devops - Bug #6701

ceph-deploy osd prepare on directory path fails: OSError: [Errno 18] Invalid cross-device link

10/31/2013 11:05 PM - Mark Kirkwood

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

Description

Ceph version is 0.71-234-g1f02d00 built from src on bunti 13.10.

The desired setup is osd data in /data2/cephdata journal on /dev/sda9

```
$ sudo ceph-deploy -v osd prepare zmori:/data2/cephdata:/dev/sda9
[ceph_deploy.cli][INFO  ] Invoked (1.2.7): /usr/bin/ceph-deploy -v osd prepare zmori:/data2/cephdata:/dev/sda9
[ceph_deploy.osd][DEBUG ] Preparing cluster ceph disks zmori:/data2/cephdata:/dev/sda9
[zmori][DEBUG ] connected to host: zmori
[zmori][DEBUG ] detect platform information from remote host
[zmori][DEBUG ] detect machine type
[ceph_deploy.osd][INFO  ] Distro info: Ubuntu 13.10 saucy
[ceph_deploy.osd][DEBUG ] Deploying osd to zmori
[zmori][DEBUG ] write cluster configuration to /etc/ceph/{cluster}.conf
[ceph_deploy.osd][ERROR ] OSError: [Errno 18] Invalid cross-device link
[ceph_deploy][ERROR ] GenericError: Failed to create 1 OSDs
```

The problem appears to be the data device path, as attempting to prepare an osd with just /data2/cephdata gives the same error.

I'm using ceph deploy from git, I note that a checkout from 2013-10-03 NZST does **not** have this issue. The 'Invalid cross-device link' started popping up around 2013-10-17 NZST, and is present in current master (2013-11-01 NZST)

History

#1 - 10/31/2013 11:08 PM - Mark Kirkwood

Omitted the probably significant fact that /data2 is a partition in a different disk from /var

#2 - 11/01/2013 01:17 AM - Mark Kirkwood

The particular issue is caused by os.rename in ceph_deploy/hosts/remotes.py line 54. replacing that with shutil.move seems to be the usual solution - however this brings to light another issue:

```
zmori][INFO  ] Running command: sudo udevadm trigger --subsystem-match=block --action=add
[ceph_deploy.osd][DEBUG ] Preparing host zmori disk /data2/cephdata/ journal None activate False
[zmori][INFO  ] Running command: sudo ceph-disk-prepare --fs-type xfs --cluster ceph -- /data2/cephdata/
[zmori][ERROR ] ceph-disk: Error: getting cluster uuid from configuration failed
[zmori][ERROR ] Traceback (most recent call last):
[zmori][ERROR ]   File "/home/markir/develop/python/ceph-deploy/ceph_deploy/lib/remoto/process.py", line 68, in run
[zmori][ERROR ]     reporting(conn, result, timeout)
[zmori][ERROR ]   File "/home/markir/develop/python/ceph-deploy/ceph_deploy/lib/remoto/log.py", line 13, in reporting
[zmori][ERROR ]     received = result.receive(timeout)
[zmori][ERROR ]   File "/home/markir/develop/python/ceph-deploy/ceph_deploy/lib/remoto/lib/execnet/gateway_base.py", line 455, in receive
[zmori][ERROR ]     raise self._getremoteerror() or EOFError()
[zmori][ERROR ] RemoteError: Traceback (most recent call last):
[zmori][ERROR ]   File "/home/markir/develop/python/ceph-deploy/ceph_deploy/lib/remoto/lib/execnet/gateway_base.py", line 806, in executetask
```

```
[zmori][ERROR ]     function(channel, **kwargs)
[zmori][ERROR ]   File "", line 35, in _remote_run
[zmori][ERROR ] RuntimeError: command returned non-zero exit status: 1
[zmori][ERROR ]
[zmori][ERROR ]
[ceph_deploy.osd][ERROR ] Failed to execute command: ceph-disk-prepare --fs-type xfs --cluster ceph -- /data2/cephdata/
[ceph_deploy][ERROR ] GenericError: Failed to create 1 OSDs
```

Looks like non-whole device setups are being broken here.

#3 - 11/02/2013 05:28 PM - Mark Kirkwood

Further on this (post the os.rename -> shutil.move), the next problem is:

```
[ERROR] ceph-disk: Error: getting cluster uuid from configuration failed
```

This is because the config file has been reduced to zero bytes, e.g: after mon create:

```
$ ls -l /etc/ceph/ceph.conf
-rw-r--r-- 1 root root 187 Nov  3 13:22 /etc/ceph/ceph.conf
```

after attempting osd prepare:

```
$ ls -l /etc/ceph/ceph.conf
-rw------- 1 root root 0 Nov  3 13:23 /etc/ceph/ceph.conf
```

I'll see if I can figure out why...

#4 - 11/02/2013 05:40 PM - Mark Kirkwood

I'm possibly causing the issue using shutil.move (can't see how mind you)...

