Alibaba Group

MyRocks: Best Practice at Alibaba

Percona Live 2017

April 26, 2017

Agenda

- ✓ Brief MyRocks introduction
- ✓ Performance test
- ✓ Benefit we got
- ✓ MyRocks improvements we made
- ✓ Where to go?

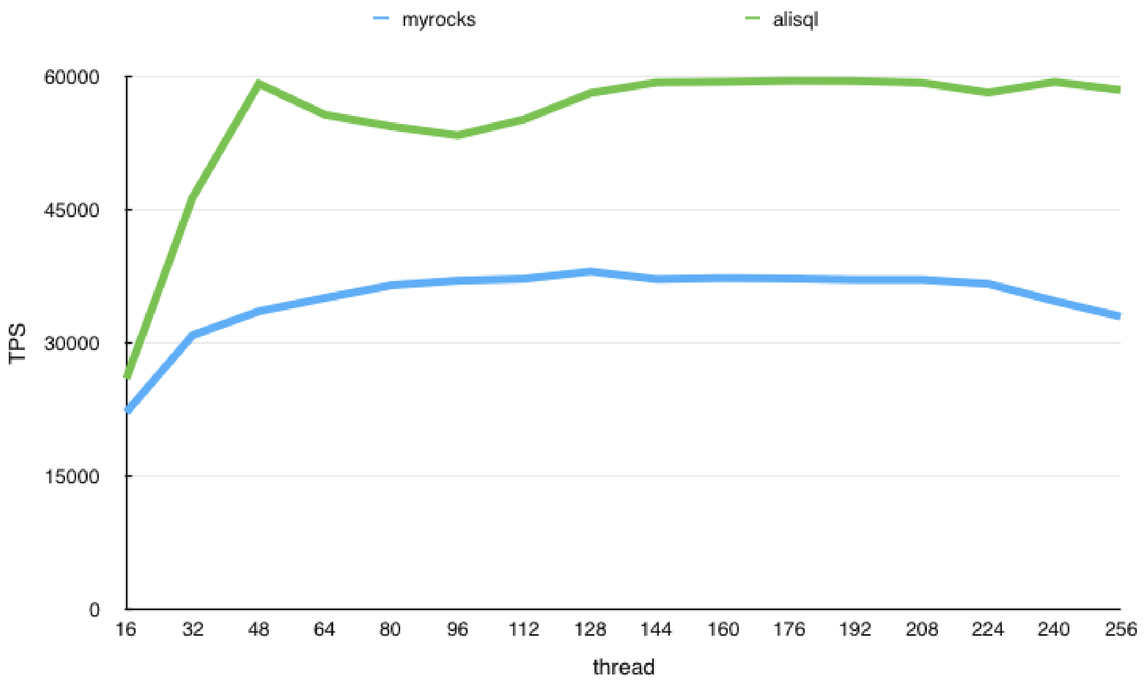
✓ Brief MyRocks introduction

1. RocksDB uses LSM tree
2. RocksDB is written friendly with high data compression
3. RocksDB is widely used as K-V engine or storage engine
4. The Alibaba database team is active in the RocksDB open source community

