bluefs性能测试

测试环境和测试方法

1.测试环境

nvmessd:930GB

内存：54GB

CPU:Intel Xeon Processor (Skylake) 2.4GHz，12线程

内核：Linux 3.10.0-957.el7.x86_64

ceph版本：12.2.4

2.测试方法

将rocksdb中db_bench模块相关代码移植到ceph的test模块中，利用ObjectStore::create方法创建一个BlueStore对象，利用该BlueStore实例构造一个BlueFS用户态文件系统和BlueRocksEnv（继承于rocksdb::Env)全局对象env；最后将该env导出到db_bench中即可，实现将db_bench中所有的I/O操作通过env对象转发到ceph的bluefs上。

使用db_bench分别测试fillseq，readseq和readrandom三个benchmark，同时在bluefs和本地文件系统XFS上进行对比测试，并且分别在单线程和8线程的情况下进行测试比较。

测试脚本示例如下：

```
./db_bench -compression_type=none -threads=1 \
  -benchmarks=readrandom -db=bluefs_test_db_16 \
  -use_bluefs=true -use_existing_db -duration=60 \
  -num=50000000 -value_size=16 2>/dev/null
```

完整测试脚本：

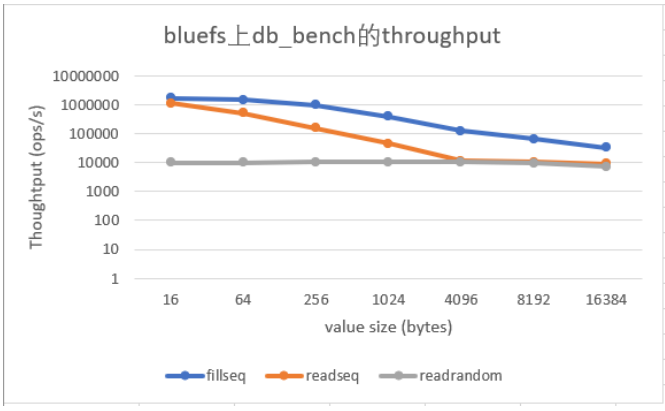
1. db_bench_bluefs.sh
2. db_bench_xfs.sh

