Ceph - Bug #1094

"mkcephfs -c /etc/ceph.conf --allhosts --mkbtrfs" finds /tmp/mkcephfs.**** directory no longer exists

05/17/2011 11:24 AM - shyamali mukherjee

Status:	Resolved	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:	Sage Weil		
Category:			
Target version:	v0.29		
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

I have used ceph0.23 for quite sometime. But now after a fresh install and build of ceph 0.27.1
I see that during "mkcephfs" init /tmp/mkcephfs.**** gets deleted intermittently. There is no apparent error.

Log:

=== osd.50 ===
pushing conf and monmap to bzt10
umount: /dev/sdc3: not mounted

WARNING! - Btrfs Btrfs v0.19 IS EXPERIMENTAL
WARNING! - see http://btrfs.wiki.kernel.org before using

fs created label (null) on /dev/sdc3
nodesize 4096 leafsize 4096 sectorsize 4096 size 281.11GB
Btrfs Btrfs v0.19
Scanning for Btrfs filesystems *** WARNING: Ceph is still under heavy development, and is only suitable for * * testing and review. Do not trust it with important data. *** ** WARNING: 'filestore btrfs snap' is enabled (for safe transactions, rollback), but btrfs does not support the SNAP_CREATE_V2 ioctl
(added in Linux 2.6.37). Expect slow btrfs sync/commit
performance.
2011-05-17 11:19:39.235658 7fce764ce6f0 created object store /data/osd50 journal /data/osd50/journal for osd50 fsid 41cddb12-6412-8a82-29f6-7d7df26fab21
creating private key for osd.50 keyring /tmp/mkcephfs.tRHuFp9666/keyring.osd.50
creating /tmp/mkcephfs.tRHuFp9666/keyring.osd.50
collecting osd.50 key === osd.51 ===
pushing conf and monmap to bzt10
umount: /dev/sdd3: not mounted

WARNING! - Btrfs Btrfs v0.19 IS EXPERIMENTAL
WARNING! - see http://btrfs.wiki.kernel.org before using

fs created label (null) on /dev/sdd3
nodesize 4096 leafsize 4096 sectorsize 4096 size 281.11GB
Btrfs Btrfs v0.19
Scanning for Btrfs filesystems *** WARNING: Ceph is still under heavy development, and is only suitable for * * testing and review. Do not trust it with important data. *** ** WARNING: 'filestore btrfs snap' is enabled (for safe transactions, rollback), but btrfs does not support the SNAP_CREATE_V2 ioctl
(added in Linux 2.6.37). Expect slow btrfs sync/commit
performance.
2011-05-17 11:22:18.164137 7f1a031586f0 created object store /data/osd51 journal /data/osd51/journal for osd51 fsid 41cddb12-6412-8a82-29f6-7d7df26fab21
creating private key for osd.51 keyring /tmp/mkcephfs.tRHuFp9666/keyring.osd.51
creating /tmp/mkcephfs.tRHuFp9666/keyring.osd.51

collecting osd.51 key

...
...

It has sucessfully created all 54 osds on 9 differnt node prior to failure below:

=== mds.0 ===
pushing conf and monmap to bzt1
creating private key for mds.0 keyring /tmp/mkcephfs.foFYOU2693/keyring.mds.0
creating /tmp/mkcephfs.foFYOU2693/keyring.mds.0
collecting mds.0 key
Building generic osdmap
2011-05-16 17:24:33.513745 7faea71146f0 common_init: unable to open config file.
highest numbered osd in /tmp/mkcephfs.foFYOU2693/conf is osd.
2011-05-16 17:24:33.519108 7f3ffa27c6f0 common_init: unable to open config file.
num osd = 1
/usr/local/bin/osdmaptool: osdmap file '/tmp/mkcephfs.foFYOU2693/osdmap'
/usr/local/bin/osdmaptool: writing epoch 1 to /tmp/mkcephfs.foFYOU2693/osdmap
bufferlist::write_file(/tmp/mkcephfs.foFYOU2693/osdmap): failed to open file: error 2: No such file or directory
osdmaptool: error writing to '/tmp/mkcephfs.foFYOU2693/osdmap': error 2: No such file or directory
rm: cannot remove `/tmp/mkcephfs.foFYOU2693/*': No such file or directory

It seems that directory and files were created moment ago.. but no longer there.

