 3 oids in gc list --include-all

s3cmd s3://testdat2
1. empty output
2. radosgw-admin gc list --include-all
   [   { "tag": "default.4292.0", "time": "2014-12-16 10:58:47.729479", "objs": [   { "pool": ".rgw.buckets", "oid": "default.4292.1__shadow_.0ZO019kstnvzHUTrGW0zz1sqdl08Efm_1", "key": ""},   { "pool": ".rgw.buckets", "oid": "default.4292.1__shadow_.0ZO019kstnvzHUTrGW0zz1sqdl08Efm_2", "key": ""},   { "pool": ".rgw.buckets", "oid": "default.4292.1__shadow_.0ZO019kstnvzHUTrGW0zz1sqdl08Efm_3", "key": ""}]}

#12 - 12/15/2014 06:44 PM - Richard Chan
Revert rgw_gc_max_objs to 32, problem remains.

#13 - 12/16/2014 01:36 AM - Richard Chan
civetweb seems to "corrupt" the gc list.
1. rm one object (10M size) with civetweb, 3 oids in gc list
2. rm 2nd object with apache/fastcgi, no new oids appear

When the gc list is clear, delete two 10M objects with apache/fastcgi:
6 oids appear

#14 - 12/16/2014 08:33 AM - Yehuda Sadeh
- Priority changed from Normal to High

Can you provide logs to both your apache and civetweb gateways, when removing objects? (set 'debug rgw = 20', and 'debug ms = 1'). Also, if you’re running 'radosgw-admin gc process' command, then logs for that (same log settings, could do --log-to-stderr)

#15 - 12/16/2014 04:23 PM - Richard Chan
- File access_log added
- File client.radosgw.gw1.log added

Logs with good gc behaviour:
1. rm 1 object # 3 oids expected
   [root@cephadm ~]# radosgw-admin gc list --include-all 2>/dev/null | grep oid
   "oid": "default.7460.1__shadow_.FLJn2SDW9cQ6o7W_D5j7FSIXiXg0v0d_1",
   "oid": "default.7460.1__shadow_.FLJn2SDW9cQ6o7W_D5j7FSIXiXg0v0d_2",
   "oid": "default.7460.1__shadow_.FLJn2SDW9cQ6o7W_D5j7FSIXiXg0v0d_3",
2. rm 2nd object: 6 oids expected
   [root@cephadm ~]# radosgw-admin gc list --include-all 2>/dev/null | grep oid
   "oid": "default.7460.1__shadow_.FLJn2SDW9cQ6o7W_D5j7FSIXiXg0v0d_1",
Suddenly civetweb is working, huh!!??

1. Delete 3 objects in succession

   [   [ "tag": "default.7460.27", "time": "2014-12-17 08:37:05.149612", "objs": [   { "pool": ".rgw.buckets", "oid": "default.7460.1__shadow_.qPur-x0VhCPNAZ9Dmhz2eaH9dhCwQH_1", "key": []},   { "pool": ".rgw.buckets", "oid": "default.7460.1__shadow_.qPur-x0VhCPNAZ9Dmhz2eaH9dhCwQH_2", "key": []},   { "pool": ".rgw.buckets", "oid": "default.7460.1__shadow_.qPur-x0VhCPNAZ9Dmhz2eaH9dhCwQH_3", "key": []}]],  [ "tag": "default.7460.25", "time": "2014-12-17 08:35:42.837098", "objs": [   { "pool": ".rgw.buckets", "oid": "default.7460.1__shadow_.V4tNCUJugxkYAOjtaoZSK0i5UQulP2M_1", "key": []},   { "pool": ".rgw.buckets", "oid": "default.7460.1__shadow_.V4tNCUJugxkYAOjtaoZSK0i5UQulP2M_2", "key": []},   { "pool": ".rgw.buckets", "oid": "default.7460.1__shadow_.V4tNCUJugxkYAOjtaoZSK0i5UQulP2M_3", "key": []}]],  [ "tag": "default.7460.26", "time": "2014-12-17 08:36:02.235228", "objs": [   { "pool": ".rgw.buckets", "oid": "default.7460.1__shadow_.KcaqRDcqua8JhHay-E4hyF3E2L8WSr4_1", "key": []},   { "pool": ".rgw.buckets", "oid": "default.7460.1__shadow_.KcaqRDcqua8JhHay-E4hyF3E2L8WSr4_2", "key": []},   { "pool": ".rgw.buckets", "oid": "default.7460.1__shadow_.KcaqRDcqua8JhHay-E4hyF3E2L8WSr4_3", "key": []}]]]
The corruption occurs if civetweb is the first thing that touches a new rgw setup.

