MyRocks Storage Engine Status Update

Sergei Petrunia <sergey@mariadb.com>

MariaDB Meetup
New York
February, 2018

Plan

- What MyRocks is
- How it is provided in upstream
- Packaging MyRocks in MariaDB
- MyRocks for non-myrocks users.

What is MyRocks

- (See other talks for a long answer)
- Short answer:
  - Better compression
  - Better (lower) write amplification
    - Less SSD wear
    - Higher write throughput
  - Developed and used @ Facebook

Plan

- What MyRocks is
- How it is provided in upstream
- Packaging MyRocks in MariaDB
- MyRocks for non-myrocks users.

MyRocks lives in Facebook's MySQL branch

- github.com/facebook/mysql-5.6
  - Will call this "FB/MySQL"
- MyRocks lives there in storage/rocksdb
- FB/MySQL is easy to use if you are Facebook
  - Not so easy if you are not

FB/mysql-5.6 – user perspective

- No binaries, no packages
  - Compile yourself from source
    - Dependencies, etc.
- No releases
  - (Is the latest git revision ok?)
- Has extra features
  - e.g. extra counters “confuse” monitoring tools.

FB/mysql-5.6 – dev perspective

- Targets a CentOS-type OS
  - Compiler, cmake version, etc.
  - Others may or may not [periodically] work
    - MariaDB/Percona file pull requests to fix
- Special command to compile
  - https://github.com/facebook/mysql-5.6/wiki/Build-Steps
- Special command to run tests
  - Test suite assumes a big machine
    - Some tests even a release build

Bringing MyRocks to a wider audience

- Two porting efforts
  - MariaDB 10.2
  - Percona Server 5.7
- Porting considerations
  - Providing Packages
  - Changing “in-house” experience to be user-friendlier
  - Decoupling from FB-only features
  - Coupling with features of your version (MariaDB 10.2 or MySQL 5.7)

Plan

- What MyRocks is
- How it is provided in upstream
- Putting MyRocks into MariaDB
- MyRocks for non-myrocks users.

MyRocks in MariaDB

- MyRocks is a loadable plugin with its own Maturity level.
- Available in MariaDB 10.2+
  - MyRocks itself is the same across 10.2 and 10.3
    - New related feature in 10.3: “Per-engine mysql.gtid_slave_pos”
- Releases
  - April, 2017: MyRocks is Alpha (added to MariaDB 10.2.5 RC)
  - January, 2018: MyRocks is Beta
  - (very soon): MyRocks is RC

Keeping up to date with FB/MySQL

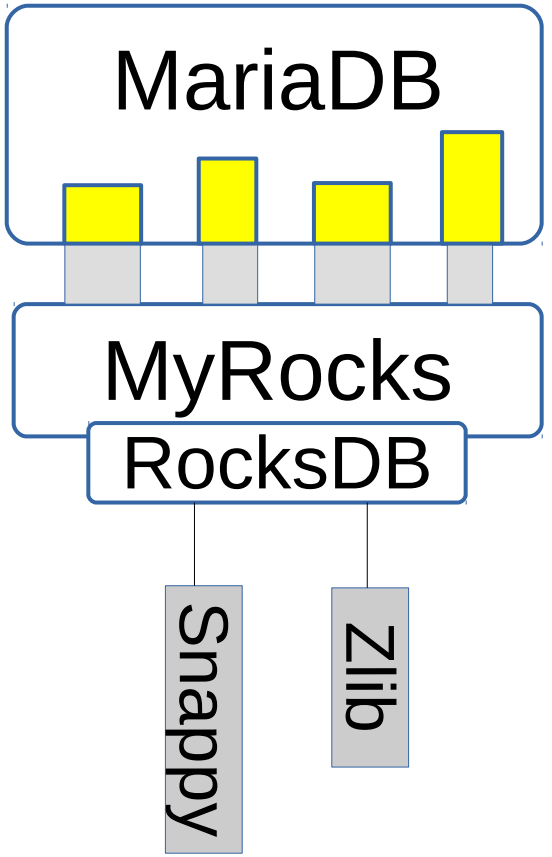
- MyRocks continues to evolve @ Facebook
- New changes are periodically merged into MariaDB
- “Merge tree” approach
  - Can view the merge status at https://github.com/MariaDB/mergetrees/commits/merge-myrocks
- Merging is still a manual process
  - But the amount of effort is reasonable

Plan

- What MyRocks is
- How it is provided in upstream
- Putting MyRocks into MariaDB
- Packaging MyRocks
- MyRocks for non-myrocks users.