Sample config file:

```
; global
[global]
; enable secure authentication

;auth supported = cephx
        max open files = 131072
        log file = /scratch/ceph_bup/ceph.log



; monitors
;  You need at least one.  You need at least three if you want to
;  tolerate any node failures.  Always create an odd number.
[mon]
mon data = /scratch/ceph_bup/mon$id

; some minimal logging (just message traffic) to aid debugging
        debug ms = 1
        debug mon = 20



[mon.0]
host = bzt1
mon addr = 192.168.2.101:6789
; mds
;  You need at least one.  Define two to get a standby.
[mds]
; where the mds keeps it's secret encryption keys
debug ms = 1
debug mds = 20

[mds.0]
host = bzt1

; osd
;  You need at least one.  Two if you want data to be replicated.
;  Define as many as you like.
[osd]
```

```
debug ms = 1
debug osd = 10
debug journal = 20
debug filestore = 20
osd journal = /data/osd$id/journal
osd journal size = 1000 ; journal size, in megabytes
; This is where the btrfs volume will be mounted.
osd data = /data/osd$id

osd.0]
host = bzt2
btrfs devs = /dev/sda3
[osd.1]
host = bzt2
btrfs devs = /dev/sdb3
[osd.2]
host = bzt2
btrfs devs = /dev/sdc3

[osd.3]
host = bzt2
btrfs devs = /dev/sdd3
[osd.4]
host = bzt2
btrfs devs = /dev/sde3
[osd.5]
host = bzt2
btrfs devs = /dev/sdf3

; next host bz3

[osd.6]
host = bzt3
btrfs devs = /dev/sda3
[osd.7]
host = bzt3
btrfs devs = /dev/sdb3
[osd.8]
host = bzt3
btrfs devs = /dev/sdc3

[osd.9]
host = bzt3
btrfs devs = /dev/sdd3
[osd.10]
host = bzt3
btrfs devs = /dev/sde3
[osd.11]
host = bzt3
```

Associated revisions

Revision ecb7c961 - 05/19/2011 04:29 PM - Sage Weil

mkcephfs: pick rdir based on whether current daemon is local or not

We need to pick $rdir as local or remote inside the for name loop.

Fixes: #1094
Signed-off-by: Sage Weil <sage@newdream.net>

History

#1 - 05/17/2011 12:35 PM - Sage Weil

can you run mkcephfs with -x (bash -x mkcephfs <regular args>) so we can tell exactly what it's doing?

#2 - 05/17/2011 12:36 PM - Sage Weil

- Status changed from New to 4

#3 - 05/17/2011 02:55 PM - shyamali mukherjee

It is happening due to check_host fails to identify this host as "localhost"

Here is what happened:

On a different node sequence of events:

```
+ for name in '$what'
+ echo osd.53
+ cut -c 1-3
+ type=osd
+ echo osd.53
+ cut -c 4-
+ sed 's/^\.//'
id=53
+ num=53
+ name=osd.53
+ check_host
+ ./cconf -c /etc/ceph.conf -n osd.53 host
host=bzt10
+ '[' bzt10 = localhost ']'
+ ssh=
+ rootssh=
+ sshdir=/usr/src/ceph/latest/ceph-0.26/src
+ get_conf user '' user
+ var=user
+ def=
+ key=user
+ shift
+ shift
+ shift
+ '[' -z '' ']'
+ '[' 0 -eq 1 ']'
+ ./cconf -c /etc/ceph.conf -n osd.53 user
+ eval echo -n ''
++ echo -n
eval 'user="""
+ user=
'[' -n bzt10 ']'
+ '[' bzt10 '!=' bz1 ']'
+ '[' 1 -eq 0 ']'
+ '[' -z '' ']'
+ ssh='ssh bzt10'
+ rootssh='ssh root@bzt10'
+ get_conf sshdir /usr/src/ceph/latest/ceph-0.26/src 'ssh path'
+ var=sshdir
+ def=/usr/src/ceph/latest/ceph-0.26/src
+ key='ssh path'
+ shift
+ shift
+ shift
+ '[' -z '' ']'
+ '[' 0 -eq 1 ']'
+ ./cconf -c /etc/ceph.conf -n osd.53 'ssh path'
+ eval echo -n /usr/src/ceph/latest/ceph-0.26/src
++ echo -n /usr/src/ceph/latest/ceph-0.26/src
eval 'sshdir="/usr/src/ceph/latest/ceph-0.26/src"'
eval 'sshdir="/usr/src/ceph/latest/ceph-0.26/src"'
+ sshdir=/usr/src/ceph/latest/ceph-0.26/src
echo '=== osd.53 === ' === osd.53 ===
+ return 0
+ '[' -n 'ssh bzt10' ']'
+ echo pushing conf and monmap to bzt10
pushing conf and monmap to bzt10
+ do_cmd 'mkdir -p /tmp/mkcephfs.ERNtP17746'
+ '[' -z 'ssh bzt10' ']'
```