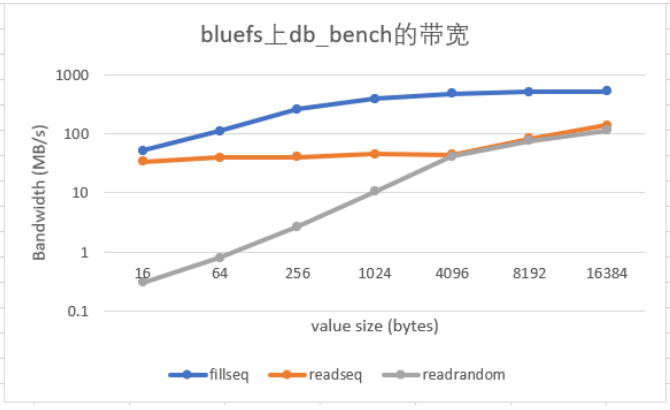
测试结果

0.完整数据：db_bench_bluefs_vs_xfs_result.xlsx。

1. db_bench运行在bluefs上的性能表现，其中ceph.conf采用默认配置：

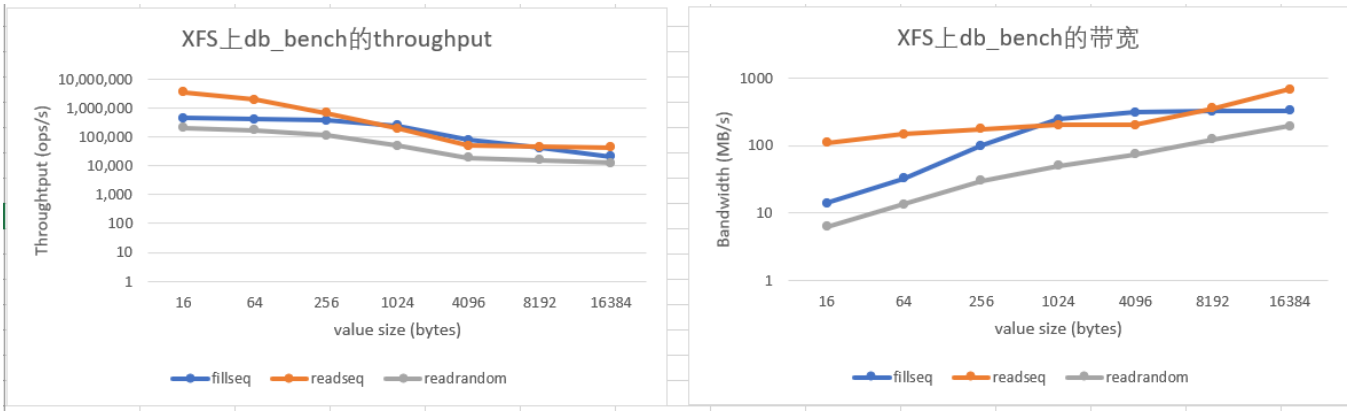
	16	64	256	1024	4096	8192	16384
fillseq	1708942(ops/s), 52.2 (MB/s)	1471561, 112.3	1005054, 260.7	397025, 393.8	123170, 483	66310, 519.1	33879, 529.9
readseq	1116664, 34.1	522274, 39.8	158200, 41	45332, 45	11268, 44.2	10610, 83.1	8968, 140.3
readrandom	10214, 0.3	9958, 0.8	10572, 2.7	10595, 10.5	10681, 41.9	9809, 76.8	7274, 113.8



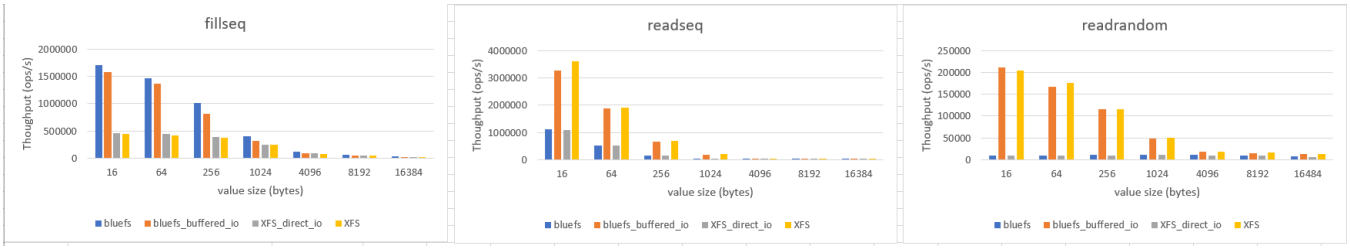


2．db_bench运行在XFS上的性能表现（注：db_bench参数值默认使用文件系统缓存）：

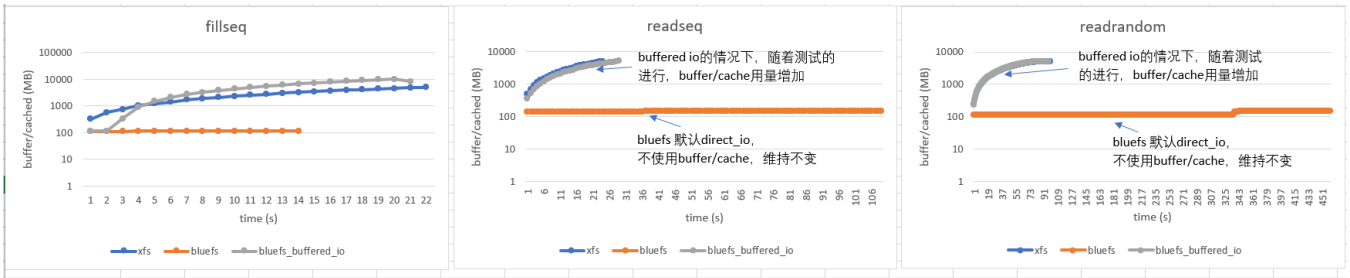
	16	64	256	1024	4096	8192	16384
fillseq	450596 (ops/s), 13.8 (MB/s)	421381, 32.1	378429, 98.2	246607, 244.6	78609, 308.3	40940, 320.5	21262, 332.6
readseq	3605057, 110	1925935, 146.9	680949, 176.6	200972, 199.3	51055, 200.2	45131, 353.3	43770, 684.6
readrandom	203875, 6.2	176118, 13.4	116083, 30.1	49938, 49.5	18706, 73.4	15712, 123	12438, 194.5



