Galera 4

in

MariaDB 10.4

And a little bit in MySQL

Seppo Jaakola
Codership
Seppo Jaakola

- CEO Codership
- Developer role, 15 yrs with MySQL engineering

Background:
- DBMS Engineering
- Data Security
Codership

- Technology Company, focus in R&D
- Since 2007
- Replication for Open Source Databases
- Galera Cluster
- Business through MySQL / MariaDB Support & Consulting
- Stable continuous growth, team growing 3 → 20
Agenda

- Galera Cluster Overview
- Galera 4 in 10.4 Features
- Upgrading 10.3 Cluster to 10.4
- Galera 4 Streaming Replication
- Additional Galera 4 Feature Road Map
Galera Replication

- Generic Replication Plugin for database servers
- Uses Replication API to interact with DBMS (wsrep API project in github)
- DBMS and Galera plugin must have same wsrep API version
Galera Replication Versions

- Major versions 1..2..3
- Current production head version 3.26
- wsrep API versions 1..25
- Next major version is Galera 4 and using wsrep API #26
- wsrep API change requires rolling upgrade path
Galera 4 in MariaDB 10.4

- MariaDB 10.4 RC has Galera 4 Replication
- wsrep API #26
  - Impacts upgrading
- wsrep patch integration in 10.4 codebase
  - Impacts code stability
  - bug fix & new features turnover
Galera 4 in MariaDB 10.4

- Streaming Replication
  - Support for large transactions
  - Platform for other new features
- Group commit support
- 10.4 Backup locks
  - mariabackup SST with “light weight lock”
Streaming Replication

- Originally developed for supporting huge transactions
- In Galera 3, transaction processes in master node until commit time
- For large transactions, the write size will be big, and is hard to handle
- There are means to prevent too large transactions
  - `wsrep_max_ws_size`
Streaming Replication

- Streaming replication is new technology developed for Galera Replication 4 to enable running transaction of unlimited size in cluster
- Transaction size limits will remain, and cluster can still reject too large transactions
Streaming Replication

- Transaction is replicated, gradually in small fragments, during transaction processing
  - i.e. before actual commit, we replicate a number of small size fragments
- Size threshold for fragment replication is configurable
- Replicated fragments are applied in slave threads preserving transaction’s state in all cluster nodes
  - Fragments hold locks in all nodes and cannot be conflicted later
Streaming Replication

Huge transaction

Begin
Update, update, update....

Node A

Trx

Node B

Galera Replication
Streaming Replication

Huge transaction

Node A

Node B

Update, update, update....

Trx

SR Trx

WS

Galera Replication
Streaming Replication

Huge transaction

Update, update, update....

Node A

Node B

Trx

SR Trx

WS

Galera Replication
Streaming Replication

Huge transaction

commit

Node A

Node B

Trx

SR Trx

WSC

Galera Replication
Streaming Replication

Huge transaction

OK

Node A

Node B

Galera Replication
Fragment Applying

SR transaction pool

- SR#1 THD
- SR#2 THD
- SR#n THD

WS
- SR trx :2
- CF: 0

certification

applier

apply

applier

applier
Fragment Applying

SR transaction pool

WS
SR trx :2
CF: 0

certification

applier

apply

applier

Pull THD

applier

SR#1 THD

SR#2 THD

SR#n THD
Fragment Applying

SR transaction pool

WS
SR trx :2
CF: 0

SR#2 THD

ev→apply_event()
...
ev->apply_event()
wsrep_SR_store->append_frag_apply()
Fragment Applying

SR transaction pool

applier

WS
SR trx :2
CF: 0

OK

Push THD

SR#1 THD
SR#2 THD
SR#n THD
Fragment Committing

SR transaction pool

SR#1 THD
SR#2 THD
SR#n THD

applier
applier
applier

commit
Pull THD

WS
SR trx :2
CF: 1
certification

www.galeracluster.com
Fragment Committing

SR transaction pool

trans_commit()
wsrep_SR_store->append_frag_commit()

commit

WS
SR trx :2
CF: 1

SR#2 THD

applier

SR#1 THD

SR#n THD

WS
SR trx :2
CF: 1

applier

www.galeracluster.com
Fragment Committing

SR transaction pool

applier

SR#1 THD

SR#1nTHD

WS
SR trx :2
CF: 1

OK
## Configuring Streaming Replication

### Session variables and can be dynamically set

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
</table>
| `wsrep_trx_fragment_unit` | Unit metrics for fragmenting, options are:  
  - bytes: WS size in bytes  
  - rows: # of rows modified  
  - statements: # of SQL statements issued |
| `wsrep_trx_fragment_size` |  
  - Threshold size (in units), when fragment will be replicated  
  - 0 = no streaming |
Using Streaming Replication

