Hi,

while stopping all OSDs on a host I get for some seconds slow ops. Sometimes this don't happen and mostly it works as expected on stopping a single OSD.

I found the osd_fast_shutdown parameter which is per default true. In this case the OSD don't announce its shutdown to the mons before it stops.

If I stop all OSDs on a host (24 OSDs) I see the following several hundred times (of course with different OSD numbers) in my ceph.log: cluster [DBG] osd.317 reported immediately failed by osd.202

The first message like this is about 5 to 7 seconds after stopping the OSD. After down detection the cluster starts its peering and after this all is good. But the long time causes slow ops.

If i set the option osd_fast_shutdown to false I will see more or less immediately the following in the ceph.log (one message per OSD):

cluster [INF] osd.837 marked itself down

In this case the whole process of detection and peering is much faster and I don't get any slow ops.

I found this two PRs:
https://github.com/ceph/ceph/pull/31677
https://github.com/rook/rook/pull/4328

After reading the conversation I don't get it why the fast_shutdown shout be faster and better (related to the down detection) than so normal one. But I can understand that it is not necessary to stop all subsystems.

I wonder if it would make sense to send the shutdown message to the mons before stopping the OSD even in the fast shutdown process.

What do you think?

My Cluster:
Nautilus (v14.2.11)
44 Nodes
1056 OSDs
each node is connected via 2x10G LACP Channel

Manuel
Hi Manuel,

Would you be able to test a patch for this issue?

If so, what OS and ceph packages/version you run?

cheers,
Mauricio

Test steps, plus testing with vstart, run-make-check, and qa/run-standalone look good.

https://github.com/ceph/ceph/pull/38909

If the PR is merged, can this be backported to Octopus/Nautilus? (I can’t update the fields.) Thanks!

If Status changed from New to Fix Under Review
- Pull request ID set to 38909

If the PR is merged, can this be backported to Octopus/Nautilus? (I can’t update the fields.) Thanks!

Subject changed from nautilus: OSD: shutdown of a OSD Host causes slow requests to OSD: shutdown of a OSD Host causes slow requests
- Backport set to octopus, nautilus

Backport changed from octopus, nautilus to pacific, octopus, nautilus

Project changed from Ceph to RADOS
- Category deleted (OSD)

The master PR has been merged.

Can someone update Status to Pending Backport, please?

Thanks!

Status changed from Fix Under Review to Pending Backport
#10 - 03/09/2021 07:35 PM - Backport Bot
- Copied to Backport #49681: octopus: OSD: shutdown of a OSD Host causes slow requests added

#11 - 03/09/2021 07:35 PM - Backport Bot
- Copied to Backport #49682: nautilus: OSD: shutdown of a OSD Host causes slow requests added

#12 - 03/09/2021 07:35 PM - Backport Bot
- Copied to Backport #49683: pacific: OSD: shutdown of a OSD Host causes slow requests added

#13 - 03/09/2021 07:40 PM - Mauricio Oliveira
Igor, thanks.

I'd like to / can work on submitting the backport PRs, if that's OK.

In the future, if I want to open backport tracker issues with the script, is it possible to get access so not to get ForbiddenError?
(I tried slightly before it was done, and hit this.)

INFO:root:Processing issue list ->46978<-
INFO:root:Processing 1 issues with status Pending Backport
Traceback (most recent call last):
  File "./src/script/backport-create-issue", line 343, in <module>
    ...
redminelib.exceptions.ForbiddenError: Requested resource is forbidden

#14 - 03/10/2021 06:47 AM - Konstantin Shalygin
Mauricio, just make a backport PR at GitHub, we'll attach it to tracker later.

#15 - 03/10/2021 07:57 AM - Loïc Dachary
Hi Mauricio,

You are welcome to join the Stable Release team on IRC at #ceph-backports to discuss and resolve the issue you have with the script.

Cheers

#16 - 03/10/2021 01:08 PM - Mauricio Oliveira
Hey Konstantin and Loïc,

Understood; thanks!

#17 - 03/15/2021 10:42 PM - Sage Weil
- Status changed from Pending Backport to Resolved
Since this issue is resolved and only the pacific backport was done, I assume it means the octopus & nautilus backports are no longer needed and can be cancelled.

- Backport changed from pacific, octopus, nautilus to pacific

Hi @singuliere_,

Could you please revert the backport field to include Octopus and Nautilus?

Such backports have been done (see 'Related issues'), but are waiting reviews. Only Pacific has been merged at this time perhaps due to release timing/focus.

There's no other comment or discussion in the backport bugs or PRs that suggest these being no longer needed; so I'd like to ask for the backports to be kept.

Thank you!

@Mauricio, I was update issue backports and status.

Thanks, Konstantin!

While running with --resolve-parent, the script "backport-create-issue" noticed that all backports of this issue are in status "Resolved" or "Rejected".