Ceph - Bug #13627

mon: last seen election epoch not honored in win_standalone_election()

10/28/2015 06:20 AM - huanwen ren

Status:	Resolved	**Start date:**	10/28/2015
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:	Monitor	**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:	other	**Reviewed:**	
Tags:		**Affected Versions:**	v0.94.2
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor		

Description

The Monitor of epoch version is set to 1, when number by multiple downgraded to a mon
The sample:

```
[[root@node181 ~]# ceph mon_status -f json-pretty

{ "name": "node173",
  "rank": 0,
  "state": "leader",
  "election_epoch": 4,
  "quorum": [
        0,
        1],
  "outside_quorum": [],
  "extra_probe_peers": [
        "10.118.202.181:6789\/0"],
  "sync_provider": [],
  "monmap": { "epoch": 1,
      "fsid": "7028dd47-e1bb-4dd7-9954-818bfa8e43fe",
      "modified": "0.000000",
      "created": "0.000000",
      "mons": [
            { "rank": 0,
              "name": "node173",
              "addr": "10.118.202.173:6789\/0"},
            { "rank": 1,
              "name": "node181",
              "addr": "10.118.202.181:6789\/0"}]}}
[root@node181 ~]# ceph mon_status -f json-pretty

{ "name": "node173",
  "rank": 0,
  "state": "leader",
  "election_epoch": 4,
  "quorum": [
        0,
        1],
  "outside_quorum": [],
  "extra_probe_peers": [
        "10.118.202.181:6789\/0"],
  "sync_provider": [],
  "monmap": { "epoch": 1,
      "fsid": "7028dd47-e1bb-4dd7-9954-818bfa8e43fe",
      "modified": "0.000000",
      "created": "0.000000",
      "mons": [
            { "rank": 0,
              "name": "node173",
              "addr": "10.118.202.173:6789\/0"},
            { "rank": 1,
              "name": "node181",
              "addr": "10.118.202.181:6789\/0"}]}}
[root@node181 ~]# ceph mon remove node173
removed mon.node173 at 10.118.202.173:6789/0, there are now 1 monitors
[root@node181 ~]# ceph mon_status -f json-pretty
2015-10-28 10:04:09.787510 7f3fcc7f8700  0 -- :/1022052 >> 10.118.202.173:6789/0 pipe(0x7f3fc40230
d0 sd=3 :0 s=1 pgs=0 cs=0 l=1 c=0x7f3fc4023360).fault

{ "name": "node181",
  "rank": 0,
  "state": "leader",
  "election_epoch": 5,
  "quorum": [
        0],
  "outside_quorum": [],
  "extra_probe_peers": [
        "10.118.202.173:6789\/0"],
  "sync_provider": [],
  "monmap": { "epoch": 2,
      "fsid": "7028dd47-e1bb-4dd7-9954-818bfa8e43fe",
      "modified": "2015-10-28 09:59:44.616088",
      "created": "0.000000",
      "mons": [
            { "rank": 0,
              "name": "node181",
              "addr": "10.118.202.181:6789\/0"}]}}
[root@node181 ~]# service ceph stop mon
=== mon.node181 ===
Stopping Ceph mon.node181 on node181...kill 11170...done
[root@node181 ~]# service ceph stop mon^C
[root@node181 ~]# service ceph start mon
=== mon.node181 ===
Starting Ceph mon.node181 on node181...
Running as unit run-24187.service.
Starting ceph-create-keys on node181...
[root@node181 ~]# ceph mon_status -f json-pretty
2015-10-28 10:06:17.657206 7f8ac43cc700  0 -- :/1024231 >> 10.118.202.173:6789/0 pipe(0x7f8abc0230
d0 sd=3 :0 s=1 pgs=0 cs=0 l=1 c=0x7f8abc023360).fault

{ "name": "node181",
  "rank": 0,
  "state": "leader",
  "election_epoch": 1,
  "quorum": [
        0],
  "outside_quorum": [],
  "extra_probe_peers": [],
  "sync_provider": [],
  "monmap": { "epoch": 2,
      "fsid": "7028dd47-e1bb-4dd7-9954-818bfa8e43fe",
      "modified": "2015-10-28 09:59:44.616088",
      "created": "0.000000",
      "mons": [
            { "rank": 0,
              "name": "node181",
              "addr": "10.118.202.181:6789\/0"}]}}
[root@node181 ~]#
```

After I remove node173 Monitor, check the Monitor in the cluster election_epoch is "5"
But I restart node181 Monitor node, check the cluster Monitor election_epoch into "1"

I think no matter whether to restart the only Mon node in a cluster, all should be in the cluster Mon election_epoch set for the final state, as in the example below the Mon election_epoch should be "5"

Associated revisions

Revision 43ba8200 - 10/31/2015 05:12 AM - renhwztetecs

mon:honour last seen election epoch in win_standalone_election()

add the elector.init() and elector.get_epoch() into Monitor::win_standalone_election() to initialise the new election epoch with last election epoch+1

Fixes: #13627
Signed-off-by: huanwen ren <ren.huanwen@zte.com.cn>

History

#1 - 10/28/2015 07:11 AM - huanwen ren

https://github.com/ceph/ceph/pull/6407

#2 - 10/28/2015 08:56 AM - Joao Eduardo Luis

- Category set to Monitor

Can you actually point us to an incorrect behavior arising from this?

I have never considered this a bug. The election epoch does not need to be durable, and making it so without a really good reason (say, an actual bug) is adding unnecessary complexity.

#3 - 10/28/2015 09:02 AM - Joao Eduardo Luis

- Subject changed from The Monitor of epoch version is set to 1, when number by multiple downgraded to a mon to mon: last seen election epoch not honored in win_standalone_election()

- Status changed from New to Verified

Nevermind, I figured what you actually mean. The monitor is not honoring its last seen election epoch when running on a single-monitor cluster.

This does not cause problems (given one is on a single-monitor cluster), but yeah... it's weird.

The pull request should go through some QA, but looks okay.

#4 - 10/28/2015 09:36 AM - huanwen ren

@Category
election_epoch can be used to record the change of the mon state, if has been election_epoch = 1, then you lose the effect

We have an application scenario: calamari need through election_epoch sense the change of mon

#5 - 10/28/2015 09:42 AM - huanwen ren

@Joao Luis

election_epoch can be used to record the change of the mon state, if has been election_epoch = 1, then you lose the effect

We have an application scenario: calamari need through election_epoch sense the change of mon

#6 - 11/11/2015 02:16 PM - Sage Weil

- Status changed from Verified to Resolved