+ '[' 0 -eq 1 ']'
+ ssh bzt10 'cd /usr/src/ceph/latest/ceph-0.26/src ; ulimit -c unlimited ; mkdir -p /tmp/mkcephfs.ERNtP17746'
+ scp -q /tmp/mkcephfs.ERNtP17746/conf bzt10:/tmp/mkcephfs.ERNtP17746
+ scp -q /tmp/mkcephfs.ERNtP17746/monmap bzt10:/tmp/mkcephfs.ERNtP17746
+ '[' 1 -eq 1 ']'
+ '[' osd = osd ']'
+ do_root_cmd 'mkcephfs -d /tmp/mkcephfs.ERNtP17746 --prepare-osdfs osd.53'
+ '[' -z 'ssh bzt10' ']'
+ '[' 0 -eq 1 ']'
+ ssh root@bzt10 'cd /usr/src/ceph/latest/ceph-0.26/src ; ulimit -c unlimited ; mkcephfs -d /tmp/mkcephfs.ERNtP17746 --prepare-osdfs osd.53'
umount: /dev/sdf3: not mounted

WARNING! - Btrfs Btrfs v0.19 IS EXPERIMENTAL
WARNING! - see http://btrfs.wiki.kernel.org before using

fs created label (null) on /dev/sdf3
nodesize 4096 leafsize 4096 sectorsize 4096 size 281.11GB
Btrfs Btrfs v0.19
Scanning for Btrfs filesystems
+ do_cmd 'mkcephfs -d /tmp/mkcephfs.ERNtP17746 --init-daemon osd.53'
+ '[' -z 'ssh bzt10' ']'
+ '[' 0 -eq 1 ']'
+ ssh bzt10 'cd /usr/src/ceph/latest/ceph-0.26/src ; ulimit -c unlimited ; mkcephfs -d /tmp/mkcephfs.ERNtP17746 --init-daemon osd.53' * **WARNING: Ceph is still under heavy development, and is only suitable for * * testing and review. Do not trust it with important data. * **
WARNING: 'filestore btrfs snap' is enabled (for safe transactions,
rollback), but btrfs does not support the SNAP_CREATE_V2 ioctl
(added in Linux 2.6.37). Expect slow btrfs sync/commit
performance.
2011-05-17 14:23:33.567023 7f890959d6f0 created object store /data/osd53 journal /data/osd53/journal for osd53 fsid e340c4e9-6ae1-6224-654d-c513bc333df0
creating private key for osd.53 keyring /tmp/mkcephfs.ERNtP17746/keyring.osd.53
creating /tmp/mkcephfs.ERNtP17746/keyring.osd.53
+ '[' -n 'ssh bzt10' ']'
+ echo collecting osd.53 key
collecting osd.53 key
+ scp -q bzt10:/tmp/mkcephfs.ERNtP17746/key.osd.53 /tmp/mkcephfs.ERNtP17746
+ do_cmd 'rm -r /tmp/mkcephfs.ERNtP17746'
+ '[' -z 'ssh bzt10' ']'
+ '[' 0 -eq 1 ']'
+ ssh bzt10 'cd /usr/src/ceph/latest/ceph-0.26/src ; ulimit -c unlimited ; rm -r /tmp/mkcephfs.ERNtP17746'

++++++++++++++++ Now on my local node I could create first OSD but then due to above "rm -r" +++ we fail to get conf file second time === osd.54 ===
+ return 0
+ '[' -n 'ssh bzt1' ']'
+ echo pushing conf and monmap to bzt1
pushing conf and monmap to bzt1
pushing conf and monmap to bzt1
+ do_cmd 'mkdir -p /tmp/mkcephfs.ERNtP17746'
+ '[' -z 'ssh bzt1' ']'
+ '[' 0 -eq 1 ']'
+ ssh bzt1 'cd /usr/src/ceph/latest/ceph-0.26/src ; ulimit -c unlimited ; mkdir -p /tmp/mkcephfs.ERNtP17746'
+ scp -q /tmp/mkcephfs.ERNtP17746/conf bzt1:/tmp/mkcephfs.ERNtP17746
+ scp -q /tmp/mkcephfs.ERNtP17746/monmap bzt1:/tmp/mkcephfs.ERNtP17746
+ '[' 1 -eq 1 ']'
+ '[' osd = osd ']'
+ do_root_cmd 'mkcephfs -d /tmp/mkcephfs.ERNtP17746 --prepare-osdfs osd.54'
+ '[' -z 'ssh bzt1' ']'
+ '[' 0 -eq 1 ']'
+ ssh root@bzt1 'cd /usr/src/ceph/latest/ceph-0.26/src ; ulimit -c unlimited ; mkcephfs -d /tmp/mkcephfs.ERNtP17746 --prepare-osdfs osd.54'
umount: /dev/sda3: not mounted

