Synchrone multi-Master Replikation für MySQL

DOAG SIG-MySQL 2013, München

Oli Sennhauser

Senior MySQL Consultant, FromDual GmbH

oli.sennhauser@fromdual.com

Über FromDual GmbH

- FromDual bietet neutral und unabhängig:
  - Beratung für MySQL und Galera
  - Support für MySQL und Galera
  - Remote-DBA Dienstleistungen
  - MySQL Schulungen
- Partner der Open Database Alliance (ODBA.org)
- Oracle Silver Partner (OPN)

www.fromdual.com

Inhalt

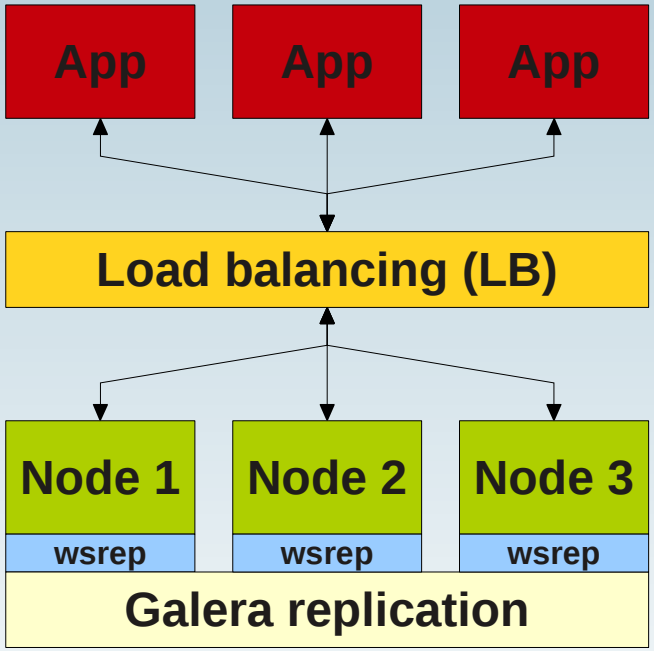
Galera Cluster

- Was ist Galera?
- Warum Galera?
- Charakteristik
- Set-up
- Konfiguration
- Starten und Stoppen
- SST
- Information

Was ist Galera Cluster

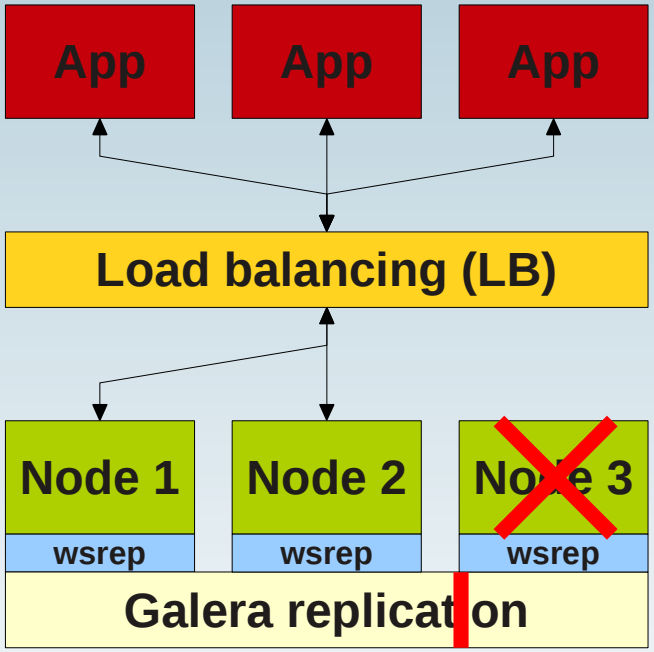
- Galera Cluster für MySQL ist:

  Synchrone multi-Master Replikation



Galera HA Cluster

- Geht ein Knoten kaputt:



Charakteristik von Galera

- Synchrone Replikation
- Basiert auf der InnoDB SE
- Echte aktiv-aktiv multi-Master Topologie
- Lesen UND Schreiben auf irgendeinen Knoten möglich
- Automatische Kontrolle der teilnehmenden Knoten
- Echtes paralleles Replizieren auf Zeileneben (RBR)
- Kein Slave-Lag
- Keine verlohrenen Transaktionen
- Lese-Skalierbarkeit (read Scale-Out!)
- Erhöhter Schreibdurchsatz + gefahrloser Einsatz von SSD
- Vorsicht vor Hotspots
- Höhere Wahrscheinlichkeit von Deadlocks
- Initiale Voll-Synchronisation blockiert 1 Knoten → 3 Knoten (`mysqldump`)

Warum Galera Cluster?

