Ceph - Bug #11454

pg not being marked undersized in EC pools when k+m > number of hosts with ruleset-failure-domain set to host

04/22/2015 10:42 PM - Kyle Bader

Status: Resolved
Priority: Urgent
Assignee: Kyle Bader
Category: 
Target version: 
Source: other
Backport: firefly
Regression: No
Severity: 3 - minor

Description
With a k=4, m=2, ruleset-failure-domain=host erasure code profile I am able to create a pool and have its pgs reach active+clean status (albeit very slowly) with 5 hosts. Instead of going active+clean the pool's pgs should remain active+degraded until a 6th host is added to the cluster.

[cphuser@mgmt example]$ ceph -s
cluster 52829f26-6dba-493a-bf4b-b87f1f28187f
health HEALTH_OK
monmap e4: 1 mons at [mgmt=172.27.50.50:6789/0], election epoch 1, quorum 0 mgmt
osdmap e43422: 178 osds: 178 up, 178 in
pgmap v266351: 1024 pgs, 1 pools, 0 bytes data, 0 objects
40711 MB used, 484 TB / 484 TB avail
1024 active+clean

[cphuser@mgmt example]$ ceph osd erasure-code-profile get myec
directory=/usr/lib64/ceph/erasure-code
k=4
m=2
plugin=jerasure
ruleset-failure-domain=host
technique=reed_sol_van

[cphuser@mgmt example]$ ceph osd tree | grep host | wc -l
5

[cphuser@mgmt example]$ ceph pg dump | grep 2147483647 | wc -l
dumped all in format plain
1024

[cphuser@mgmt example]$ ceph pg dump | grep clean | wc -l
dumped all in format plain
1024

Associated revisions
Revision e1d57730 - 04/23/2015 04:51 PM - Guang Yang
PG::actingset should be used when checking the number of acting OSDs for a given PG.
Signed-off-by: Guang Yang <yguang@yahoo-inc.com>
(cherry picked from commit 19be358322be48faaf17b28054619a8b5e7d403b)

Conflicts:
src/osd/PG.cc   PG::get_backfill_priority() doesn't exist in firefly
Variation in code related to no "undersized" state in firefly
History

#1 - 04/22/2015 10:42 PM - Kyle Bader
[cephuser@mgmt example]$ ceph -v
ceph version 0.80.8 (69eaad7f8308f21573c604f121956e64679a52a7)
We'll want to backport this to firefly!

#2 - 04/22/2015 10:45 PM - Kyle Bader

```
{ 
  "rule_id": 27,
  "rule_name": "foo",
  "ruleset": 27,
  "type": 3,
  "min_size": 3,
  "max_size": 20,
  "steps": [
    { "op": "set_chooseleaf_tries", "num": 5 },
    { "op": "take", "item": -1 },
    { "op": "chooseleaf_indep", "num": 0, "type": "host" },
    { "op": "emit" }
  ]
}
```

#3 - 04/23/2015 03:08 PM - Sage Weil
- Assignee set to David Zafman
- Priority changed from Normal to Urgent
- Backport changed from Firefly to firefly

this likely only affects firefly since this code was totally rewritten for giant/hammer.

#4 - 04/23/2015 04:11 PM - David Zafman
- Status changed from New to In Progress

Fixed in:

commit 19be358322be48faa17b28054619a8b5e7d403b
Author: Guang Yang <yguang@yahoo-inc.com>
Date: Mon Sep 29 08:21:10 2014 +0000

PG::actingset should be used when checking the number of acting OSDs for a given PG.
Signed-off-by: Guang Yang &lt;yguang@yahoo-inc.com&gt;

Since actingset doesn't include CRUSH_ITEM_NONE from acting, it would be the count needed to determine "degraded" state
#5 - 04/23/2015 07:20 PM - David Zafman
- Status changed from In Progress to Testing
- Assignee changed from David Zafman to Kyle Bader

Created pull request marked needs-qa:

- firefly backport https://github.com/ceph/ceph/pull/4453

#6 - 05/13/2015 07:27 PM - Loic Dachary
- Regression set to No

- master https://github.com/ceph/ceph/pull/2616

#7 - 05/13/2015 07:27 PM - Loic Dachary
- Status changed from Testing to Resolved

#8 - 05/13/2015 07:29 PM - Loic Dachary

$ commit=e1d5773 ; picked_from=$(git show --no-patch --pretty=%b $commit | perl -ne 'print if(s/.*cherry picked from commit \(w+\).*$/\1/)) ; diff -u --ignore-matching-lines "^[+-]" <(git show $picked_from) <(git show $commit)

@@ -13,23 +21,12 @@
// degraded?
   - if (get_osdmap()->get_pg_size(info.pgid.pgid) > acting.size()) {
     state_set(PG_STATE_DEGRADED);
-       state_set(PG_STATE_UNDERSIZED);
-     }
-@@ -1915,8 +1915,8 @@ unsigned PG::get_backfill_priority()
-   // undersized: 200 + num missing replicas
-   if (is_undersized()) { 
-     assert(pool.info.size > acting.size());
-     return 200 + (pool.info.size - acting.size());
-+    assert(pool.info.size > actingset.size());
-+    return 200 + (pool.info.size - actingset.size());
-    }
-   }
-   // degraded: baseline degraded
@@ -6266,13 +6266,11 @@ boost::statechart::result PG::RecoveryState::Active::react(const AdvMap& advmap)
+@@ -6404,13 +6404,11 @@ boost::statechart::result PG::RecoveryState::Active::react(const AdvMap& advmap)
++    if (pg->get_osdmap()->get_pg_size(pg->info.pgid.pgid) <= pg->actingset.size())
++     pg->state_set(PG_STATE_DEGRADED);
++   else
++     pg->state_set(PG_STATE_DEGRADED);