#5 - 11/02/2013 06:25 PM - Mark Kirkwood

I now know why the original error is happening. My previous musings were not really on the mark (as it were):

consider the df output on the workstation:

```
Filesystem     1K-blocks     Used Available Use% Mounted on
/dev/sda3        3871400   320504   3334528   9% /
/dev/sda5       15616412    40648  14759432   1% /tmp
```

The ceph-deploy code circa version 1.3 is calling hosts/remotes.py:write_conf, which is:
- seeing if /etc/ceph/ceph.conf exists (it does)
- checking if it is different from ceph-deploy's conf (it is it seems [1])
- creates a temp file, writes the conf to that and renames the result to /etc/ceph/ceph.conf

That last step is failing because os.rename will not rename from /tmp to / (etc) i.e accross filesystems. As I said aboce, the usual fix for that is shutil.move - I am having trouble getting that to work (I'll investigate why).

As an minor aside, I'm wondering why, instead of the temp file stuff we don't just do:

```
*** remotes.py.orig    2013-11-03 14:25:19.589216186 +1300
--- remotes.py    2013-11-03 14:25:07.933515056 +1300
***************
*** 50,56 ****
              if old != conf and not overwrite:
                  raise RuntimeError(err_msg)
          tmp_file.write(conf)
!         os.rename(tmp_file.name, path)
          return
      if os.path.exists('/etc/ceph'):
          with open(path, 'w') as f:
--- 50,58 ----
              if old != conf and not overwrite:
                  raise RuntimeError(err_msg)
          tmp_file.write(conf)
!         with open(path, 'w') as fw:
!             fw.truncate()
!             fw.write(conf)
          return
      if os.path.exists('/etc/ceph'):
          with open(path, 'w') as f:
```

[1] I'm a little puzzled whay it is, since ceph-deploy is the only thing touching it at this point...

#6 - 11/02/2013 07:54 PM - Mark Kirkwood

Figured out what the issue with shutil.move was - needed to close the temp file before moving. Not an issue with os.rename as I think it is using a hard link under the covers. This presumably means shutil.move is not atomic..sigh. But her;s the woking patch anyway:

```
*** remotes.py.orig     2013-11-03 14:25:19.589216186 +1300
--- remotes.py  2013-11-03 15:37:52.654655203 +1300
***************
*** 1,6 ****
--- 1,7 ----
  import errno
  import socket
  import os
+ import shutil
  import tempfile
  import platform

***************
*** 50,56 ****
              if old != conf and not overwrite:
                  raise RuntimeError(err_msg)
          tmp_file.write(conf)
!         os.rename(tmp_file.name, path)
          return
      if os.path.exists('/etc/ceph'):
          with open(path, 'w') as f:
--- 51,58 ----
              if old != conf and not overwrite:
                  raise RuntimeError(err_msg)
          tmp_file.write(conf)
!         tmp_file.close()
!         shutil.move(tmp_file.name, path)
          return
      if os.path.exists('/etc/ceph'):
          with open(path, 'w') as f:
```

#7 - 11/04/2013 08:05 PM - Ian Colle

- Status changed from New to Fix Under Review

#8 - 11/04/2013 08:06 PM - Ian Colle

- Assignee set to Alfredo Deza

#9 - 11/05/2013 06:04 AM - Alfredo Deza

Thanks for the ticket and the resolution Mark!

Would you mind sending a pull request to https://github.com/ceph/ceph-deploy ? That way your contribution is saved there :)

Make sure you sign the commit with `-s` as well!

#10 - 11/05/2013 06:25 PM - Mark Kirkwood

Done.

#11 - 11/06/2013 05:31 AM - Alfredo Deza

- Status changed from Fix Under Review to Resolved

Pull Request opened: https://github.com/ceph/ceph-deploy/pull/126

And merged into ceph-deploy's master branch with hash: aeaaf11

#12 - 11/16/2013 07:35 PM - Mark Kirkwood

Managed to provoke this again, this time creating a keyring for an osd on a host that is not a monitor. The tiggering factor seems to be /tmp being a seperate filesystem. Experimental patch here that seems to fix it (shutil.move again)

https://github.com/markir9/ceph-deploy/commit/826433886a4f1215e1dcd07d57c13f43a2b12153

#13 - 11/19/2013 12:55 PM - Alfredo Deza

There was a PR addressing the problem for using shutil.move and I just opened another one to fix the missing `close()` call

https://github.com/ceph/ceph-deploy/pull/137/files

#14 - 11/19/2013 12:56 PM - Alfredo Deza

- Status changed from Resolved to Fix Under Review

#15 - 11/19/2013 03:59 PM - Mark Kirkwood

This will be fine for temporary files opened with 'delete=False' - if we start using delete=True then they will be possibly destroyed before we can copy them.

I did wonder if simply flushing the temp file before moving it might work, but only thought of that **after** doing the close + move patch!

#16 - 11/21/2013 12:33 PM - Alfredo Deza

`delete=True` is the default, and we are explicitly setting that flag to `delete=False` because of that reason.

Would it be reasonable to just add a comment that warns about this behavior and close this? Or do you propose something else?

#17 - 11/21/2013 01:39 PM - Mark Kirkwood

Yeah, I think a note is fine.

#18 - 11/26/2013 07:54 AM - Alfredo Deza

- Status changed from Fix Under Review to Resolved

#19 - 11/26/2013 08:23 AM - Alfredo Deza

Added a comment to that function, Hash: 2d9c452332d51f550abb2a189c1a3621a20c504a