Ceph - Bug #20765

bluestore: mismatched uuid in bdev_label after unclean shutdown

07/25/2017 11:40 AM - Tomasz Torcz

Status:	Can't reproduce	**% Done:**	0%
Priority:	Normal	**Spent time:**	0.00 hour
Assignee:			
Category:	OSD		
Target version:	v12.1.0		
Source:		**Reviewed:**	
Tags:		**Affected Versions:**	v12.1.0
Backport:		**ceph-qa-suite:**	
Regression:	No	**Pull request ID:**	
Severity:	3 - minor	**Crash signature:**	

Description

Few hours after creation, one of my Bluestore OSDs was killed by OOM. It won't start now. This is on luminous RC, with debug osd = 20
debug bluestore = 20
debug bluefs = 20
debug rocksdb = 5
:

2017-07-25 13:25:25.600040 7f2b7fe2bd00 0 ceph version 12.1.1 (f3e663a190bf2ed12c7e3cda288b9a159572c800) luminous (rc), process (unknown), pid 8911
2017-07-25 13:25:25.600163 7f2b7fe2bd00 5 object store type is bluestore
2017-07-25 13:25:25.600563 7f2b7fe2bd00 10 bluestore(/var/lib/ceph/osd/ceph-1) set_cache_shards 1
2017-07-25 13:25:25.618159 7f2b7fe2bd00 0 pidfile_write: ignore empty --pid-file
2017-07-25 13:25:25.622461 7f2b7fe2bd00 10 bluestore(/var/lib/ceph/osd/ceph-1) _set_csum csum_type crc32c
2017-07-25 13:25:25.633659 7f2b7fe2bd00 0 load: jerasure load: lrc load: isa
2017-07-25 13:25:25.633868 7f2b7fe2bd00 1 bdev create path /var/lib/ceph/osd/ceph-1/block type kernel
2017-07-25 13:25:25.633907 7f2b7fe2bd00 1 bdev(0x3076b12900 /var/lib/ceph/osd/ceph-1/block) open path /var/lib/ceph/osd/ceph-1/block
2017-07-25 13:25:25.634890 7f2b7fe2bd00 1 bdev(0x3076b12900 /var/lib/ceph/osd/ceph-1/block) open size 319967006720 (0x4a7f851000, 297 GB) block_size 4096 (4096 B) rotational
2017-07-25 13:25:25.634909 7f2b7fe2bd00 10 bluestore(/var/lib/ceph/osd/ceph-1/block) _read_bdev_label
2017-07-25 13:25:26.075750 7f2b7fe2bd00 10 bluestore(/var/lib/ceph/osd/ceph-1/block) _read_bdev_label got bdev(osd_uuid dd3e6c5e-da75-433b-b619-25dc9cc031a4 size 0x4a7f851000 btime 2017-07-24 18:50:28.542135 desc main)
2017-07-25 13:25:26.075837 7f2b7fe2bd00 -1 bluestore(/var/lib/ceph/osd/ceph-1/block) _check_or_set_bdev_label bdev /var/lib/ceph/osd/ceph-1/block fsid dd3e6c5e-da75-433b-b619-25dc9cc031a4 does not match our fsid 85177bde-19e4-497a-b824-d3425a161264
2017-07-25 13:25:26.075845 7f2b7fe2bd00 1 bdev(0x3076b12900 /var/lib/ceph/osd/ceph-1/block) close
2017-07-25 13:25:26.150037 7f2b7fe2bd00 2 osd.1 0 init /var/lib/ceph/osd/ceph-1 (looks like hdd)
2017-07-25 13:25:26.150063 7f2b7fe2bd00 10 bluestore(/var/lib/ceph/osd/ceph-1) set_cache_shards 5
2017-07-25 13:25:26.150083 7f2b7fe2bd00 1 bluestore(/var/lib/ceph/osd/ceph-1) _mount path /var/lib/ceph/osd/ceph-1
2017-07-25 13:25:26.150217 7f2b7fe2bd00 1 bdev create path /var/lib/ceph/osd/ceph-1/block type kernel
2017-07-25 13:25:26.150225 7f2b7fe2bd00 1 bdev(0x3076b12b40 /var/lib/ceph/osd/ceph-1/block) open path /var/lib/ceph/osd/ceph-1/block
2017-07-25 13:25:26.152227 7f2b7fe2bd00 1 bdev(0x3076b12b40 /var/lib/ceph/osd/ceph-1/block) open size 319967006720 (0x4a7f851000, 297 GB) block_size 4096 (4096 B) rotational
2017-07-25 13:25:26.152254 7f2b7fe2bd00 10 bluestore(/var/lib/ceph/osd/ceph-1/block) _read_bdev_label
2017-07-25 13:25:26.154030 7f2b7fe2bd00 10 bluestore(/var/lib/ceph/osd/ceph-1/block) _read_bdev_label got bdev(osd_uuid dd3e6c5e-da75-433b-b619-25dc9cc031a4 size 0x4a7f851000 btime 2017-07-24 18:50:28.542135 desc main)
2017-07-25 13:25:26.154061 7f2b7fe2bd00 -1 bluestore(/var/lib/ceph/osd/ceph-1/block) _check_or_set_bdev_label bdev /var/lib/ceph/osd/ceph-1/block fsid dd3e6c5e-da75-433b-b619-25dc9cc031a4 does not match our fsid 85177bde-19e4-497a-b824-d3425a161264
2017-07-25 13:25:26.154066 7f2b7fe2bd00 1 bdev(0x3076b12b40 /var/lib/ceph/osd/ceph-1/block) close
2017-07-25 13:25:26.415124 7f2b7fe2bd00 -1 osd.1 0 OSD:init: unable to mount object store
2017-07-25 13:25:26.415165 7f2b7fe2bd00 -1 ESC[0;31m ** ERROR: osd init failed: (5) Input/output errorESC[0m

