RADOS - Backport #38850

upgrade: 1 nautilus mon + 1 luminous mon can't automatically form quorum

03/22/2019 10:19 AM - Tim Serong

Status:	Resolved	**Spent time:**	0.00 hour
Priority:	Immediate		
Assignee:	Joao Eduardo Luis		
Target version:	v14.2.2		
Release:	nautilus		

Description

Seen while upgrading Luminous (12.2.10) to Nautilus (14.2.0). Three mon hosts, four osd hosts. The process was:

- Shutdown mon1 (quorum in now mon2+mon3, both Luminous)
- Upgrade mon1 to Nautilus
- Start mon1 again. mon1 joins cluster, `ceph health` reports all three mons OK
- Shutdown mon2 (leaving mon1 = Nautilus and mon3 = Luminous)
- `ceph health` is now broken (eventually times out)
- mon1 logs repeat:

```
2019-03-22 09:50:21.634 7f31906d1700  1 mon.mon1@0(electing) e1  peer v1:172.16.1.13:6789/0 release  < min_mon_release, or missing features 0
```

- mon3 logs repeat:

```
2019-03-22 09:51:59.644672 7fa9f589a700 -1 mon.mon3@2(probing) e1 handle_probe missing features, have 4611087853745930235, required 0, missing 0
```

This means that the cluster is effectively down until you're able to complete the upgrade of mon2.

Curiously, on mon1 (Nautilus):

```
# ceph daemon mon.$(hostname) mon_status|grep min_mon_release
        "min_mon_release": 12,
        "min_mon_release_name": "luminous",
```

So why is it comlaining about release < min_mon_release?

Even more interesting, I can run this on the Luminous mon:

```
# ceph daemon mon.$(hostname) quorum enter
started responding to quorum, initiated new election
```

...and **bam** a few seconds later, we're in business again:

```
# ceph status
  cluster:
    id:     44e4a575-5c31-3c61-88c5-001ea49e8aaa
    health: HEALTH_WARN
            1/3 mons down, quorum mon1,mon3

  services:
    mon: 3 daemons, quorum mon1,mon3, out of quorum: mon2
    mgr: mon3(active), standbys: mon1
    osd: 30 osds: 30 up, 30 in

  data:
    pools:   1 pools, 512 pgs
    objects: 0 objects, 0B
    usage:   30.3GiB used, 567GiB / 597GiB avail
    pgs:     512 active+clean
```

Not that "quorum enter" doesn't help if run from the Nautilus mon, it only works when run from the Luminous mon.

History

#1 - 03/22/2019 10:22 AM - Ricardo Dias

- Assignee set to Joao Eduardo Luis

#2 - 03/22/2019 11:05 AM - Tim Serong

Just to clarify slightly -- I know the upgrade instructions in the Nautilus release announcement say to "upgrade monitors by installing the new packages and restarting the monitor daemons", but this quick way of upgrading is not always possible; depending on what distro you're using, you may have to upgrade the base OS before you can install the new Nautilus packages, which means that each node is going to be down for quite some time (at least several minutes, maybe many tens of minutes or longer).

#3 - 03/22/2019 11:07 AM - Joao Eduardo Luis

- Category set to Correctness/Safety

- Regression changed from No to Yes

#4 - 03/25/2019 10:07 AM - Lars Marowsky-Brée

Agreed, my expectation would be that we can maintain quorum during the entire upgrade period. Even discounting OS upgrades, restarts can go wrong, nodes can fail etc. Maintaining availability is crucial.

#5 - 04/04/2019 08:32 PM - Greg Farnum

This is super weird; the only other recent reference I see to min_mon_release is https://github.com/ceph/ceph/pull/27107 but that looks like just an output bug that we hit when going from Luminous to master/"Octopus" releases...

#6 - 05/16/2019 11:31 AM - Joao Eduardo Luis

I have been working on it, able to reproduce, just unable yet to pin down the cause.

Reproducing basically takes the following steps:

1. 3 monitors on luminous (a, b, c)
2. shutdown mon.a, let quorum form (2 luminous monitors, b and c)
3. upgrade mon.a to nautilus; quorum forms with a, b, and c.
4. shutdown mon.b; quorum is unable to form between mon.a and mon.c

I'm trying to figure out which paths are involved here, and why that's happening. Evidence points to feature mismatch, but unable to pinpoint why just yet.

#7 - 05/28/2019 09:41 AM - Joao Eduardo Luis

- Priority changed from Normal to Immediate

- Backport set to nautilus

backport PR to nautilus: https://github.com/ceph/ceph/pull/28262

#8 - 05/31/2019 08:19 PM - Yuri Weinstein

Joao Eduardo Luis wrote:

> backport PR to nautilus: https://github.com/ceph/ceph/pull/28262

merged

#9 - 06/01/2019 10:23 AM - Nathan Cutler

- Tracker changed from Bug to Backport

- Status changed from New to Resolved

- Target version set to v14.2.2

- Release set to nautilus