How to reproduce:
1. rmpool .users* .rgw*
2. recreate new rgw instance
3. make sure civetweb is the only interaction with the object store. create new user, create bucket, create objects.
4. delete two objects in succession

Expected 6 oids:

```
[root@cephadm ~]# radosgw-admin gc list --include-all 2>/dev/null| grep oid
"oid": "default.7634.1__shadow__e6kccdTEMgBAdi0wTHGykZvDmkt_Ck2_1",
"oid": "default.7634.1__shadow__e6kccdTEMgBAdi0wTHGykZvDmkt_Ck2_2",
"oid": "default.7634.1__shadow__e6kccdTEMgBAdi0wTHGykZvDmkt_Ck2_3",
```

#18 - 12/16/2014 05:44 PM - Richard Chan

- File client.radosgw.gw1.log added

Weird - the problem occurs with objects created by civetweb.

How to reproduce:
1. create new rgw instance
2. create user, bucket
3. create 4 10m objects: 2 by civetweb, 2 by fastcgi
4. delete the four objects. observe each deletion of fastcgi object adds 3 oids to gc list
deletion of civetweb objects adds 3 oids (the first object deleted) and no more.

#19 - 12/16/2014 05:54 PM - Richard Chan

1. The issue is reproducible only on objects created by civetweb. The bucket can be created by either civetweb or fastcgi.
   Only oids of the first deleted civetweb object deleted appear in the gc list.
   All oids of deleted fastcgi objects appear in the gc list.
2. Add 100 objects by fastcgi, delete by fastcgi/civetweb.
   No issues here — 300 oids added to gc list
3. Add 100 objects by civetweb, delete by fastcgi/civetweb.
   Only 3 oids added to gc list.

11/25/2015
Finally..mini test case.

1. New rgw instance
2. Create bucket + 4 x 10m objects (2 by fastcgi, 2 by civetweb)
3. Delete 4 objects

s3cmd -c saturn.cfg mb s3://test

1. this uses port 80, fastcgi
   s3cmd -c saturn80.cfg put test10m s3://test/data1
   s3cmd -c saturn80.cfg put test10m s3://test/data2
2. this uses port 7480, civetweb
   s3cmd -c saturn.cfg put test10m s3://test/data3
   s3cmd -c saturn.cfg put test10m s3://test/data4 ##
   s3cmd -c saturn.cfg ls

2014-12-17 02:09  10485760   s3://test/data1
2014-12-17 02:09  10485760   s3://test/data2
2014-12-17 02:09  10485760   s3://test/data3
2014-12-17 02:09  10485760   s3://test/data4 ##

for k in 1 2 3 4; do
   s3cmd -c saturn.cfg rm s3://test/$k
done

How reproducible:
Always

Result:
9 oids in gc list

Expected:
12 oids in gc list

---

I'm facing with the same issue and I realised when I restart the radosgw the gc became clean but the shadow objects still the ceph pool, see below:

root@cephgw0001---> radosgw-admin gc list
[   {"tag": "default.1552327.7",
    "time": "2015-01-26 18:42:39.532908",
    "objs": [        {"pool": ".rgw.buckets.replicated,"},