Disk is available:

1. blkid /var/lib/ceph/osd/ceph-1/block
   /var/lib/ceph/osd/ceph-1/block: PARTLABEL="ceph block" PARTUUID="973dff0a-63c6-436d-b72e-72af0d4453b7"

I saw such errors previously on 10.x - after ungraceful reboots, bluestore OSD weren't usable anymore and had to be recreated. This is first time I see this on 12.x, but it makes OSD unusable.

History

#1 - 07/26/2017 03:01 AM - Sage Weil

- Subject changed from Bluestore OSD won't start after unclean shutdown to bluestore: mismatched uuid in bdev_label after unclean shutdown

- Status changed from New to Need More Info

- Priority changed from Normal to Urgent

Interesting! I'm not sure how it happened but I susepct it's easy to fix. Can you catalog all of the /var/lib/ceph/osd/*/block* symlinks on your system, and also do

```
ceph-bluestore-tool show-label --dev $device
```

for every device (main or wal) in the system? Somewhere the streams got crossed...

Thanks!

#2 - 07/26/2017 05:12 AM - Tomasz Torcz

This is my experimental setup, so only few harddisks, each with 2 partitions (metadata and block storage). No separate WALs etc.

```
 1   hdd 0.29099         osd.1       down        0 1.00000
 3   hdd 0.07269         osd.3         up  1.00000 1.00000
 4   hdd 0.29099         osd.4         up  1.00000 1.00000
 5   hdd 0.29099         osd.5         up  1.00000 1.00000
```

```
[root@rolling ceph]# ls -l /var/lib/ceph/osd/*/block
lrwxrwxrwx. 1 ceph ceph 58 Jul 24 18:50 /var/lib/ceph/osd/ceph-1/block -> /dev/disk/by-partuuid/973dff0a-63c6-436d-b72e-72af0d4453b7
lrwxrwxrwx. 1 ceph ceph 58 Jul 24 18:52 /var/lib/ceph/osd/ceph-3/block -> /dev/disk/by-partuuid/2028f2a4-d1cf-46dd-b2e6-278e4062dd14
lrwxrwxrwx. 1 ceph ceph 58 Jul 24 18:52 /var/lib/ceph/osd/ceph-4/block -> /dev/disk/by-partuuid/ed424caa-a327-4b8c-bba5-eed922a30833
lrwxrwxrwx. 1 ceph ceph 58 Jul 25 18:01 /var/lib/ceph/osd/ceph-5/block -> /dev/disk/by-partuuid/50b0b5f8-acc3-4ad0-85c3-96633efd2735
```