✓ Some of the issues we were facing at Alibaba

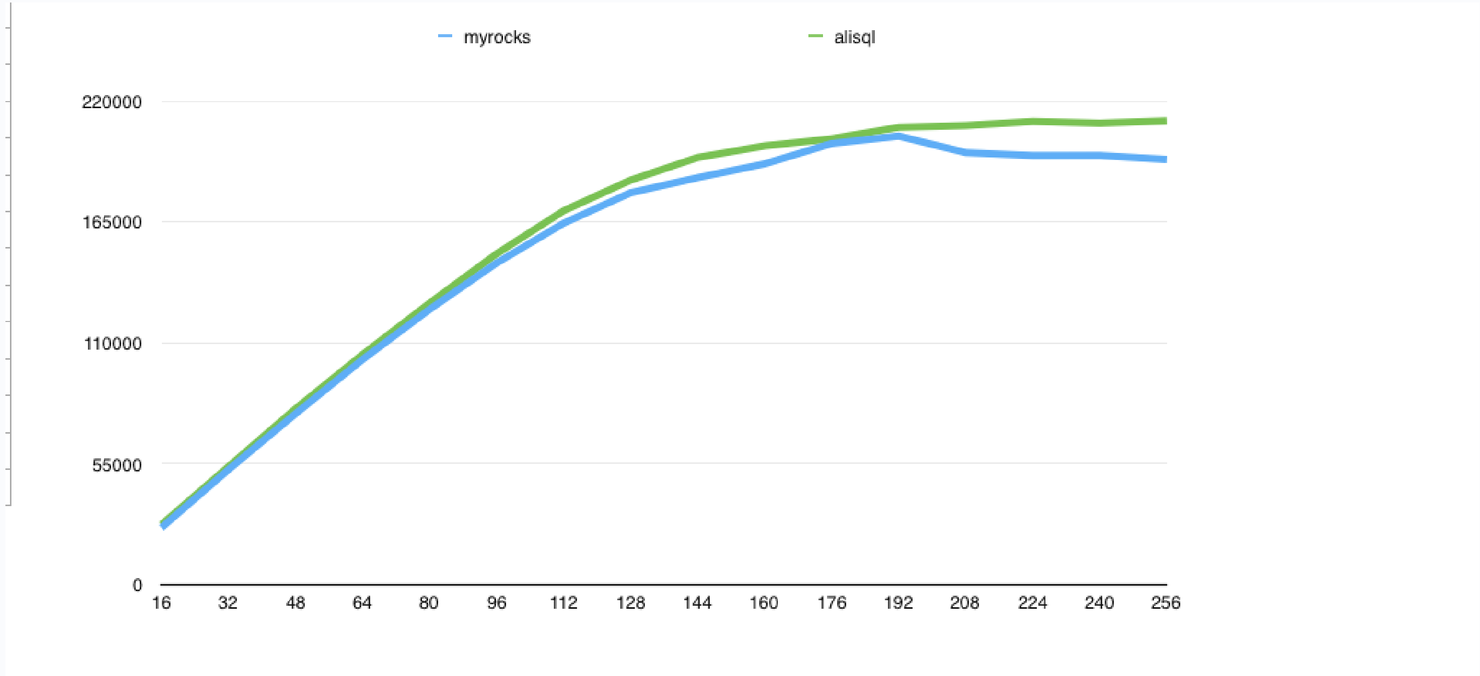
1. Write amplification issue on SSD
2. Data files keep growing, and DBAs have to deal with space issue every day

✓ Performance evaluation



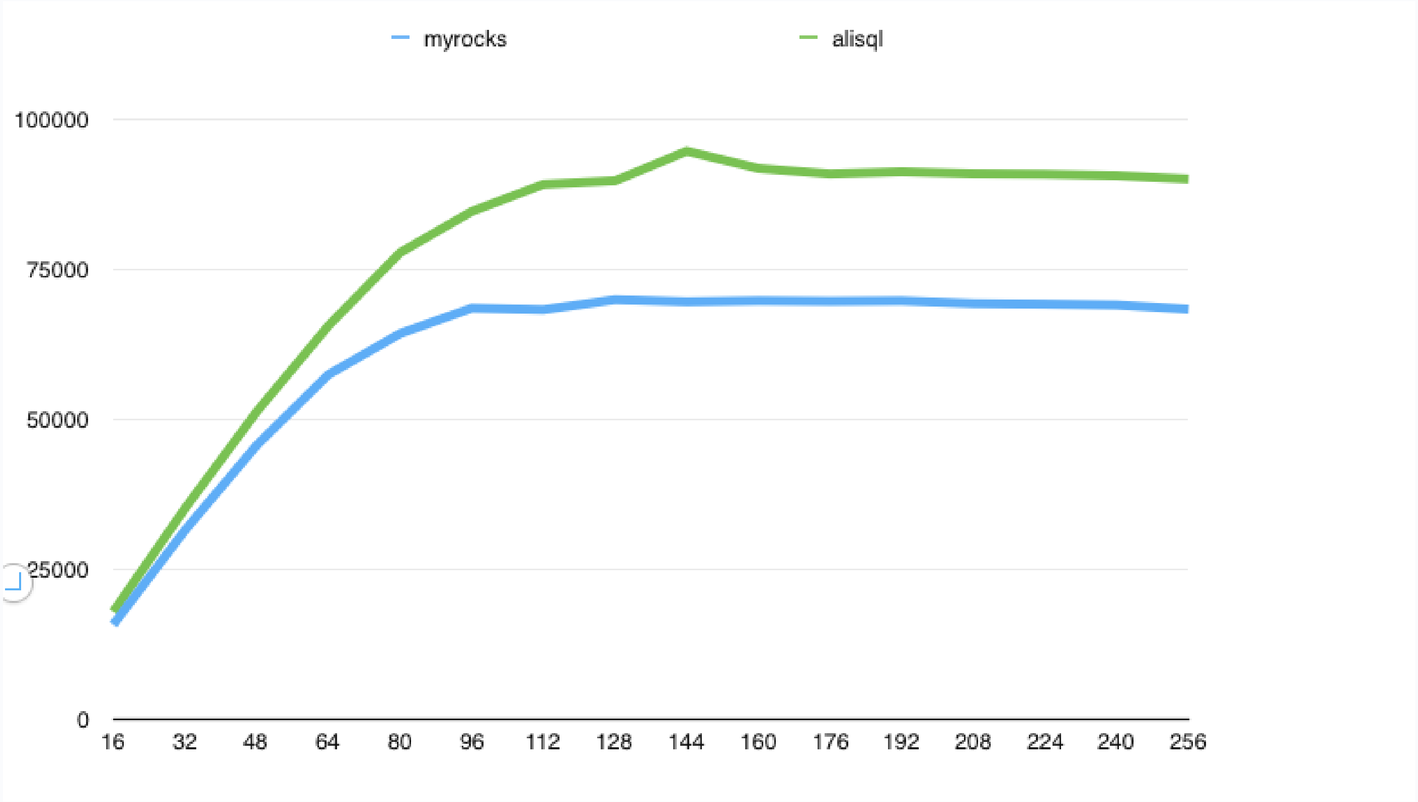
Insert performance meet the basic requirement, while there is still room to improve.