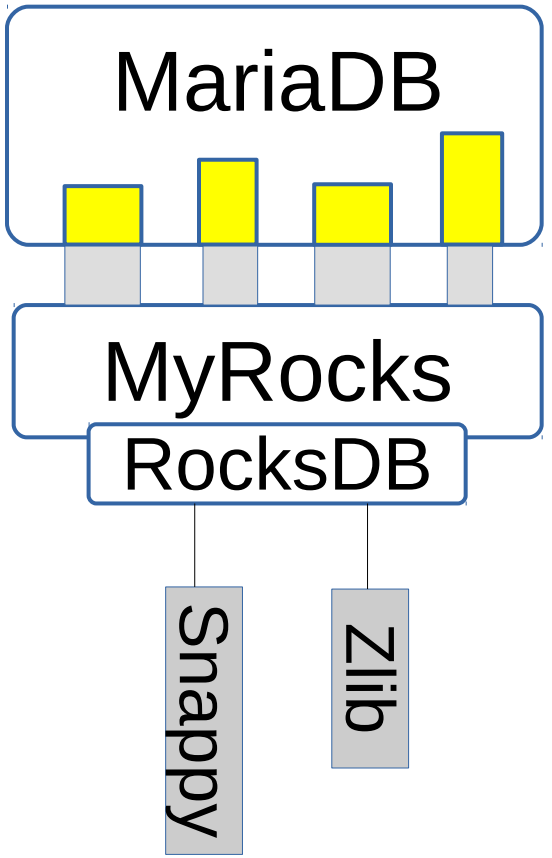
MyRocks and RocksDB library

- MyRocks is tied RocksDB@revno
  - RocksDB is a github submodule
  - No compatibility with other versions
- RocksDB is compiled together with MyRocks
- And linked-in statically
- Distros have a RocksDB package
  - Not using it.



Compression libraries

- RocksDB uses compression libraries
  - Snappy, Zlib, Bzip, LZ4, ZStandard
- Distros strongly prefer you to use OS packages
  - e.g. don’t ship your own bzip
- MariaDB’s MyRocks package has proper dependencies.