> ceph-bluestore-tool show-label --dev $device
>
> for every device (main or wal) in the system? Somewhere the streams got crossed...

```
# for b in /var/lib/ceph/osd/*/block; do echo $b; ceph-bluestore-tool show-label --dev $b; done
```

```
/var/lib/ceph/osd/ceph-1/block
```

```
infering bluefs devices from bluestore path
```

```
action show-label
[
    {
        "osd_uuid": "dd3e6c5e-da75-433b-b619-25dc9cc031a4",
        "size": 319967006720,
        "btime": "2017-07-24 18:50:28.542135",
        "description": "main"
    }
]

/var/lib/ceph/osd/ceph-3/block

infering bluefs devices from bluestore path
action show-label
[
    {
        "osd_uuid": "e13102b6-73eb-4ef6-8f4f-7ee7eed33099",
        "size": 79920435200,
        "btime": "2017-07-24 18:52:09.760262",
        "description": "main"
    }
]

/var/lib/ceph/osd/ceph-4/block
infering bluefs devices from bluestore path
action show-label
[
    {
        "osd_uuid": "9536ce96-4765-42c2-96a1-ebe14cd3356d",
        "size": 319967006720,
        "btime": "2017-07-24 18:52:30.411049",
        "description": "main"
    }
]

/var/lib/ceph/osd/ceph-5/block
infering bluefs devices from bluestore path
action show-label
[
    {
        "osd_uuid": "b3b0939f-0364-4b46-9530-9f5d397b82fb",
        "size": 319967006720,
        "btime": "2017-07-25 18:02:02.935156",
        "description": "main"
    }
]
```

#3 - 07/26/2017 05:15 AM - Tomasz Torcz

Also:

```
[root@rolling osd]# grep fsid /etc/ceph/ceph.conf
fsid = 85177bde-19e4-497a-b824-d3425a161264

[root@rolling osd]# pwd
/var/lib/ceph/osd

[root@rolling osd]# cat ceph-*/ceph_fsid
85177bde-19e4-497a-b824-d3425a161264
85177bde-19e4-497a-b824-d3425a161264
85177bde-19e4-497a-b824-d3425a161264
85177bde-19e4-497a-b824-d3425a161264

[root@rolling osd]# cat ceph-*/block_uuid
973dff0a-63c6-436d-b72e-72af0d4453b7
2028f2a4-d1cf-46dd-b2e6-278e4062dd14
ed424caa-a327-4b8c-bba5-eed922a30833
50b0b5f8-acc3-4ad0-85c3-96633efd2735
```

#4 - 07/26/2017 01:55 PM - Sage Weil

one last piece of info: 'ceph osd dump' output

#5 - 07/26/2017 01:55 PM - Sage Weil

how did you provision these osds?

#6 - 07/26/2017 02:07 PM - Tomasz Torcz

Provisioning was basic: wipefs -a all the partitions and the whole disk at the end. Then