✓ Performance evaluation



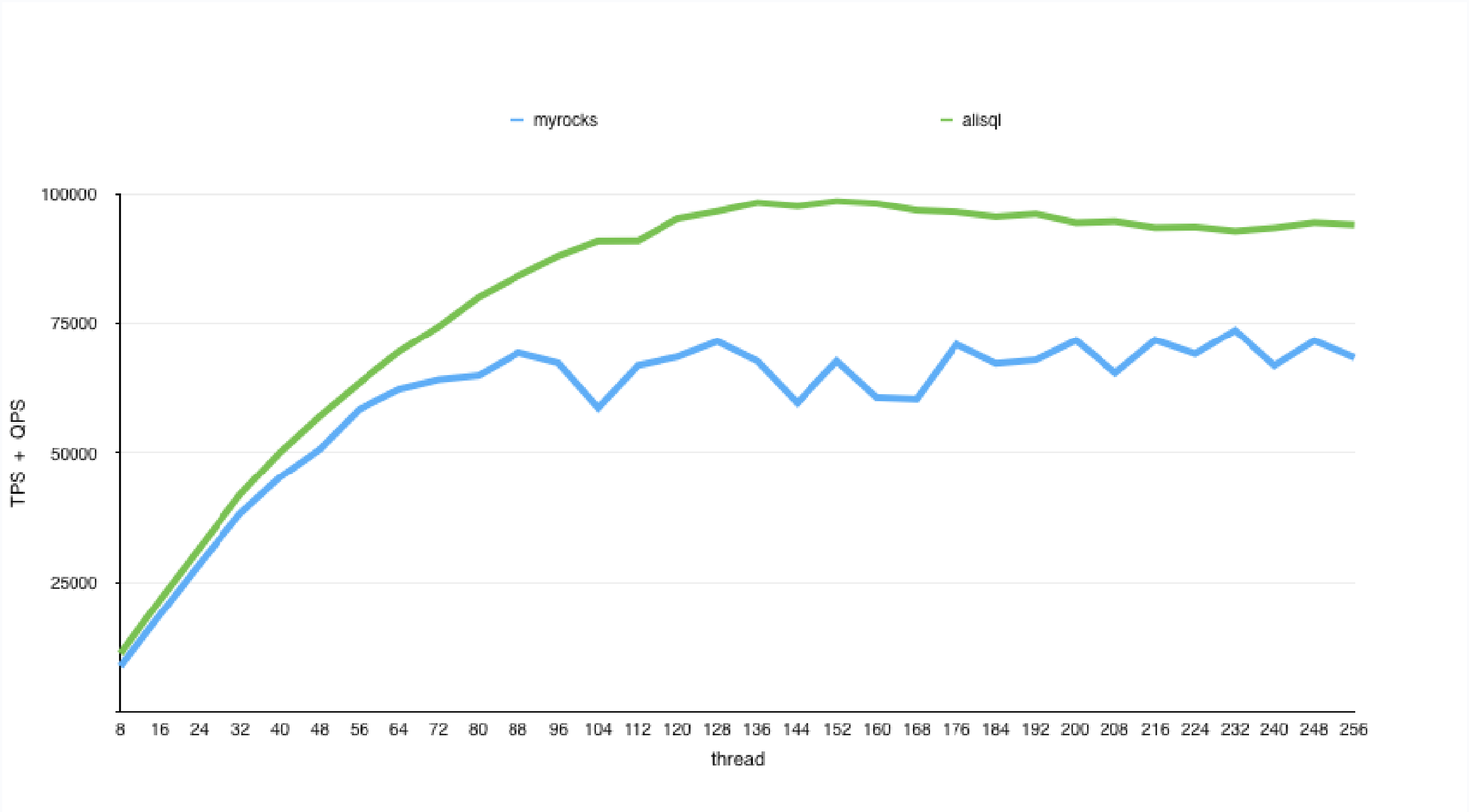
Simple query is not slow with the help of bloom filter.

