Ceph-deploy - Bug #13833

ceph-deploy osd activate doesn't change ownership of journal partitions

11/19/2015 12:39 PM - David Riedl

Status:	Closed	**Start date:**	11/19/2015
Priority:	High	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			
Source:	other	**Severity:**	2 - major
Tags:	osd activate journal	**Reviewed:**	
Backport:		**Affected Versions:**	1.5.28
Regression:	No	**ceph-qa-suite:**	ceph-deploy

Description

The command
ceph-deploy osd activate ceph01:/dev/sda1:/dev/sdd1
fails with

```
[ceph01][WARNIN] INFO:ceph-disk:Running command: /usr/bin/ceph --cluster ceph --name client.bootst
rap-osd --keyring /var/lib/ceph/bootstrap-osd/ceph.keyring mon getmap -o /var/lib/ceph/tmp/mnt.pmH
Ruu/activate.monmap
[ceph01][WARNIN] 2015-11-19 11:22:53.974765 7f1a06852700  0 -- :/3225863658 >> 10.20.60.10:6789/0
pipe(0x7f19f8062590 sd=4 :0 s=1 pgs=0 cs=0 l=1 c=0x7f19f805c1b0).fault
[ceph01][WARNIN] got monmap epoch 16
[ceph01][WARNIN] INFO:ceph-disk:Running command: /usr/bin/ceph-osd --cluster ceph --mkfs --mkkey -
i 0 --monmap /var/lib/ceph/tmp/mnt.pmHRuu/activate.monmap --osd-data /var/lib/ceph/tmp/mnt.pmHRuu
--osd-journal /var/lib/ceph/tmp/mnt.pmHRuu/journal --osd-uuid de162e24-16b6-4796-b6b9-774fdb8ec234
 --keyring /var/lib/ceph/tmp/mnt.pmHRuu/keyring --setuser ceph --setgroup ceph
[ceph01][WARNIN] 2015-11-19 11:22:57.237096 7fb458bb7900 -1 filestore(/var/lib/ceph/tmp/mnt.pmHRuu
) mkjournal error creating journal on /var/lib/ceph/tmp/mnt.pmHRuu/journal: (13) Permission denied
[ceph01][WARNIN] 2015-11-19 11:22:57.237118 7fb458bb7900 -1 OSD::mkfs: ObjectStore::mkfs failed wi
th error -13
[ceph01][WARNIN] 2015-11-19 11:22:57.237157 7fb458bb7900 -1  ** ERROR: error creating empty object
 store in /var/lib/ceph/tmp/mnt.pmHRuu: (13) Permission denied
[ceph01][WARNIN] ERROR:ceph-disk:Failed to activate
[ceph01][WARNIN] DEBUG:ceph-disk:Unmounting /var/lib/ceph/tmp/mnt.pmHRuu
[ceph01][WARNIN] INFO:ceph-disk:Running command: /bin/umount -- /var/lib/ceph/tmp/mnt.pmHRuu
[ceph01][WARNIN] Traceback (most recent call last):
[ceph01][WARNIN]   File "/usr/sbin/ceph-disk", line 3576, in <module>
[ceph01][WARNIN]     main(sys.argv[1:])
[ceph01][WARNIN]   File "/usr/sbin/ceph-disk", line 3530, in main
[ceph01][WARNIN]     args.func(args)
[ceph01][WARNIN]   File "/usr/sbin/ceph-disk", line 2424, in main_activate
[ceph01][WARNIN]     dmcrypt_key_dir=args.dmcrypt_key_dir,
[ceph01][WARNIN]   File "/usr/sbin/ceph-disk", line 2197, in mount_activate
[ceph01][WARNIN]     (osd_id, cluster) = activate(path, activate_key_template, init)
[ceph01][WARNIN]   File "/usr/sbin/ceph-disk", line 2360, in activate
[ceph01][WARNIN]     keyring=keyring,
[ceph01][WARNIN]   File "/usr/sbin/ceph-disk", line 1950, in mkfs
[ceph01][WARNIN]     '--setgroup', get_ceph_user(),
[ceph01][WARNIN]   File "/usr/sbin/ceph-disk", line 349, in command_check_call
[ceph01][WARNIN]     return subprocess.check_call(arguments)
[ceph01][WARNIN]   File "/usr/lib64/python2.7/subprocess.py", line 542, in check_call
[ceph01][WARNIN]     raise CalledProcessError(retcode, cmd)
[ceph01][WARNIN] subprocess.CalledProcessError: Command '['/usr/bin/ceph-osd', '--cluster', 'ceph'
, '--mkfs', '--mkkey', '-i', '0', '--monmap', '/var/lib/ceph/tmp/mnt.pmHRuu/activate.monmap', '--o
sd-data', '/var/lib/ceph/tmp/mnt.pmHRuu', '--osd-journal', '/var/lib/ceph/tmp/mnt.pmHRuu/journal',
 '--osd-uuid', 'de162e24-16b6-4796-b6b9-774fdb8ec234', '--keyring', '/var/lib/ceph/tmp/mnt.pmHRuu/
keyring', '--setuser', 'ceph', '--setgroup', 'ceph']' returned non-zero exit status 1
[ceph01][ERROR ] RuntimeError: command returned non-zero exit status: 1
[ceph_deploy][ERROR ] RuntimeError: Failed to execute command: ceph-disk -v activate --mark-init s
ystemd --mount /dev/sda1
```