ceph-disk prepare /dev/sdi
ceph-disk activate /dev/sdi1

Using ceph-base-12.1.1-1.fc27.x86_64 . No special options – I've used defaults.

```
ceph osd dump

epoch 142
fsid 128d924d-38ab-4d9b-a4fe-74a59e3ca69c
created 2017-07-24 18:35:10.268075
modified 2017-07-26 11:33:12.249007
flags sortbitwise
crush_version 16
full_ratio 0.95
backfillfull_ratio 0.9
nearfull_ratio 0.85
require_min_compat_client jewel
min_compat_client jewel
```

```
require_osd_release luminous
pool 3 'syfs_meta' replicated size 2 min_size 1 crush_rule 0 object_hash rjenkins pg_num 32 pgp_num 32 last_ch
ange 141 flags hashpspool stripe_width 0
pool 4 'syfs_deta' replicated size 2 min_size 1 crush_rule 0 object_hash rjenkins pg_num 128 pgp_num 128 last_
change 142 flags hashpspool stripe_width 0
max_osd 6
osd.1 down out weight 0 up_from 9 up_thru 58 down_at 61 last_clean_interval [0,0) [2001:470:71:68d:8e72:8e99:d
fb3:c684]:6805/16006 [2001:470:71:68d:8e72:8e99:dfb3:c684]:6806/16006 [2001:470:71:68d:8e72:8e99:dfb3:c684]:68
07/16006 [2001:470:71:68d:8e72:8e99:dfb3:c684]:6808/16006 autoout,exists dd3e6c5e-da75-433b-b619-25dc9cc031a4
osd.3 up   in  weight 1 up_from 118 up_thru 138 down_at 115 last_clean_interval [79,114) [2001:470:71:68d:8e72
:8e99:dfb3:c684]:6801/10426 [2001:470:71:68d:8e72:8e99:dfb3:c684]:6802/10426 [2001:470:71:68d:8e72:8e99:dfb3:c
684]:6803/10426 [2001:470:71:68d:8e72:8e99:dfb3:c684]:6804/10426 exists,up e13102b6-73eb-4ef6-8f4f-7ee7eed3309
9
osd.4 up   in  weight 1 up_from 136 up_thru 138 down_at 132 last_clean_interval [122,131) [2001:470:71:68d:613
3:8356:fec3:93b6]:6800/6944 [2001:470:71:68d:6133:8356:fec3:93b6]:6801/6944 [2001:470:71:68d:6133:8356:fec3:93
b6]:6802/6944 [2001:470:71:68d:6133:8356:fec3:93b6]:6803/6944 exists,up 9536ce96-4765-42c2-96a1-ebe14cd3356d
osd.5 up   in  weight 1 up_from 127 up_thru 138 down_at 125 last_clean_interval [87,124) [2001:470:71:68d:8e72
:8e99:dfb3:c684]:6809/10655 [2001:470:71:68d:8e72:8e99:dfb3:c684]:6810/10655 [2001:470:71:68d:8e72:8e99:dfb3:c
684]:6811/10655 [2001:470:71:68d:8e72:8e99:dfb3:c684]:6812/10655 exists,up b3b0939f-0364-4b46-9530-9f5d397b82f
b
pg_temp 3.0 [5]
pg_temp 3.4 [5]
pg_temp 3.d [5]
pg_temp 3.e [5]
pg_temp 3.f [5]
pg_temp 3.10 [5]
pg_temp 3.12 [5]
pg_temp 3.13 [5]
pg_temp 3.14 [5]
pg_temp 3.15 [3]
pg_temp 3.16 [3]
pg_temp 3.19 [3]
pg_temp 3.1c [5]
pg_temp 3.1d [5]
pg_temp 3.1f [5]
pg_temp 4.2 [5]
pg_temp 4.3 [5]
pg_temp 4.4 [5]
pg_temp 4.5 [5]
pg_temp 4.7 [5]
pg_temp 4.8 [5]
pg_temp 4.9 [5]
pg_temp 4.a [5]
pg_temp 4.b [5]
pg_temp 4.f [3]
pg_temp 4.10 [3]
pg_temp 4.12 [3]
pg_temp 4.17 [5]
pg_temp 4.19 [5]
pg_temp 4.20 [5]
pg_temp 4.21 [5]
pg_temp 4.24 [5]
pg_temp 4.26 [5]
pg_temp 4.27 [5]
pg_temp 4.28 [5]
pg_temp 4.2a [5]
pg_temp 4.2b [3]
pg_temp 4.2c [3]
pg_temp 4.2d [5]
pg_temp 4.2f [5]
pg_temp 4.31 [5]
pg_temp 4.32 [5]
pg_temp 4.36 [5]
pg_temp 4.38 [5]
pg_temp 4.3c [5]
pg_temp 4.41 [5]
pg_temp 4.42 [5]
pg_temp 4.4a [5]
pg_temp 4.4c [5]
pg_temp 4.4d [5]
pg_temp 4.4e [5]
pg_temp 4.56 [5]
pg_temp 4.57 [5]
pg_temp 4.5a [5]
pg_temp 4.5d [5]
```

```
pg_temp 4.5e [5]
pg_temp 4.62 [5]
pg_temp 4.64 [3]
pg_temp 4.65 [5]
pg_temp 4.67 [5]
pg_temp 4.6a [5]
pg_temp 4.6d [3]
pg_temp 4.70 [5]
pg_temp 4.72 [5]
pg_temp 4.73 [5]
pg_temp 4.74 [5]
pg_temp 4.7d [5]
pg_temp 4.7e [5]
pg_temp 4.7f [5]
```

#7 - 07/26/2017 07:12 PM - Tomasz Torcz

Ukhm. There was something completely messed up with my installation. FS UUID shown by 'ceph -s' was different than the one in /etc/ceph/ceph.conf. The value in osd/ceph-1/ceph_fsid was wrong, too, and osd/ceph-1/fsid was wrong, too (different than in stored in ceph-1/block). I don't know how this happened (I wasn't careful enough erounh ceph-deploy?) but after sorting out all the vales - the OSD starts, runs recover and seems to be fine.

Sorry for wasting your time on this ticket. But as I mentioned, I saw similar issues on other cluster when the bluestore node was rebooted hard. If I see it again, I will gather full logs and open a new ticket. This one can be closed.

#8 - 07/26/2017 07:51 PM - Sage Weil

- Priority changed from Urgent to Normal

yeah, i'm trying to figure out how the 'fsid' file in each osd data dir came to be wrong. if you're able to reproduce this, please let me know!

it may be that having a ceph.conf in teh ceph-deploy directory that didn't match teh one on the node you were deploy an osd to contributed? just guessing.

#9 - 08/29/2017 07:39 PM - Sage Weil

- Status changed from Need More Info to Can't reproduce