✓ Performance evaluation



Range scan is slower than InnoDB, while it's still acceptable.

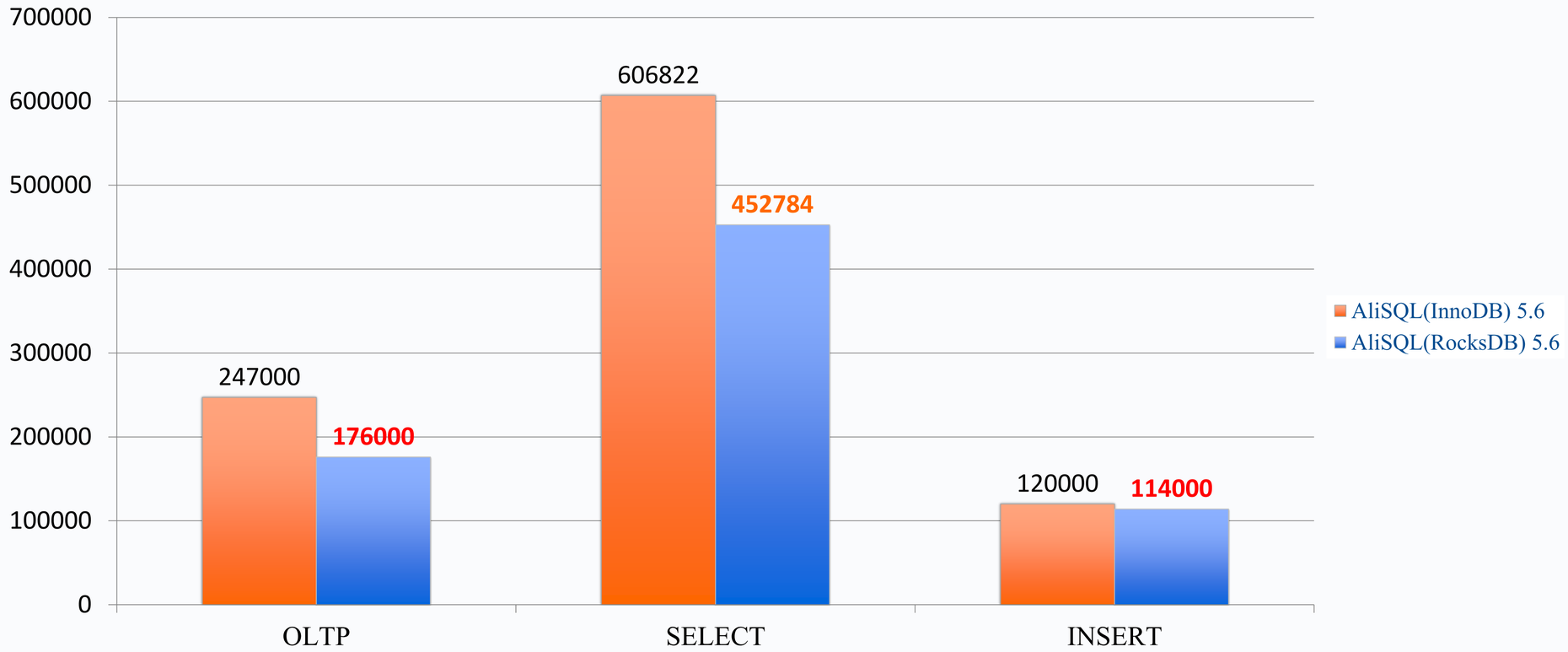
✓ Performance evaluation



OLTP workload is not that good.