I fixed the issue by manually chown the ssd partitions for my journal with

chown ceph:ceph /dev/sdd1

After that the activate command worked flawlessly.

History

#1 - 11/26/2015 09:43 AM - Jonas Keidel

Is it possible to get a persistent fix? Running chown just fix that temporary.

#2 - 11/26/2015 10:01 AM - David Riedl

Jonas Keidel wrote:

> Is it possible to get a persistent fix? Running chown just fix that temporary.

I found a permanent solution and the underlying problem.

The underlying problem was, that the partitions I created by hand hadn't the right partition UID for CEPH partitions. The new UDEV rules, coming with the infernalis release also change the access rights for all partitions with these UIDs.

The permanent solution for that issue is to change the UID of the partitions with this command

sgdisk -t 1:45B0969E-9B03-4F30-B4C6-B4B80CEFF106 /dev/sdd

I still haven't upgraded to infernalis yet and haven't tested this yet.

I observed the following behavior of ceph-deploy though:
When I create a new OSD with a journal on a different disk, ceph-deploy prepare automatically creates both partitions with the right partition UID. But if I use ceph-deploy prepare on a blank OSD and an already created journal partition (SSD with several journal partitions), ceph-deploy doesn't change the UID of that particular partition.

I think ceph-deploy should automatically change the UID of the already created partition.

#3 - 11/26/2015 10:02 AM - David Riedl

PS: Thanks to the ceph-users mailing list. They helped me alot!

#4 - 11/30/2015 07:52 PM - Alfredo Deza

Does the same happen with `osd create` ? (vs. prepare and then activate)

#5 - 03/31/2016 06:31 PM - Nate Curry

David Riedl wrote:

> The permanent solution for that issue is to change the UID of the partitions with this command
>
> sgdisk -t 1:45B0969E-9B03-4F30-B4C6-B4B80CEFF106 /dev/sdd
>
> I still haven't upgraded to infernalis yet and haven't tested this yet.

I am having the same problem with Jewel and the temporary fix of changing ownership of the journal partitions for works for me as well. I just wanted to clarify that you set all of your journal disks in your cluster to that same UID: 45B0969E-9B03-4F30-B4C6-B4B80CEFF106

#6 - 05/22/2016 05:52 PM - Christian Sarrasin

Alfredo Deza wrote:

> Does the same happen with `osd create` ? (vs. prepare and then activate)

It does for me with:

ceph version 10.2.1 (3a66dd4f30852819c1bdaa8ec23c795d4ad77269)
ceph-deploy 1.5.32

Mind you I'm in a slightly different boat since my journal partitions are MBR rather than GPT so the above workaround doesn't work (and somehow neither does

```
chown ceph:disk /dev/vda2 /dev/vda3
```

(where my manually-created journal partitions live). Going to try and convert the whole setup to GPT (is this actually a requirement and if so, is it documented anywhere?)

Cheers!