- shyamali **********
  + for name in '$what'
  + echo osd.54
  + cut ~~-c 1-3~~
  ~~+ type=osd~~
  ~~+ echo osd.54~~
  ~~+ cut -c 4~~
  + sed 's/^\.//'
  id=54
  + num=54
  + name=osd.54
  + check_host
  + ./cconf -c /etc/ceph.conf -n osd.54 host

```
host=bzt1
+ '[' bzt1 = localhost ']'
+ ssh=
+ rootssh=
+ sshdir=/usr/src/ceph/latest/ceph-0.26/src
+ get_conf user '' user
+ var=user
+ def=
+ key=user
+ shift
+ shift
+ shift
+ '[' -z '' ']'
+ '[' 0 -eq 1 ']'
+ ./cconf -c /etc/ceph.conf -n osd.54 user
+ eval echo -n ''
++ echo -n
eval 'user=""'
+ user=
'[' -n bzt1 ']'
+ '[' bzt1 '!=' bz1 ']'
+ '[' 1 -eq 0 ']'
+ '[' -z '' ']'
+ ssh='ssh bzt1'
+ rootssh='ssh root@bzt1'
+ get_conf sshdir /usr/src/ceph/latest/ceph-0.26/src 'ssh path'
+ var=sshdir
+ def=/usr/src/ceph/latest/ceph-0.26/src
+ key='ssh path'
+ shift
+ shift
+ shift
+ '[' -z '' ']'
+ '[' 0 -eq 1 ']'
+ ./cconf -c /etc/ceph.conf -n osd.54 'ssh path'
+ eval echo -n /usr/src/ceph/latest/ceph-0.26/src
++ echo -n /usr/src/ceph/latest/ceph-0.26/src
eval 'sshdir="/usr/src/ceph/latest/ceph-0.26/src"'
+ sshdir=/usr/src/ceph/latest/ceph-0.26/src
echo '=== osd.54 === '

WARNING! - Btrfs Btrfs v0.19 IS EXPERIMENTAL
WARNING! - see http://btrfs.wiki.kernel.org before using

fs created label (null) on /dev/sda3
nodesize 4096 leafsize 4096 sectorsize 4096 size 281.11GB
Btrfs Btrfs v0.19
Scanning for Btrfs filesystems
failed to read /dev/hdb
failed to read /dev/fd0u800
failed to read /dev/fd0
+ do_cmd 'mkcephfs -d /tmp/mkcephfs.ERNtP17746 --init-daemon osd.54'
+ '[' -z 'ssh bzt1' ']'
+ '[' 0 -eq 1 ']'
+ ssh bzt1 'cd /usr/src/ceph/latest/ceph-0.26/src ; ulimit -c unlimited ; mkcephfs -d /tmp/mkcephfs.ERNtP17746 --init-daemon osd.54'  * WARNING: Ceph is still under heavy development, and is only suitable for *  *  testing and review.  Do not trust it with important data.  *  **
WARNING: 'filestore btrfs snap' is enabled (for safe transactions,
rollback), but btrfs does not support the SNAP_CREATE_V2 ioctl
(added in Linux 2.6.37).  Expect slow btrfs sync/commit
performance.
2011-05-17 14:23:35.913476 7f27c4d4e6f0 created object store /data/osd54 journal /data/osd54/journal for osd54 fsid
e340c4e9-6ae1-6224-654d-c513bc333df0
creating private key for osd.54 keyring /tmp/mkcephfs.ERNtP17746/keyring.osd.54
creating /tmp/mkcephfs.ERNtP17746/keyring.osd.54
+ '[' -n 'ssh bzt1' ']'
+ echo collecting osd.54 key
collecting osd.54 key
+ scp -q bzt1:/tmp/mkcephfs.ERNtP17746/key.osd.54 /tmp/mkcephfs.ERNtP17746
+ do_cmd 'rm -r /tmp/mkcephfs.ERNtP17746'
+ '[' -z 'ssh bzt1' ']'
+ '[' 0 -eq 1 ']'
+ ssh bzt1 'cd /usr/src/ceph/latest/ceph-0.26/src ; ulimit -c unlimited ; rm -r /tmp/mkcephfs.ERNtP17746'
```

