Ceph - Feature #15346

add error handle for leveldb or rocksdb in bluestore or filestore

04/01/2016 02:31 AM - Xinxin Shu

Status:	Resolved	**Start date:**	04/01/2016
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	0.00 hour
Source:	other	**Reviewed:**	
Tags:		**Affected Versions:**	
Backport:		**Pull request ID:**	

Description

recently i tested bluestore, i met error that rocksdb corruptted(submit_transaction always return -1)but ceph does not handle this error.

History

#1 - 04/08/2016 03:12 AM - shasha lu

I met the error too. (submit_transaction_sync always return -1)

I added logs in RocksDBStore::submit_transaction_sync , found out rocksdb return error:
2016-04-08 08:46:02.064797 7f1093600700 -1 rocksdb: submit_transaction_sync IO error: /opt/ceph-bluestore/var/lib/blue/osd/osd4/db/024586.log: No such file or directory

My ceph version is 10.0.5
BlueStore_bluefs = false

rocksdb include these files:

```
[root@lj20 db]# pwd
/opt/ceph-bluestore/var/lib/blue/osd/osd4/db
[root@lj20 db]# ls
006585.sst  019654.sst  020975.sst  021884.sst  022248.sst  023086.sst  023565.sst  024015.sst  024205.sst  024421.sst  024487.sst  024523.sst
024540.sst  024556.sst  024571.sst  LOCK
015462.sst  019826.sst  020976.sst  021886.sst  022249.sst  023087.sst  023566.sst  024018.sst  024206.sst  024422.sst  024488.sst  024525.sst
024541.sst  024557.sst  024572.sst  MANIFEST-021513
016094.sst  019921.sst  021164.sst  021998.sst  022310.sst  023088.sst  023567.sst  024019.sst  024207.sst  024423.sst  024489.sst  024526.sst
024543.sst  024558.sst  024573.sst  OPTIONS-021510
016181.sst  020179.sst  021165.sst  021999.sst  022311.sst  023197.sst  023568.sst  024054.sst  024268.sst  024424.sst  024508.sst  024527.sst
024544.sst  024559.sst  024574.sst  OPTIONS-021516
016183.sst  020502.sst  021166.sst  022000.sst  022312.sst  023198.sst  023780.sst  024055.sst  024269.sst  024425.sst  024511.sst  024528.sst
024545.sst  024560.sst  024575.sst
016224.sst  020503.sst  021167.sst  022001.sst  022313.sst  023309.sst  023781.sst  024080.sst  024270.sst  024427.sst  024512.sst  024529.sst
024546.sst  024561.sst  024576.sst
016386.sst  020504.sst  021334.sst  022002.sst  022403.sst  023313.sst  023782.sst  024082.sst  024353.sst  024455.sst  024513.sst  024530.sst
024547.sst  024562.sst  024577.sst
017622.sst  020725.sst  021386.sst  022003.sst  022829.sst  023314.sst  023783.sst  024135.sst  024399.sst  024479.sst  024514.sst  024531.sst
024548.sst  024563.sst  024578.sst
017845.sst  020726.sst  021387.sst  022005.sst  022969.sst  023498.sst  023784.sst  024136.sst  024400.sst  024480.sst  024516.sst  024532.sst
024549.sst  024564.sst  024579.sst
017846.sst  020727.sst  021388.sst  022066.sst  022970.sst  023499.sst  023962.sst  024137.sst  024401.sst  024481.sst  024517.sst  024534.sst
024550.sst  024565.sst  024580.sst
017912.sst  020728.sst  021453.sst  022067.sst  022971.sst  023500.sst  023964.sst  024186.sst  024402.sst  024482.sst  024518.sst  024535.sst
024551.sst  024566.sst  024581.sst
019294.sst  020790.sst  021609.sst  022068.sst  023081.sst  023560.sst  023965.sst  024200.sst  024404.sst  024483.sst  024519.sst  024536.sst
024552.sst  024567.sst  024584.log
019295.sst  020854.sst  021670.sst  022070.sst  023082.sst  023561.sst  023966.sst  024201.sst  024405.sst  024484.sst  024520.sst  024537.sst
024553.sst  024568.sst  024585.sst
019420.sst  020973.sst  021671.sst  022245.sst  023083.sst  023562.sst  023967.sst  024202.sst  024406.sst  024485.sst  024521.sst  024538.sst
024554.sst  024569.sst  CURRENT
019421.sst  020974.sst  021883.sst  022246.sst  023084.sst  023564.sst  023968.sst  024204.sst  024407.sst  024486.sst  024522.sst  024539.sst
024555.sst  024570.sst  IDENTITY
[root@lj20 db]# ls ~~al | grep 'log'~~
~~-rw-r--r-~~ 1 root root 4015151 Apr  8 08:46 024584.log
```

#2 - 04/13/2016 05:29 AM - Yang Dongsheng

Hi guys, I am trying to fix this problem by introducing a mechanism of _txc_abort(). It will reply an -EIO to client if we met any problem in transaction submitting.

does that sounds good enough?

#3 - 04/14/2016 08:40 AM - shasha lu

This is rocksdb's bug.
Rocksdb report:
2016-04-14 14:37:13.705520 7f0d4f4e3700 2 rocksdb: Waiting after background compaction error: IO error:
/opt/ceph-bluestore/var/lib/blue/osd/osd1/db/407669.log: No such file or directory, Accumulated background error counts: 1

Return -EIO to client maybe is not proper.

#4 - 04/14/2016 10:00 AM - Yang Dongsheng

shasha lu wrote:

> This is rocksdb's bug.
> Rocksdb report:
> 2016-04-14 14:37:13.705520 7f0d4f4e3700 2 rocksdb: Waiting after background compaction error: IO error:
> /opt/ceph-bluestore/var/lib/blue/osd/osd1/db/407669.log: No such file or directory, Accumulated background error counts: 1

I am not sure is that a bug in socksdb here, but I am sure objectstore has no method to abort a transaction. That's what I am working for.

> Return -EIO to client maybe is not proper.

Why not, please consider this scenario, there is something wrong in the device of rocksdb and we got an EIO from writing data into it. Then we return a -EIO to user, why not proper?

#5 - 05/21/2016 03:05 AM - Yang Dongsheng

https://github.com/ceph/ceph/pull/8599

Hi, guys, does this commit solve it?

#6 - 05/21/2016 03:54 AM - shasha lu

yes, osd will abort when submit_transaction met error.

BTW, the default bluestore rocksdb options is
bluestore_rocksdb_options =
compression=kNoCompression,max_write_buffer_number=16,min_write_buffer_number_to_merge=3,recycle_log_file_num=16

I remove the option recycle_log_file_num=16, with ceph.conf
bluestore_rocksdb_options = compression=kNoCompression,max_write_buffer_number=16,min_write_buffer_number_to_merge=3

The error no longer appears.
Maybe the options recycle_log_file_num should be killed in config_opts.h

#7 - 06/21/2017 02:24 AM - Sage Weil

- Status changed from New to Resolved