MariaDB's MyRocks package

```
Package: mariadb-plugin-rocksdb
Source: mariadb-10.3
Version: 10.3.5+maria~artful
Architecture: amd64
Maintainer: MariaDB Developers <maria-developers@lists.launchpad.net>
Installed-Size: 13227
Depends: mariadb-server-10.3 (= 10.3.5+maria~artful), libc6 (>= 2.17),
        liblz4-1 (>= 0.0~r130), libsnappy1v5, libstdc++6 (>= 7), zlib1g (>= 1:1.1.4)
Recommends: python-mysqldb
Breaks: mariadb-rocksdb-engine-10.2, mariadb-rocksdb-engine-10.3
Replaces: mariadb-rocksdb-engine-10.2, mariadb-rocksdb-engine-10.3
Section: database
Priority: optional
Homepage: http://mariadb.org/
Description: RocksDB storage engine for MariaDB...

*** Contents:
drwxr-xr-x root/root         0 2018-02-23 13:42 ./
drwxr-xr-x root/root         0 2018-02-23 13:42 ./etc/
drwxr-xr-x root/root         0 2018-02-23 13:42 ./etc/mysql/
drwxr-xr-x root/root         0 2018-02-23 13:42 ./etc/mysql/mariadb.conf.d/
-rw-r--r-- root/root        40 2018-02-23 13:42 ./etc/mysql/mariadb.conf.d/rocksdb.cnf
drwxr-xr-x root/root         0 2018-02-23 13:42 ./usr/
drwxr-xr-x root/root         0 2018-02-23 13:42 ./usr/bin/
-rwxr-xr-x root/root     24704 2018-02-23 11:01 ./usr/bin/myrocks_hotbackup
-rwxr-xr-x root/root   4103224 2018-02-23 13:42 ./usr/bin/mysql_ldb
-rwxr-xr-x root/root   4094992 2018-02-23 13:42 ./usr/bin/sst_dump
drwxr-xr-x root/root         0 2018-02-23 13:42 ./usr/lib/
drwxr-xr-x root/root         0 2018-02-23 13:42 ./usr/lib/mysql/
drwxr-xr-x root/root         0 2018-02-23 13:42 ./usr/lib/mysql/plugin/
-rw-r--r-- root/root   5298792 2018-02-23 13:42 ./usr/lib/mysql/plugin/ha_rocksdb.so
drwxr-xr-x root/root         0 2018-02-23 13:42 ./usr/share/
drwxr-xr-x root/root         0 2018-02-23 13:42 ./usr/share/doc/
drwxr-xr-x root/root         0 2018-02-23 13:42 ./usr/share/doc/mariadb-plugin-rocksdb/
-rw-r--r-- root/root       513 2018-02-23 13:42 ./usr/share/doc/mariadb-plugin-rocksdb/changelog.gz
-rw-r--r-- root/root      2501 2018-02-23 11:01 ./usr/share/doc/mariadb-plugin-rocksdb/copyright
```

Compare with Percona Server's package

```
Package: percona-server-rocksdb-5.7
Source: percona-server-5.7
Version: 5.7.21-20-1.trusty
Architecture: amd64
Maintainer: Percona Server Development Team <mysql-dev@percona.com>
Installed-Size: 132079
Depends: percona-server-server-5.7 (= 5.7.21-20-1.trusty)
Section: database
Priority: extra
Homepage: http://www.percona.com/software/percona-server/
Description: MyRocks storage engine plugin for Percona Server
 .
 MyRocks is a storage engine for Percona Server which incorporates RocksDB
 library optimized for fast storage and space efficiency.
 .
 This package includes the MyRocks/RocksDB plugin library.

*** Contents:
drwxr-xr-x root/root         0 2018-02-16 14:00 ./
drwxr-xr-x root/root         0 2018-02-16 14:00 ./usr/
drwxr-xr-x root/root         0 2018-02-16 14:00 ./usr/share/
drwxr-xr-x root/root         0 2018-02-16 14:00 ./usr/share/doc/
drwxr-xr-x root/root         0 2018-02-16 14:00 ./usr/share/doc/percona-server-rocksdb-5.7/
-rw-r--r-- root/root      1959 2018-02-16 10:57 ./usr/share/doc/percona-server-rocksdb-5.7/copyright
-rw-r--r-- root/root       781 2018-02-16 12:35 ./usr/share/doc/percona-server-rocksdb-5.7/changelog.Debian.gz
drwxr-xr-x root/root         0 2018-02-16 14:01 ./usr/bin/
-rwxr-xr-x root/root   5539400 2018-02-16 14:01 ./usr/bin/sst_dump
-rwxr-xr-x root/root   5547592 2018-02-16 14:01 ./usr/bin/ldb
-rwxr-xr-x root/root   5551688 2018-02-16 14:01 ./usr/bin/mysql_ldb
drwxr-xr-x root/root         0 2018-02-16 13:57 ./usr/lib/
drwxr-xr-x root/root         0 2018-02-16 13:57 ./usr/lib/mysql/
drwxr-xr-x root/root         0 2018-02-16 14:01 ./usr/lib/mysql/plugin/
-rw-r--r-- root/root   7095880 2018-02-16 14:01 ./usr/lib/mysql/plugin/ha_rocksdb.so
drwxr-xr-x root/root         0 2018-02-16 13:57 ./usr/lib/mysql/plugin/debug/
-rw-r--r-- root/root 111465379 2018-02-16 13:03 ./usr/lib/mysql/plugin/debug/ha_rocksdb.so
```