- First OSD created sucessfully but then we did rm -rf ***** so osd.55 will fail

```
+ for name in '$what'
+ echo osd.55
+ cut -c 1-3
+ type=osd
+ echo osd.55
+ cut -c 4-
+ sed 's/^\.//'
id=55
+ num=55
+ name=osd.55

+ num=55
+ name=osd.55
+ check_host
+ ./cconf -c /etc/ceph.conf -n osd.55 host
host=bzt1
+ '[' bzt1 = localhost ']'
+ ssh=
+ rootssh=
+ sshdir=/usr/src/ceph/latest/ceph-0.26/src
+ get_conf user '' user
+ var=user
+ def=
+ key=user
+ shift
+ shift
+ shift
+ '[' -z '' ']'
+ '[' 0 -eq 1 ']'
+ ./cconf -c /etc/ceph.conf -n osd.55 user
+ eval echo -n ''
++ echo -n
eval 'user=""'
+ user=
'[' -n bzt1 ']'
+ '[' bzt1 '!=' bz1 ']'
+ '[' 1 -eq 0 ']'
+ '[' -z '' ']'
+ ssh='ssh bzt1'
+ rootssh='ssh root@bzt1'
+ get_conf sshdir /usr/src/ceph/latest/ceph-0.26/src 'ssh path'
+ var=sshdir
+ def=/usr/src/ceph/latest/ceph-0.26/src
+ key='ssh path'
+ shift
+ shift
+ shift
+ '[' -z '' ']'
+ '[' 0 -eq 1 ']'
+ ./cconf -c /etc/ceph.conf -n osd.55 'ssh path'
+ eval echo -n /usr/src/ceph/latest/ceph-0.26/src
++ echo -n /usr/src/ceph/latest/ceph-0.26/src
eval 'sshdir="/usr/src/ceph/latest/ceph-0.26/src"'
+ sshdir=/usr/src/ceph/latest/ceph-0.26/src
echo '=== osd.55 === ' === osd.55 ===
+ return 0
+ '[' -n 'ssh bzt1' ']'
+ echo pushing conf and monmap to bzt1
pushing conf and monmap to bzt1
+ do_cmd 'mkdir -p /tmp/mkcephfs.ERNtP17746'
+ '[' -z 'ssh bzt1' ']'
+ '[' 0 -eq 1 ']'
+ ssh bzt1 'cd /usr/src/ceph/latest/ceph-0.26/src ; ulimit -c unlimited ; mkdir -p /tmp/mkcephfs.ERNtP17746'
+ scp -q /tmp/mkcephfs.ERNtP17746/conf bzt1:/tmp/mkcephfs.ERNtP17746
/tmp/mkcephfs.ERNtP17746/conf: No such file or directory
+ rm '/tmp/mkcephfs.ERNtP17746/*'
rm: cannot remove `/tmp/mkcephfs.ERNtP17746/*': No such file or directory

====== I will look at check_host script  ======
```

#4 - 05/18/2011 02:36 PM - Sage Weil

- Assignee set to Sage Weil

- Target version set to v0.29

I see the problem. Can you please test commit:0efd51dede578e2cc8c68e1a55d1468a06eef83e (the wip-mkcephfsb branch) and see if that fixes things?

Thanks!

#5 - 05/19/2011 09:35 AM - shyamali mukherjee

Sage,
Thanks! There is one more thing I had to change. But I see that it is fixed in your latest code.
maxosd=`$CCONF ~~c $conf -l osd | grep -v ^osd\$ | cut -c 5~~ | sort -n | tail -1`
echo " highest numbered osd in $conf is osd.$maxosd"

Previously it was "cut 4-" and it failed.

#6 - 05/19/2011 09:39 AM - Sage Weil

- Status changed from 4 to Resolved

Thanks for testing!

#7 - 05/19/2011 10:12 AM - Sage Weil

- translation missing: en.field_story_points set to 1

- translation missing: en.field_position set to 1

- translation missing: en.field_position changed from 1 to 673