3．bluefs 和 XFS之间的性能（吞吐量）对比结果如下，其中bluefs表示将ceph.conf中bluefs_buffered_io采用默认值false，bluefs_buffered_io表示将bluefs_buffered_io设置为true进行测试；xfs_direct_io表示在db_bench中设置use_direct_reads参数为true采用direct io，xfs表示使用文件系统缓存。下同：

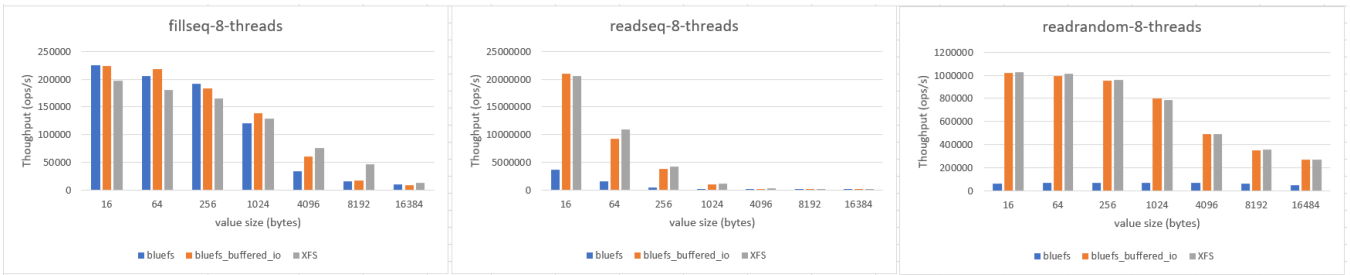


从图中可以发现，bluefs和XFS_direct_io，bluefs_buffered_io和XFS，性能相当；有操作系统缓存的情况下性能优于没有操作系统缓存，对操作系统中缓存大小进行监控，选择value size为1024字节db总大小为5GB左右的测试用例进行测试，得到如下结果，由此结果可见，打开操作系统缓存时，系统的内存使用量随着测试的进行，不断增高：



注：db size 5GB左右，采用value size为1024字节进行测试

4.八线程并发I/O测试，设置db_bench的threads参数为8进行测试，此处列出bluefs，bluefs_buffered_io和XFS之间的性能（吞吐量，throughput）对比，完整具体数据见excel表格：



总结

1. 在fillseq测试中，bluefs的性能比一直比XFS的高，随着value大小的增加，bluefs的性能优势越来越小。
2. 在读测试下，bluefs的性能比XFS低，随着value大小的增加，两者之间的差距逐渐变小。
3. readrandom测试中，bluefs的throughput随着value大小的增加，基本保持不变，维持在10000ops/s左右。
4. 在单线程测试中，bluefs_buffered_io性能表现和XFS在value大小的趋势上类似，同时bluefs_buffered_io的性能总是略小于XFS。
5. 8线程测试情况下，fillseq中在value大小小于1024字节时，bluefs的性能优于XFS，大于1024字节时bluefs的性能比XFS差。readseq中，bluefs的性能也比XFS的要差，两者差距随着value大小的增加逐渐变小。readrandom中，和单线程类似，bluefs的throughput基本维持不变的状态，XFS随着value大小的增加throughput下降，两者之间的差距也逐渐缩小，但bluefs始终比XFS的性能低。bluefs_buffered_io性能表现和XFS相当。

6. 同等条件下，bluefs和xfs_direct_io之间性能接近，bluefs_buffered_io和xfs之间性能也几乎相同。换句话说，在db_bench下，目前bluefs和XFS本地文件系统性能相当。

参考文献

1. 一种基于LSM 树的键值存储系统性能优化方法 http://crad.ict.ac.cn/EN/article/downloadArticleFile.do?attachType=PDF&id=3996
2. https://www.pdl.cmu.edu/PDL-FTP/FS/CMU-PDL-19-102.pdf