- No dependencies
- They bundle lz4 and zstd

MariaDB

Plan

- What MyRocks is
- How it is provided in upstream
- Putting MyRocks into MariaDB
- Packaging MyRocks
- MyRocks for non-myrocks users
  - Data loading
  - Replication
  - Backup

Plan

- MyRocks for non-myrocks users
  - Data loading
  - Replication
  - Backup

Data loading – good news

- It’s a write-optimized storage engine
- The same data takes less space on disk
- Data loading is faster
- …
- Do not have my own benchmark data, yet
  - See Facebook’s talks

Data loading – bad news

- Limitation: Transaction must fit in memory

```
mysql> ALTER TABLE big_table ENGINE=RocksDB;
ERROR 2013 (HY000): Lost connection to MySQL server during query
```

- Need to use special settings for loading data

```
mysql> set rocksdb_bulk_load=1;
```

- See https://github.com/facebook/mysql-5.6/wiki/data-loading
- Some settings make behavior non-transactional

Safety settings

- Avoid run-away memory usage and OOM killer:

```
mysql> set rocksdb_max_row_locks=10000;

mysql> alter table t10 engine=rocksdb;

ERROR 4067 (HY000): Status error 10 received from RocksDB: Operation
aborted: Failed to acquire lock due to max_num_locks limit
```

- This is useful after data loading, too.

Plan

- MyRocks for non-myrocks users
  - Data loading
  - Replication
  - Backup

MyRocks only supports Row-Based

- MyRocks' highest isolation level is "snapshot isolation", that is REPEATABLE-READ
- Statement-Based Replication: slave will run statements sequentially (serializable).
- Because of this, MyRocks doesn't support SBR.
- Row-Based Replication must be used.

Gap Lock Detector

- InnoDB supports SBR due to having “Gap Locking”
- FB/MySQL has “Gap Lock Detector”
  - Detect queries that ought to do gap locking
    - Log/fail them
- MariaDB doesn’t have it (SQL level feature)
- Percona Server does have it and always returns errors:

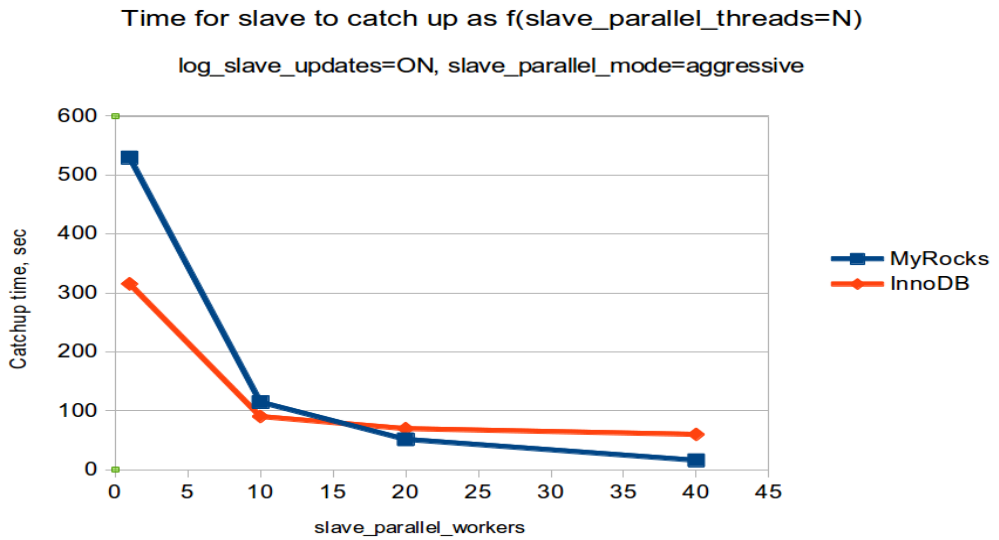
```
# log_bin=1, binlog_format=row
# optionally : set sql_log_bin=0; set rocksdb_bulk_load=1;

MySQL [test]> insert into t2 select * from t1;
ERROR 1105 (HY000): Using Gap Lock without full unique key in multi-table or multi-statement
transactions is not allowed. You need to either rewrite queries to use all unique key columns in
WHERE equal conditions, or rewrite to single-table, single-statement transaction.  Query: insert
into t2 select * from t1
```