#7 - 05/22/2016 09:09 PM - Nathan Cutler

@Nate

> I just wanted to clarify that you set all of your journal disks in your cluster to that same UID: 45B0969E-9B03-4F30-B4C6-B4B80CEFF106

Not the journal disks - just the journal partitions. Even if one journal takes up a whole disk, the disk should have a GPT partition table with a single partition in it. All journal partitions should have "partition entry type"/"Partition GUID code" 45b0969e-9b03-4f30-b4c6-b4b80ceff106.

Finding out the partition entry type is not particularly straightforward, but I've found two ways to do it:

(1) using blkid -o udev -p $JOURNAL_PARTITION_DEVICE

look for ID_PART_ENTRY_TYPE=45b0969e-9b03-4f30-b4c6-b4b80ceff106 in the output

(2) using sgdisk --info=$JOURNAL_PARTITION_NUMBER $JOURNAL_DISK

If, for example, the disk is /dev/sda and the journal is in the first partition:

sgdisk --info=1 /dev/sda

Look for "Partition GUID code" in the output. The value should be 45B0969E-9B03-4F30-B4C6-B4B80CEFF106

The udev rules in 95-ceph-osd.rules (on my system this file is in /lib/udev/rules.d) search for block devices with the relevant ID_PART_ENTRY_TYPE value (for OSD data and journal partitions) and run ceph-disk trigger on them, as well as changing the ownership of the relevant device files to ceph:ceph so the OSD daemons can access them.

@Christian: My guess is that the above mechanism will not work with MBR partition tables, since they do not have partition GUID codes - only single-byte partition types. You would be well-advised to convert your OSD data/journal disks to GPT and set the right Partition GUID codes.

#8 - 05/24/2016 07:27 PM - Christian Sarrasin

> Alfredo Deza wrote:
> Does the same happen with `osd create` ? (vs. prepare and then activate)

Further confirmed this happens with `osd create` and manually created GPT partitions.

ceph version 10.2.1 (3a66dd4f30852819c1bdaa8ec23c795d4ad77269) ceph-deploy 1.5.32

I didn't manage to get anywhere with journals on MBR partitions so it seems this isn't supported.

#9 - 07/12/2016 12:05 PM - Alfredo Deza

- Status changed from New to Closed

#10 - 10/21/2016 02:20 AM - kid hualing

but i meet this bug again :(

i deploy my ceph jewel cluster on 3 virtualbox vms by ceph-deploy.

i must 'chown ceph:ceph MY-CEPH-VOLUME' manually.

but it works for 2 vms, there still 1 vm does not work, always show me the same error message:

"ERROR: error creating empty object store in /var/lib/ceph/tmp/XXXXX: (13) Permission denied".

```
[root@b ceph-deploy-home]# ceph --version
ceph version 10.2.3 (ecc23778eb545d8dd55e2e4735b53cc93f92e65b)
```

#11 - 10/21/2016 10:18 AM - Nathan Cutler

@kid: Please check (and fix) the partion GUID codes of your data and journal partitions as described in http://tracker.ceph.com/issues/13833#note-7

Then the udev rules will set the permissions correctly at boot-time.

#12 - 06/27/2017 05:12 PM - Kyle Bader

This still happens when using ceph-disk against partitions created with parted:

```
[root@qcttwcoec27 ~]# ceph -v
ceph version 10.2.7-27.el7cp (e0d2d4f2fac9d95a26486121257255260bbec8d5)
[root@qcttwcoec27 ~]# rpm -qa ceph-common
ceph-common-10.2.7-27.el7cp.x86_64
```

Perhaps it should be fixed in ceph-disk?

#13 - 06/27/2017 08:12 PM - Nathan Cutler

@Kyle: can you reproduce it using ceph-disk alone, i.e. without ceph-deploy?

#14 - 06/27/2017 10:10 PM - Kyle Bader

That's how I encountered it and found this issue. I haven't attempted to reproduce it yet.

#15 - 12/20/2018 01:31 PM - Janek Bevendorff

I am having this or a similar issue with the current Docker image. I cannot for the life of me get my OSDs to activate. Changing permissions on partitions has no effect, since the Docker container changes them back the moment I start it. Monitors and managers start fine, I can create the OSDs, but the moment they are to be activated using the Docker image, I get

```
** ERROR: error creating empty object store in /var/lib/ceph/tmp/mnt.f2vcgU: (13) Permission denied
```

Any ideas? This issue appears to be pretty old.