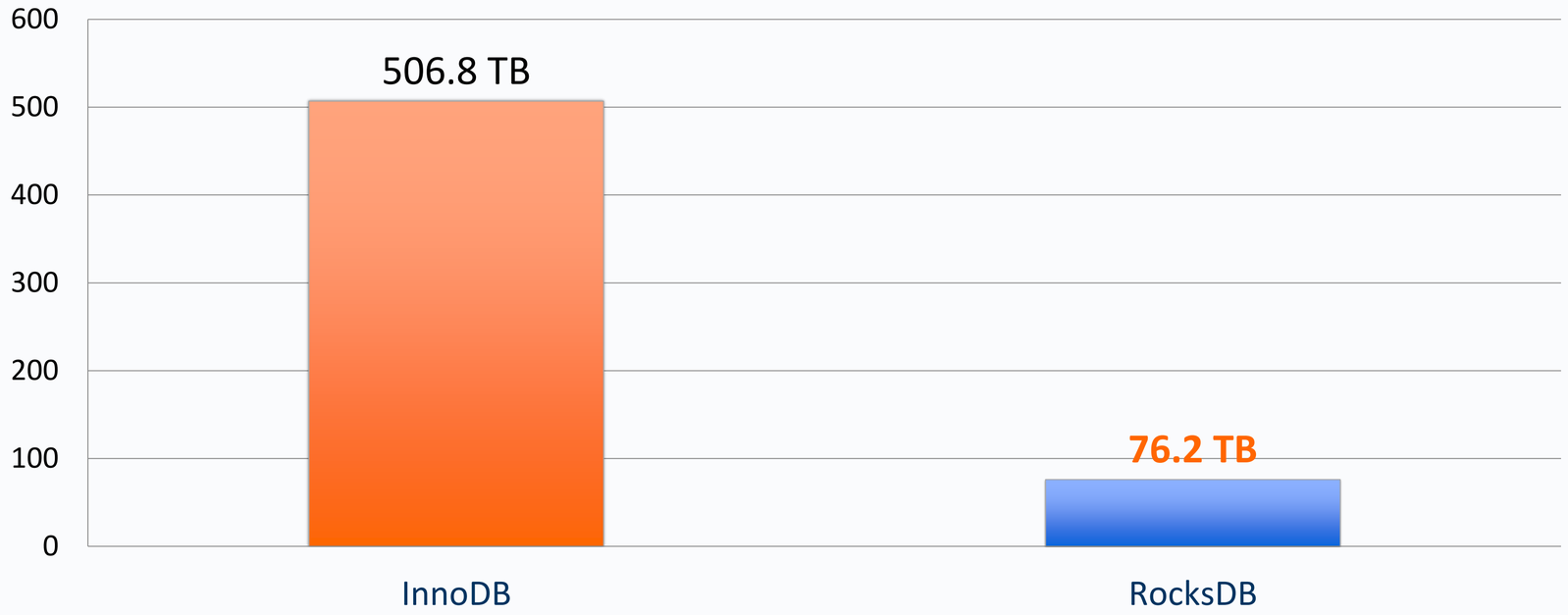
What if we combine MyRocks with AliSQL

✓ Performance Comparison



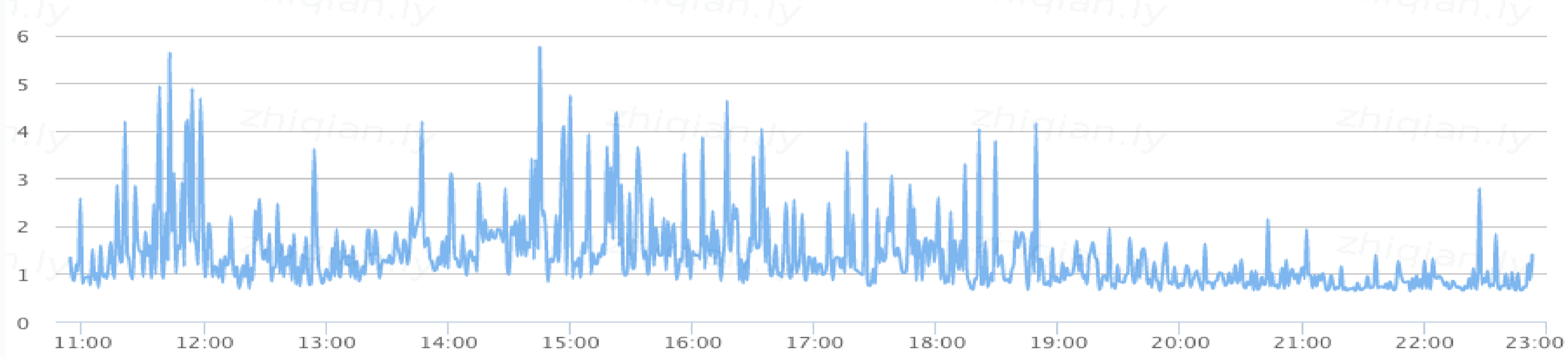
Sysbench：AliSQL(RocksDB) vs AliSQL(InnoDB)

✓ Real online system scenario

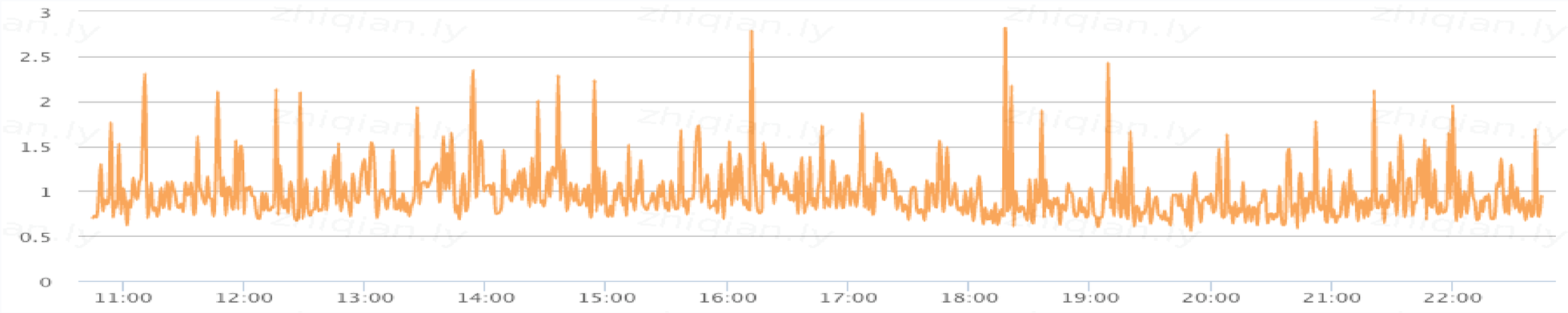


With very small RT trade-off, reduced storage to almost 1/7, saving $$$$$$ storage investment;

✓ RT comparison



AliSQL(RocksDB)



AliSQL(InnoDB)

✓ The benefit we got out of MyRocks

1. Including Taobao, AliExpress, Cainiao logistics, offline search and many other business
2. Average storage size is about 1/3 ~ 1/8 of InnoDB, according to our usage
3. Huge opportunity to combine the advantage of LSM tree with B+ tree

✓ MyRocks Known Limitations

1. Not supported online DDL.
2. Large transaction is not supported well.
3. Many parameters can not be set dynamically, like rocksdb_write_sync.
4. In-memory insert performance is limited by compaction.
5. Lack of sophisticated statistics information.
6. Like, count, order by, group by is slower than innodb.
7. Delete or update operation will affect the efficiency of the query.

✓ Additional issues encountered

1. Unordered data write is significantly slower than ordered data

2. Performance is very poor under high write workload

3. Large table DDL (add index or field) may cause OOM

4. Commit is very heavy which makes the whole commit process very slow

✓ Improvements we made - bug fixed

1. rocksdb group commit #481
2. alter table drop key cause mysqld crash #602
3. use index_merge cause mysqld crash #604
4. commit_in_middle cause mysqld crash #608
5. inserts hang in slowdown #610
6. SIGSEGV on RocksJava #1267