- Master-Slave Replikation
  - Nicht multi-Master, asynchron, Inkonsistenzen möglich
- Master-Master Replikation
  - Eine Art von multi-Master, asynchron, Inkonsistenzen, Konflikte
- MHA, MMM (v1, v2), Tungsten
  - Basiert auf MySQL Replikation, somit siehe dort
- MySQL Cluster
  - Nicht wie InnoDB, Know-How, Network-DB! In-Memory DB
- Aktiv/passiv failover Cluster
  - Know-How, unbeschäftigte Ressourcen
- Schooner
  - Teuer (besteht aus Galera!)

Set-up

- 3 Knoten sind empfohlen
- Oder: 2 + 1 (2 `mysqld` + `garbd`) → SST!!!
- 2 Knoten → Split-Brain!
- Codership MySQL + Galera Plug-in (wsrep)

Konfiguration

- my.cnf (galera.conf, wsrep.conf)

```
default_storage_engine          = InnoDB
binlog_format                   = row
innodb_autoinc_lock_mode        = 2
innodb_locks_unsafe_for_binlog = 1

innodb_flush_log_at_trx_commit = 0

query_cache_size                = 0
query_cache_type                = 0

wsrep_provider                  = …/lib/plugin/libgalera_smm.so

wsrep_cluster_address           = "gcomm://"

wsrep_cluster_name              = 'Galera Cluster'
wsrep_node_name                 = 'Node 1'

wsrep_sst_method                = mysqldump
wsrep_sst_auth                  = sst:secret
```

1. Knoten starten

- Start mit `wsrep_provider = none`
- Erstellen des `sst` Users:

```
GRANT ALL PRIVILEGES ON *.* TO 'sst'@'%' IDENTIFIED BY 'secret';
GRANT ALL PRIVILEGES ON *.* TO 'sst'@'localhost' IDENTIFIED BY 'secret';
```

- Neustart mit (`gcomm://` bildet neuen Cluster!):

```
wsrep_provider        = /usr/local/.../lib/plugin/libgalera_smm.so
wsrep_cluster_address = "gcomm://"
```

Snapshot State Transfer (SST)

- Initiale Synchronisation zwischen 1. und weiteren Knoten
- User für SST ist per default root!
  - Wir empfehlen: Eigenen Benutzer anlegen
- SST Methoden:

  `mysqldump`, `rsync`, (`xtrabackup`, LVM?)
- SST mit `mysqldump` und `rsync` blockiert Donor! (→ 3 Knoten)
- Mit v2.0: Incremental State Transfer (IST)

SST

- SST ist ein Push vom Donor.
- MySQL Error Log:

```
120131 16:26:42 [Note] WSREP: Quorum results:
        conf_id    = 4,
        members    = 2/3 (joined/total)
120131 16:26:44 [Note] WSREP: Node 2 (Node C) requested state transfer from '*any*'.
                              Selected 0 (Node A)(SYNCED) as donor.
120131 16:26:44 [Note] WSREP: Shifting SYNCED -> DONOR/DESYNCED (TO: 2695744)
120131 16:27:10 [Note] WSREP: 2 (Node C): State transfer from 0 (Node A) complete.
120131 16:27:10 [Note] WSREP: Member 2 (Node C) synced with group.
120131 16:27:10 [Note] WSREP: 0 (Node A): State transfer to 2 (Node C) complete.
120131 16:27:10 [Note] WSREP: Shifting DONOR/DESYNCED -> JOINED (TO: 2695744)
120131 16:27:10 [Note] WSREP: Member 0 (Node A) synced with group.
120131 16:27:10 [Note] WSREP: Shifting JOINED -> SYNCED (TO: 2695744)
120131 16:27:10 [Note] WSREP: Synchronized with group, ready for connections
```

Weitere Knoten starten

- 2. & 3. Knoten:
  - Starten mit `wsrep_provider = none`
  - Mindestens remote SST User erstellen: `'sst'@'%'`
- Neustart mit:

```
wsrep_provider        = /usr/local/.../lib/plugin/libgalera_smm.so
wsrep_cluster_address = "gcomm://192.168.42.1"
```

- Knoten führen SST durch
- Alle Schritte prüfen (auf lokalem und remote Knoten) mit:

```
SHOW GLOBAL STATUS LIKE 'wsrep_%';
```

und `tail -f error.log`

Checks


```
120131 07:37:17 mysqld_safe Starting mysqld daemon
...
120131  7:37:18 [Note] WSREP: wsrep_load(): loading provider library
                       'libgalera_smm.so'
120131  7:37:18 [Note] WSREP: Start replication
...
120131  7:37:18 [Note] WSREP: Shifting CLOSED -> OPEN (TO: 0)
120131  7:37:18 [Note] .../mysql/bin/mysqld: ready for connections.
...
120131  7:37:23 [Note] WSREP: Quorum results:
        conf_id    = 2,
        members    = 3/3 (joined/total)
```

```
SHOW GLOBAL STATUS LIKE 'wsrep%';
+----------------------------+--------------------------------------+
| Variable_name              | Value                                |
+----------------------------+--------------------------------------+
| wsrep_local_state_comment  | Synced (6)                           |
| wsrep_cluster_conf_id      | 3                                    |
| wsrep_cluster_size         | 3                                    |
| wsrep_cluster_status       | Primary                              |
| wsrep_connected            | ON                                   |
| wsrep_local_index          | 0                                    |
| wsrep_ready                | ON                                   |
+----------------------------+--------------------------------------+
```