---

#20 - 12/16/2014 06:13 PM - Richard Chan

- File client.radosgw.gw1.log added

#21 - 01/26/2015 09:14 PM - Italo Santos

I'm facing with the same issue and I realised when I restart the radosgw the gc became clean but the shadow objects still the ceph pool, see below:
"oid": "default.1507466.1__shadow_.lyntzWkaFPPgGfYXhZBuCIzl4Vd0TT_1",
"key": ""]
]}

```
root@cephgw0001:~# /etc/init.d/radosgw restart
Starting client.radosgw.gateway...
2015-01-26 19:10:48.573344 7fb6c1024780 -1 WARNING: libcurl doesn't support curl_multi_wait() 2015-01-26 19:10:48.573394 7fb6c1024780 -1 WARNING: cross zone / region transfer performance may be affected /usr/bin/radosgw is running.

root@cephgw0001:~# radosgw-admin gc list

root@cephgw0001:~# rados -p .rgw.buckets.replicated ls
default.1554751.1__shadow_.aoCI7rTnBbJKfb5jogxvLStSebJnL27_1
default.1554751.1__shadow_.lyntzWkaFPPgGfYXhZBuCIzl4Vd0TT_1

#22 - 01/27/2015 04:55 PM - Italo Santos
Also, I'm using Apache + FastCGI... exclude a possible issue with civetweb.

#23 - 04/13/2015 05:43 PM - Ian Colle
- Assignee set to Yehuda Sadeh
- Priority changed from High to Urgent

#24 - 04/13/2015 06:38 PM - Yehuda Sadeh
- Assignee changed from Yehuda Sadeh to Orit Wasserman

#25 - 04/23/2015 08:31 PM - Yehuda Sadeh
- Status changed from New to Testing

#26 - 04/24/2015 04:27 PM - Yehuda Sadeh
- Status changed from Testing to Pending Backport
- Backport set to hammer, firefly

#27 - 04/24/2015 04:29 PM - Yehuda Sadeh
merged, commit:26225980c2fd018e70033a453d635533fcdefec

#28 - 05/05/2015 09:30 PM - Yehuda Sadeh
hammer: 4e84f318de97e592e16493c67491ba0d7f8103a8 matching pull request https://github.com/ceph/ceph/pull/4515

#29 - 06/05/2015 08:26 AM - Nathan Cutler
#30 - 06/29/2015 09:34 AM - Gleb Borisov
How we can remove such objects from storage? We were running 0.94 with civetweb frontend for a couple of month and now it's about 70TiB of 

garbage in our buckets :-(

I think this recipe must be included in release notes, since many of us running same configuration.

---

#31 - 06/30/2015 12:16 AM - Yehuda Sadeh
I've been working on a tool to identify these orphan objects. It's still in the works, but I'll try to get it ready in the following days.

---

#32 - 07/09/2015 02:54 AM - Ben Hines
Yehuda Sadeh wrote:

I've been working on a tool to identify these orphan objects. It's still in the works, but I'll try to get it ready in the following days.

FWIW we were never using civetweb, always fastcgi, and i show very large difference between ceph df output vs bucket stats size. Looking forward to 

what the tool produces, though.

---

#33 - 09/13/2015 02:11 AM - George Mihaiescu
Hi Yehuda,

Do you have an update on this bug?

I'm running Ceph 0.94.3 with civetweb and I've run some tests with cosbench that clean up the data they generate, but a week later I see the data is 

still there.

I have only 20 MB of valid data in the .rgw.buckets pool, but "ceph df" reports 26 GB.

The garbage collector doesn't know of any objects that need to be deleted:

$ radosgw-admin gc list --include-all

[]

The settings used by garbage collector:

$ ceph --admin-daemon /var/run/ceph/ceph-mon.controller1.aok config show | grep gc

"xio_trace_msgcnt": "false",
"osd_pool_default_flag_nopgchange": "false",
"rgw_enable_gc_threads": "true",
"rgw_gc_max_objs": "32",
"rgw_gc_obj_min_wait": "7200",
"rgw_gc_processor_max_time": "3600",
"rgw_gc_processor_period": "3600",

$ ceph df | grep .rgw.buckets

| .rgw.buckets.index | 12 | 0 | 0 | 520T | 14 |
| .rgw.buckets.extra | 15 | 0 | 0 | 520T | 1  |
| .rgw.buckets     | 25 | 26037M | 0 | 520T | 428 |

There are 417 objects that should be removed

$ rados ls -p .rgw.buckets | grep -i -c shadow

417

I even changed rgw_gc_max_objs = 200 thinking that the default 32 would take too long to clean up maybe, then I restarted radosgw and ran
"radosgw-admin gc process", but no difference.

The short term solution would be to manually delete all these "shadow" objects.

I have three radosgw instances running behind haproxy and use 64M stripe size, if it makes any difference.

Thank you,

George

#34 - 09/14/2015 03:43 PM - Yehuda Sadeh

Is the data that you see new, or is it the same old data that leaked before running 0.94.3? What do you see when you run 'radosgw-admin gc list --include-all' right after running cosbench? Try to avoid running 'gc process' using radosgw-admin when you test this (as you modified the number of shards, you might have an issue with mismatching shard numbers).

Files

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>Date</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>config_show.txt</td>
<td>27.2 KB</td>
<td>12/15/2014</td>
<td>Richard Chan</td>
</tr>
<tr>
<td>saturn-conf.txt</td>
<td>887 Bytes</td>
<td>12/15/2014</td>
<td>Richard Chan</td>
</tr>
<tr>
<td>access_log</td>
<td>274 Bytes</td>
<td>12/16/2014</td>
<td>Richard Chan</td>
</tr>
<tr>
<td>client.radosgw.gw1.log</td>
<td>1.53 MB</td>
<td>12/16/2014</td>
<td>Richard Chan</td>
</tr>
<tr>
<td>client.radosgw.gw1.log</td>
<td>2.32 MB</td>
<td>12/16/2014</td>
<td>Richard Chan</td>
</tr>
<tr>
<td>client.radosgw.gw1.log</td>
<td>1.21 MB</td>
<td>12/16/2014</td>
<td>Richard Chan</td>
</tr>
<tr>
<td>client.radosgw.gw1.log</td>
<td>1.74 MB</td>
<td>12/16/2014</td>
<td>Richard Chan</td>
</tr>
<tr>
<td>client.radosgw.gw1.log</td>
<td>2.51 MB</td>
<td>12/16/2014</td>
<td>Richard Chan</td>
</tr>
</tbody>
</table>