✓ Improvements we made – feature we added

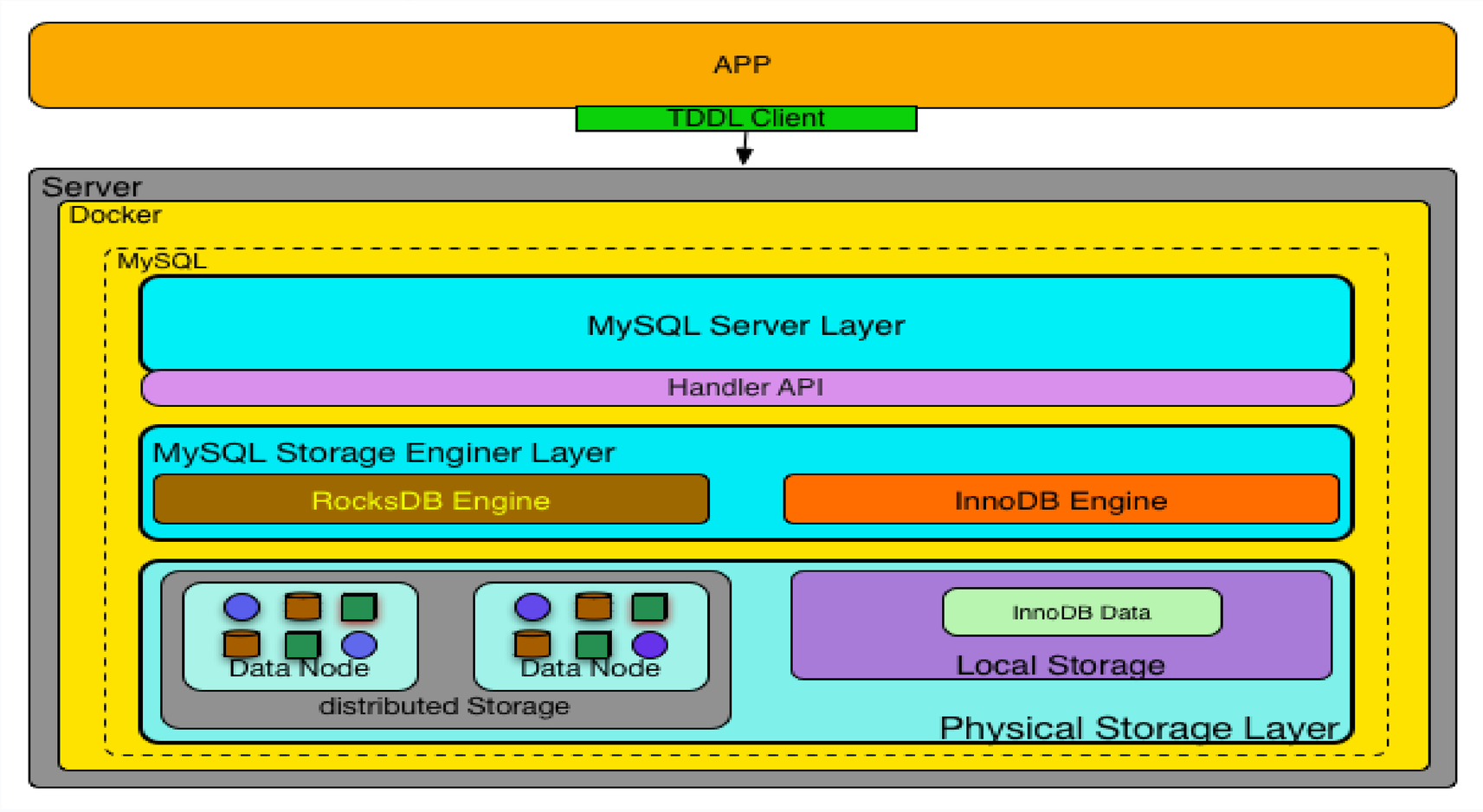
1. Extended Chinese character verification (gbk_chinese_ci, utf8_general_ci etc.)
2. Added support for thread pool function in the server layer
3. Added support for slow SQL filtering
4. PK query performance optimization
5. Master-slave table-level replication in parallel

✓ Key Take Away

1. MyRocks is not a silver bullet, cannot resolve all your issues, while it really can help you minimize your overwhelming data files by the high compression ratio
2. MyRocks still has long way to go, and it’s not so mature as InnoDB at this moment, however, we can see an active community which is crucial to the success of an Open Source project

✓ Where to go?

AliSQL(RocksDB+InnoDB) bi-engine



Contact info: Jiayi.wjy@Alibaba-inc.com

Twitter: jiayiw

Facebook: Jiayi Wang

Alibaba Group

Thank you !