ceph-volume - Bug #23337

ceph-volume lvm create fails to detect multipath devices

03/13/2018 05:06 PM - Alexander Bauer

Status:	Closed	**% Done:**	0%
Priority:	Normal		
Assignee:	Alfredo Deza		
Category:			
Target version:			
Source:		**Affected Versions:**	
Tags:		**ceph-qa-suite:**	
Backport:		**Pull request ID:**	
Regression:	No	**Crash signature (v1):**	
Severity:	3 - minor	**Crash signature (v2):**	
Reviewed:			

Description

My esteemed colleague and I are migrating our existing Ceph cluster to Luminous, and are rebuilding some of our OSDs to use BlueStore. We are using ceph-volume lvm create to re-create a recently destroyed OSD on a disk we access using multipath. The device is reported by lsblk as follows:

```
[root@ceph-disk-1 site-packages]# lsblk /dev/mapper/bay7
NAME MAJ:MIN RM  SIZE RO TYPE  MOUNTPOINT
bay7 253:5    0  7.3T  0 mpath
```

Note that the type is mpath, not disk. This creates the following output when trying to create a BlueStore-backed OSD using that storage.

```
[root@ceph-disk-1 ~]# ceph-volume lvm create --bluestore --data /dev/mapper/bay7 --block.db /dev/n
vme1n1p2 --osd-id 7
Running command: ceph-authtool --gen-print-key
Running command: ceph --cluster ceph --name client.bootstrap-osd --keyring /var/lib/ceph/bootstrap
-osd/ceph.keyring osd tree -f json
Running command: ceph --cluster ceph --name client.bootstrap-osd --keyring /var/lib/ceph/bootstrap
-osd/ceph.keyring -i - osd new 963eec50-5d44-44b0-9b7d-e9be45b44b52 7
--> Was unable to complete a new OSD, will rollback changes
--> OSD will be destroyed, keeping the ID because it was provided with --osd-id
Running command: ceph osd destroy osd.7 --yes-i-really-mean-it
 stderr: destroyed osd.7
-->  RuntimeError: Cannot use device (/dev/mapper/bay7). A vg/lv path or an existing device is nee
ded
```

Through some Python debugging, we were able to find that ceph-volume's util.is_device('/dev/mapper/bay7') disagrees that this mpath device is a usable device. On master as of this writing, the culpable code is here: https://github.com/ceph/ceph/blob/5fee2689c87dea23cab10fee2a758f4d9a5b4959/src/ceph-volume/ceph_volume/util/disk.py#L204-L206

Reproduced here:

```
# ... src/ceph-volume/util/disk.py line 204-206
    TYPE = lsblk(dev).get('TYPE')
    if TYPE:
        return TYPE == 'disk'
```

As reported by the lsblk function defined earlier in the same file, TYPE is mpath. This matches what lsblk reported earlier, but doesn't reveal that the mpath is actually fronting a disk. We were able to work around this issue for the time being by amending the above code to be

```
# ... PATCHED src/ceph-volume/util/disk.py line 204-206
    TYPE = lsblk(dev).get('TYPE')
    if TYPE:
        return TYPE == 'disk' or TYPE == 'mpath'
```

