Ceph - Feature #11692

online change of mon_osd_full_ratio and mon_osd_nearfull_ratio doesn't take effect.

05/21/2015 04:55 AM - cory gu

Status:	Resolved	**Start date:**	05/21/2015
Priority:	Normal	**Due date:**	
Assignee:	Kefu Chai	**% Done:**	0%
Category:	Monitor	**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:	Community (user)	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**Pull request ID:**	

Description

I want to verify the behavior of ceph cluster with full osd case, so I change the default full ratio with following cmds:

```
> ceph --admin-daemon /var/run/ceph/ceph-osd.0.asok config set  mon_osd_full_ratio 0.20
> ceph --admin-daemon /var/run/ceph/ceph-osd.0.asok config set  mon_osd_nearfull_ratio 0.15
>
```

Then I create big file in the associated disk. df -lh shows the disk is over 20%. I expected ceph cluster should give osd full warnning.
However, ceph -s doesn't show osd full warnnings.
The observation is online change of mon_osd_full_ratio and mon_osd_nearfull_ratio doesn't take effect.

currently we can change the runtime full_ratio settings using

```
ceph pg set_full_ratio 0.20
ceph pg set_nearfull_ratio  0.15
```

but this might confuse our user: as per http://docs.ceph.com/docs/master/rados/configuration/mon-config-ref/#storage-capacity, we are using

```
        mon osd full ratio = .80
        mon osd nearfull ratio = .70
```

for setting the full ratio, but adjusting them does not work if the PG is already created.

maybe we can let PGMonitor inherit from md_config_obs_t and watch the change of "mon osd full ratio" and "mon osd nearfull ratio", and add the new setting to the pending proposal.

Associated revisions

Revision 6a59aae0 - 12/30/2015 04:35 PM - Kefu Chai

config: complains when a setting is not tracked

- not all config items are tracked, so it does not take any effect after
  we sucessfully changed them using "ceph tell <daemon> injectargs --foo-bar 15',
  as shown by the command output:

$daemon: foo_bar = '15'
if foo-bar happens to be the one not tracked by any components in <daemon>.
in this fix, the message of
$daemon: foo_bar = '15' (unchangeable)
is returned instead. nevertheless, the config is still updated. as
"ceph daemon <daemon> config show | grep foo_bar" shows:
"foo_bar": "15"
this helps user to understand that the setting is not dynamically
changeable.

- update the test accordingly

Fixes: #11692
Signed-off-by: Kefu Chai <kchai@redhat.com>

History

#1 - 05/25/2015 06:58 AM - Kefu Chai

- Status changed from New to Verified

cory, you might want to try out

```
ceph pg set_full_ratio 0.20
ceph pg set_nearfull_ratio  0.15
```

but i am not sure why we are using different settings for changing the full_ratios at runtime, though.

#2 - 05/25/2015 06:58 AM - Kefu Chai

- Assignee set to Kefu Chai

#3 - 05/25/2015 08:28 AM - cory gu

Hi Kefu,
Thank you for your reply. Just tried your suggested input, and it works.
a few questions here:
So those settings are in pg level, does this mean all pools have the unified full ratio setting? we can't set the full raio pool by pool?

#4 - 05/26/2015 07:21 AM - Kefu Chai

cory gu wrote:

> a few questions here:
> So those settings are in pg level, does this mean all pools have the unified full ratio setting?

it's a cluster-wide setting. all pools share the same full ratio settings in the same cluster.

> we can't set the full raio pool by pool?

no, we can't. not at this moment. just out of curiosity, why would you want this to be a pool specific setting?

#5 - 05/27/2015 01:27 AM - Kefu Chai

Kefu Chai wrote:

> cory, you might want to try out
> [...]
>
> but i am not sure why we are using different settings for changing the full_ratios at runtime, though.

to identify the important settings which is supposed to be propagated to all OSDs in the cluster, we added pg set_{,near}full_ratio commands.

#6 - 05/27/2015 01:29 AM - Kefu Chai

- Status changed from Verified to Rejected

these options work as expected. so i am closing this issue. please feel free to reopen it if you think otherwise, thanks!

#7 - 07/23/2015 09:00 AM - Jan Schermer

Sorry for hijacking this issue, but IMO it should be possible to set different thresholds on each OSD, not cluster-wide.

For example if
1) the OSDs are not the same size - leaving 15% free on a 400GB OSD is 60GB and that's probably allright. But on a 1600GB OSD 15% translates to 240GB which is probably a bit too much wasted space.
2) while in a production cluster the OSDs should have the filesystems to themelves, in a lab environment or a PoC it is conceivable I'll for example run them on root filesystems of some VMs with other apps living alongside them - uniformity is a non-issue in this case and I'm forced to commit more resources because of this

#8 - 12/14/2015 07:37 AM - Kefu Chai

- Description updated

- Status changed from Rejected to Verified

#9 - 12/14/2015 07:40 AM - Kefu Chai

after a second thought: before the new setting applies to the new osdmap, and be sent to OSD nodes. it needs to be agreed by the quorum. so that might be why we made this more explicit.

#10 - 12/14/2015 08:58 PM - Greg Farnum

Yeah, I don't envision any good way to make it possible to change these values by updating the runtime config.

When thinking about alternatives, I think we could extend the config framework so that it can spit out more information when the user tries to inject a change. Then we could either mark these (and, perhaps, other non-dynamic changes) as unchangeable or at least return some kind of warning message to the user.

#11 - 12/23/2015 12:33 PM - Kefu Chai

Right, that's a good idea. if no observers is tracking a config, changing it at runtime would lead to a warning.

#12 - 12/23/2015 12:33 PM - Kefu Chai

- Tracker changed from Bug to Feature

#13 - 12/30/2015 04:38 PM - Kefu Chai

- Status changed from Verified to Need Review

https://github.com/ceph/ceph/pull/7085

#14 - 01/13/2016 02:26 AM - Sage Weil

- Status changed from Need Review to Resolved