Parallel replication

- FB/MySQL-5.6 is based on MySQL 5.6
  - Parallel replication is for different databases
- MariaDB has more advanced parallel slave
  - Conservative (group-commit based)
  - Optimistic (rolling back conflicting transactions)

Conservative mode works

- MariaDB's Group Commit works on the master
- Slave runs in parallel
- Note: different write path depending on log_slave_updates
  - ON: Does XA with slave binlog
  - OFF: Commit in order, in parallel
- Both now work, but different under the hood.



Optimistic parallel replication

- Requires support for ‘High-priority transactions”:
  - We apply trx1
  - trx1 needs a lock that trx2 is holding?
  - Roll back trx2.
- MyRocks doesn’t provide this feature
  - Can run with slave_parallel_mode=optimistic
  - But it does not provide [much] advantage over conservative.

Background: mysql.gtid_slave_pos

- **`mysql.gtid_slave_pos`** stores slave's position
- Store it in a transactional storage engine
  - After crash, we know the relay log position that matches the data
  - It's a crash-safe slave
- **`mysql.gtid_slave_pos`** uses a different engine?
  - Cross-engine transaction (slow).

Per-engine mysql.gtid_slave_pos

- The idea:
  - Have **`mysql.gtid_slave_pos_${engine}`** for each engine
  - Slave position is the biggest position in all tables.
  - Transaction affecting only MyRocks will only touch **`mysql.gtid_slave_pos_rocksdb`**
- Configuration:
  `--gtid-pos-auto-engines=engine1,engine2,...`

Per-engine mysql.gtid_slave_pos

- Available in MariaDB 10.3
- Thanks for the patch to
  - Kristian Nielsen (implementation)
  - Booking.com (request and funding)
- In MariaDB 10.2:
  - ALTER TABLE mysql.gtid_slave_pos ENGINE=RocksDB;

Special replication modes

- Read-Free Replication
  - FB/MySQL has it
  - Percona Server has it (for TokuDB initially)
  - MariaDB (currently) doesn't
- rpl_skip_tx_api
  - Server-level feature in FB/MySQL
  - Percona Server: MyRocks-specific port
  - MariaDB – doesn't have it
- Master-skip-tx-api
  - Only FB/MySQL has it (recent addition).

Plan

- MyRocks for non-myrocks users
  - Data loading
  - Replication
  - Backup

Backup for MyRocks

- FB/MySQL
  - Includes myrocks_hotbackup
- Percona Server
  - Doesn’t include it, points to FB’s myrocks_hotbackup
- MariaDB
  - Includes a [slightly] modified myrocks_hotbackup
  - Mariabackup doesn’t support MyRocks [yet?]

myrocks_hotbackup under the hood

- Operation
  - Take a RocksDB snapshot (hard link the sst files)
    - Transfer it to backup destination
  - Copy the .frm and other supplementary files
    - Don’t copy InnoDB and other files
- For the user
  - Works on a controlled MyRocks-only instance
  - Not user-friendly.

Conclusions

Conclusions

- MyRocks is RC in MariaDB
  - Aiming for GA soon
- MariaDB's improvements
  - Proper packages
  - Conservative Parallel slave
  - Per-engine mysql.gtid_slave_pos (10.3)
- Using MyRocks
  - Can get better space/write efficiency
    - And performance for write-heavy workloads
  - However one has to use special settings.

Thanks!