but I suspect this is likely not the ideal solution. After making this change, however, we re-ran our initial command, and it worked perfectly.

```
[ceph-deploy@ceph-disk-1 ~]$ sudo ceph-volume lvm create --bluestore --data /dev/mapper/bay7 --block.db /dev/nvme1n1p2 --osd-id 7
Running command: ceph-authtool --gen-print-key
Running command: ceph --cluster ceph --name client.bootstrap-osd --keyring /var/lib/ceph/bootstrap-osd/ceph.keyring osd tree -f json
Running command: ceph --cluster ceph --name client.bootstrap-osd --keyring /var/lib/ceph/bootstrap-osd/ceph.keyring -i - osd new e2b78628-53f5-4a78-870c-990d2063305a 7
Running command: vgcreate --force --yes ceph-22c44762-565e-46b8-89aa-d379b6d42b2c /dev/mapper/bay7
 stdout: Physical volume "/dev/mapper/bay7" successfully created.
 stdout: Volume group "ceph-22c44762-565e-46b8-89aa-d379b6d42b2c" successfully created
Running command: lvcreate --yes -l 100%FREE -n osd-block-e2b78628-53f5-4a78-870c-990d2063305a ceph-22c44762-565e-46b8-89aa-d379b6d42b2c
 stdout: Logical volume "osd-block-e2b78628-53f5-4a78-870c-990d2063305a" created.
Running command: ceph-authtool --gen-print-key
Running command: mount -t tmpfs tmpfs /var/lib/ceph/osd/ceph-7
Running command: chown -R ceph:ceph /dev/dm-18
Running command: ln -s /dev/ceph-22c44762-565e-46b8-89aa-d379b6d42b2c/osd-block-e2b78628-53f5-4a78-870c-990d2063305a /var/lib/ceph/osd/ceph-7/block
Running command: ceph --cluster ceph --name client.bootstrap-osd --keyring /var/lib/ceph/bootstrap-osd/ceph.keyring mon getmap -o /var/lib/ceph/osd/ceph-7/activate.monmap
 stderr: got monmap epoch 3
Running command: ceph-authtool /var/lib/ceph/osd/ceph-7/keyring --create-keyring --name osd.7 --add-key AQCl7Kda5665NRAAUBO30oCCHiSHrnJUYSDQiQ==
 stdout: creating /var/lib/ceph/osd/ceph-7/keyring
 stdout: added entity osd.7 auth auth(auid = 18446744073709551615 key=AQCl7Kda5665NRAAUBO30oCCHiSHrnJUYSDQiQ== with 0 caps)
Running command: chown -R ceph:ceph /var/lib/ceph/osd/ceph-7/keyring
Running command: chown -R ceph:ceph /var/lib/ceph/osd/ceph-7/
Running command: chown -R ceph:ceph /dev/nvme1n1p2
Running command: ceph-osd --cluster ceph --osd-objectstore bluestore --mkfs -i 7 --monmap /var/lib/ceph/osd/ceph-7/activate.monmap --keyfile - --bluestore-block-db-path /dev/nvme1n1p2 --osd-data /var/lib/ceph/osd/ceph-7/ --osd-uuid e2b78628-53f5-4a78-870c-990d2063305a --setuser ceph --setgroup ceph
--> ceph-volume lvm prepare successful for: /dev/mapper/bay7
Running command: ceph-bluestore-tool --cluster=ceph prime-osd-dir --dev /dev/ceph-22c44762-565e-46b8-89aa-d379b6d42b2c/osd-block-e2b78628-53f5-4a78-870c-990d2063305a --path /var/lib/ceph/osd/ceph-7
Running command: ln -snf /dev/ceph-22c44762-565e-46b8-89aa-d379b6d42b2c/osd-block-e2b78628-53f5-4a78-870c-990d2063305a /var/lib/ceph/osd/ceph-7/block
Running command: chown -R ceph:ceph /dev/dm-18
Running command: chown -R ceph:ceph /var/lib/ceph/osd/ceph-7
Running command: ln -snf /dev/nvme1n1p2 /var/lib/ceph/osd/ceph-7/block.db
Running command: chown -R ceph:ceph /dev/nvme1n1p2
Running command: systemctl enable ceph-volume@lvm-7-e2b78628-53f5-4a78-870c-990d2063305a
 stderr: Created symlink from /etc/systemd/system/multi-user.target.wants/ceph-volume@lvm-7-e2b78628-53f5-4a78-870c-990d2063305a.service to /usr/lib/systemd/system/ceph-volume@.service.
```

```
Running command: systemctl start ceph-osd@7
--> ceph-volume lvm activate successful for osd ID: 7
--> ceph-volume lvm activate successful for osd ID: 7
--> ceph-volume lvm create successful for: /dev/mapper/bay7
```

History

#1 - 03/13/2018 06:33 PM - Alfredo Deza

- Status changed from New to In Progress

- Assignee set to Alfredo Deza

There is no support for multipath devices by ceph-volume for the use case you are presenting. The idea here is to allow someone who doesn't care about LVM to take a whole device and make it
a single logical volume group that has a single logical volume.

It is a naive/easy approach to avoid dealing with LVM.

If you want to use multipath then you must provide ceph-volume with a logical volume created from that multipath. We don't test multipath specifically, but we should not
have any problems working with a logical volume coming from a multipath device.

It is not as easy as just allowing 'mpath' when parsing TYPE, as I explain in [0]:

```
Depending on the type of the multipath setup, if using an active/passive array as the underlying physical devi
ces, filters are required in lvm.conf to exclude the disks that are part of those underlying devices.

It is unfeasible for ceph-volume to understand what type of configuration is needed for LVM to be able to work
 in various different multipath scenarios. The functionality to create the LV for you is merely a (naive) conv
enience, anything that involves different settings or configuration must be provided by a config management sy
stem which can then provide VGs and LVs for ceph-volume to consume.
```

I am going to leave this open to address the documentation portion of it, because it should be clear that we will not support multipath to create a logical volume.

[0] https://github.com/ceph/ceph/pull/20296

#2 - 03/13/2018 06:42 PM - Alexander Bauer

Update: I've just noticed pull request https://github.com/ceph/ceph/pull/20296, which addresses this issue.

#3 - 03/14/2018 02:50 PM - Alfredo Deza

- Status changed from In Progress to Closed

As explained here, on the pull request for multipath support, and now on the ceph docs, we aren't going to allow working directly with a "raw" multipath device to create a logical volume.

If a logical volume is on top of a multipath device we should have no issues whatsoever (given correct configuration for LVM).

http://docs.ceph.com/docs/master/ceph-volume/lvm/prepare/#multipath-support

PR that documents multipath support https://github.com/ceph/ceph/pull/20878
merged commit 1cc5ad0 into master