- Due to excessive logging and elevated replication overhead, streaming replication will cause degraded transaction throughput rate.
- Best use case is to use streaming replication for cutting large transactions.
- Set fragment size to ~10K rows.
- Fragment variables are session variables and can be dynamically set.
- Intelligent application can set streaming replication on/off on need basis.
### wsrep Tables in mysql database

```
MariaDB [(none)]> show tables in mysql like 'wsrep%';
+--------------------------------------------------+
| Tables_in_mysql (wsrep%)                          |
+--------------------------------------------------+
| wsrep_cluster                                    |
| wsrep_cluster_members                            |
| wsrep_streaming_log                              |
+--------------------------------------------------+
3 rows in set (0.005 sec)
```
wsrep Tables in mysql database

```sql
MariaDB [(none)]> select * from mysql.wsrep_cluster\G
*************************** 1. row ***************************
cluster_uuid: 0be6f4d6-35da-11e9-b8a0-d6501fb08579
    view_id: 2
    view_seqno: 2
protocol_version: 4
    capabilities: 184703
1 row in set (0.005 sec)
```
### wsrep Tables in mysql database

```
MariaDB [(none)]> select * from mysql.wsrep_cluster_members;
*************************** 1. row ***************************
    node_uuid: ea306990-35b0-11e9-9841-366f30fba24f
    cluster_uuid: ea317655-35b0-11e9-9ac4-9f27c3510851
    node_name: labrador
node_incoming_address: 127.0.0.1:16000
*************************** 2. row ***************************
    node_uuid: eb29a68c-35b0-11e9-ac23-e2418aae257d
    cluster_uuid: ea317655-35b0-11e9-9ac4-9f27c3510851
    node_name: poodle
node_incoming_address: 127.0.0.1:16001
*************************** 3. row ***************************
    node_uuid: eb29e56c-35b0-11e9-a328-fe229f8ae4cb
    cluster_uuid: ea317655-35b0-11e9-9ac4-9f27c3510851
    node_name: rotweiler
node_incoming_address: 127.0.0.1:16002
3 rows in set (0.006 sec)
```
### wsrep Tables in mysql database

```sql
MariaDB [(none)]> show create table mysql.wsrep_streaming_log
*************************** 1. row ***************************
Table: wsrep_streaming_log
Create Table: CREATE TABLE `wsrep_streaming_log` ( 
  `node_uuid` char(36) NOT NULL,
  `trx_id` bigint(20) NOT NULL,
  `seqno` bigint(20) NOT NULL,
  `flags` int(11) NOT NULL,
  `frag` longblob NOT NULL,
  PRIMARY KEY (`node_uuid`, `trx_id`, `seqno`)
) ENGINE=InnoDB DEFAULT CHARSET=latin1
```

1 row in set (0.004 sec)
Rolling Upgrade
Galera Rolling Upgrades

Galera Replication

MariaDB 10.3
Galera 3

MariaDB 10.3
Galera 3

MariaDB 10.3
Galera 3

Wsrep-API 25

www.galeracluster.com
Galera Rolling Upgrades

MariaDB 10.3
Galera 3
read & write

MariaDB 10.3
Galera 3
read & write

MariaDB 10.4
Galera 4
read only

Upgrade for 10.4
Isolate node
Shutdown server
Install 10.4
Start without wsrep_provider
mysql_upgrade
Shutdown server
Start with wsrep_provider
Allow read only access

Galera Replication
Wsrep-API 25
Galera Rolling Upgrades

**MariaDB 10.3**
Galera 3
read & write

**MariaDB 10.4**
Galera 4
read only

**MariaDB 10.4**
Galera 4
read only

Upgrade for 10.4
Upgrade node 2

Galera Replication
Wsrep-API 25

www.galeracluster.com
Galera Rolling Upgrades

Upgraded cluster

MariaDB 10.4
Galera 4

MariaDB 10.4
Galera 4

Wsrep-API 26 features now enabled in replication

Upgrade for 10.4
Shutdown node 1
Allow read-write
Complete node1 upgrade

www.galeracluster.com
Galera Rolling Upgrades

Galera Replication

MariaDB 10.4
Galera 4
read & write

MariaDB 10.4
Galera 4
read & write

MariaDB 10.4
Galera 4
read & write

Wsrep-API 26
Galera Rolling Upgrades

Galera Replication

- MariaDB 10.4
  - Galera 4
  - read & write

- MariaDB 10.4
  - Galera 4
  - read & write

- MariaDB 10.4
  - Galera 4
  - read & write

- MariaDB 10.3
  - Galera 3
  - Wsrep-API 26

www.galeracluster.com
4.0 Release Status

- Part of Galera 4 feature set was merged in MariaDB 10.4 beta 2 release and is now in the 10.4 RC and later releases
- Schedule of remaining Galera 4 features is not yet confirmed
Features Missed in MariaDB 10.4

- Gcache encryption
  - Data at rest encryption for full replication pipeline
- XA transaction support
  - Will be needed in sharding cluster
- Non blocking DDL
  - Co-incides with other MariaDB non-locking DDL work
- Cluster error voting
  - Galera library feature, may be in later 10.4.* releases
Data at rest encryption

RAM

Client thread

Buffer pool

gcache

DISK

ibdata

redo log

binlog
Data at rest encryption

RAM

Client thread

Buffer pool

gcache

DISK

ibdata

redo log

binlog
Data at rest encryption
Features Missed in MariaDB 10.4

- Gcache encryption
  - Data at rest encryption for full replication pipeline
- XA transaction support
  - Will be needed in sharding cluster
- Non blocking DDL
  - Co-incides with other MariaDB non-locking DDL work
- Cluster error voting
  - Galera library feature, may be in later 10.4.* releases
XA Transactions
In Galera 3
XA Transactions with Galera 3

- XA start 'foo'
- XA start 'bar'
- XA prepare 'foo'
- XA prepare 'bar'
- XA commit 'foo'
- XA commit 'bar'

ATM

MariaDB
XA Transactions with Galera 3

- Galera replication is eager to replicate in MariaDB 2PC prepare stage
- Prepared XA transaction cannot be rolled back anymore
XA Transactions with Galera 3

XA Start

Node A

Node B

XA trans

smith

smith
XA Transactions with Galera 3

XA Insert into persons ‘jones’

Node A

Node B

XA trans

Jones

Smith

Smith
XA Transactions with Galera 3

Node A

XA Prepare

Node B

XA trans

Jones

Smith

Smith
XA Transactions with Galera 3

XA Prepare

Node A

Node B

XA trans

smith

jones

smith

WS

jones
XA Transactions with Galera 3

XA Prepare

Node A

XA trans

 smith

Node B

 apply

 smith

 jones
XA Transactions with Galera 3

Node A
- smith
- XA trans
- jones

OK

Node B
- smith
- jones

www.galeracluster.com
XA Transactions with Galera 3

Node A

Node B

XA Rollback

XA trans

smith

jones

smith

jones
XA Transactions with Galera 3

Node A

Node B

Smith

Smith Jones

Inconsistency
XA prepare Safety

- XA prepare should guarantee that transaction can commit
- However, galera multi-master conflict resolving does honor XA transactions
**XA Transactions with Galera 3**

- **XA Start**
- **Node A**
- **Node B**
- **XA trans**
- **smith**
- **smith**
UPDATE smith=kit

Node A

Node B

XA trans

kit

smith

smith
XA Transactions with Galera 3

Node A

XA trans

kit

smith

Node B

UPDATE smith=hannibal

WS

kit

hannibal
XA Transactions with Galera 3

Node A

Node B

hannibal

hannibal

OK
XA Transactions with Galera 3

XA Commit

Node A

Node B

hannibal

hannibal
XA by Streaming Replication
XA Transaction Support

XA Start

Node A

Node B

XA trans

smith

smith
XA Insert into persons ‘jones’

Node A

Node B

XA trans

XA trans

XA trans

XA trans

jones

smith

smith
XA Transaction Support

- **Node A**
  - XA trans
  - Smith

- **Node B**
  - SR trans
  - Jones

**WS**

**XA Prepare**
XA Transaction Support

XA Rollback

Node A

Node B

XA trans jones

SR trans jones

smith

smith

WS rollback
XA Transaction Support

Node A

smith

Node B

smith
XA Transaction Conflict Resolution

Node A

XA trans
kit

smith

Node B

SR trans
kit

smith

UPDATE smith=hannibal

XA Prepare
Features Missed in MariaDB 10.4

- Gcache encryption
  - Data at rest encryption for full replication pipeline
- XA transaction support
  - Will be needed in sharding cluster
- Non blocking DDL
  - Co-incides with other MariaDB non-locking DDL work
- Cluster error voting
  - Galera library feature, may be in later 10.4.* releases
MariaDB 10.5

- Galera work for 10.5 is ongoing
- Feature scoping is not complete, may come with Galera 4 or new Galera 5
Galera 4 on MySQL

- Galera 4 is integrated in MySQL 5.6 and 5.7 and 8.0 versions
- Public releases after MariaDB 10.4 GA is released
Summary

- MariaDB 10.4.5 has significant new Galera 4 features
- Streaming Replication implements new replication protocol opening possibilities for many unforeseen replication features
- Other new Galera 4 related features to come with later 10.4 and 10.5 releases
- MySQL version is coming
- Rolling cluster upgrade supported and requires care
Happy Clustering :-)

Thank you for listening!
Happy Clustering :-)