Knoten neu starten

- 2. und 3. Knoten → kein Problem
- 1. Knoten bildet neuen Cluster:

```
wsrep_cluster_address = "gcomm://
```

- Umkonfiguration erforderlich! :-(
- Daher anschliessend alle Knoten:

```
wsrep_cluster_address = "gcomm://192.168.0.1,192.168.0.2?pc.wait_prim=no"
```

Variablen

- Zur Zeit (2.2) 33 Variablen

```
SHOW GLOBAL VARIABLES LIKE 'wsrep%';
+-------------------------------+------------------------------+
| Variable_name                 | Value                        |
+-------------------------------+------------------------------+
| wsrep_cluster_address         | gcomm://                     |
| wsrep_cluster_name            | Galera-2.2 wsrep-27          |
| wsrep_max_ws_rows             | 131072                       |
| wsrep_max_ws_size             | 1073741824                   |
| wsrep_node_incoming_address   | 192.168.42.1:3306            |
| wsrep_node_name               | Node 1                       |
| wsrep_notify_cmd              |                              |
| wsrep_on                      | ON                           |
| wsrep_provider                | .../plugin/libgalera_smm.so  |
| wsrep_retry_autocommit        | 1                            |
| wsrep_slave_threads           | 1                            |
| wsrep_sst_auth                | ********                     |
| wsrep_sst_donor               |                              |
| wsrep_sst_method              | mysqldump                    |
| wsrep_sst_receive_address     | AUTO                         |
+-------------------------------+------------------------------+
```

wsrep_provider_options

- evs.debug_log_mask = 0x1;
- evs.inactive_check_period = PT0.5S
- evs.inactive_timeout = PT15S;
- evs.info_log_mask = 0;
- evs.install_timeout = PT15S;
- evs.join_retrans_period = PT0.3S;
- evs.keepalive_period = PT1S;
- evs.max_install_timeouts = 1;
- evs.send_window = 4;
- evs.stats_report_period = PT1M;
- evs.suspect_timeout = PT5S;
- evs.use_aggregate = true;
- evs.user_send_window = 2;
- evs.version = 0;
- evs.view_forget_timeout = PT5M;
- gcache.dir = ...;
- gcache.keep_pages_size = 0;
- gcache.mem_size = 0;
- gcache.name = .../galera.cache;
- gcache.page_size = 128M;
- gcache.size = 128M;
- gcs.fc_debug = 0;
- gcs.fc_factor = 0.5;
- gcs.fc_limit = 16;
- gcs.fc_master_slave = NO;
- gcs.max_packet_size = 64500;
- gcs.max_throttle = 0.25;
- gcs.recv_q_hard_limit = 9223372036854775807;
- gcs.recv_q_soft_limit = 0.25;
- gmcast.listen_addr = tcp://127.0.0.1:4567;
- gmcast.mcast_addr = ;
- gmcast.mcast_ttl = 1;
- gmcast.peer_timeout = PT3S;
- gmcast.time_wait = PT5S;
- gmcast.version = 0;
- pc.checksum = true;
- pc.ignore_quorum = false;
- pc.ignore_sb = false;
- pc.linger = PT2S;
- pc.npvo = false;
- pc.version = 0;
- protonet.backend = asio;
- protonet.version = 0;
- replicator.commit_order = 3

Status

- Zur Zeit (2.2) 40 Status Informationen

```
SHOW GLOBAL STATUS LIKE 'wsrep%';
```

  - Cluster Status
  - Performance Metriken
  - Allgemeine Informationen

```
+----------------------------+--------------------------------------+
| Variable_name              | Value                                |
+----------------------------+--------------------------------------+
| wsrep_last_committed       | 2695744                              |
| wsrep_replicated           | 1                                    |
| wsrep_replicated_bytes     | 576                                  |
| wsrep_received             | 9                                    |
| wsrep_received_bytes       | 1051                                 |
| wsrep_local_commits        | 1                                    |
| wsrep_local_send_queue     | 0                                    |
| wsrep_local_recv_queue     | 0                                    |
| wsrep_flow_control_sent    | 0                                    |
| wsrep_flow_control_recv    | 0                                    |
| wsrep_provider_version     | 22.1.1(r95)                          |
+----------------------------+--------------------------------------+
```

Betrieb

- 2 Modi:
  - Master-Master (Schreiben auf alle Knoten)
  - Master-Slave (Schreiben nur auf 1 Knoten)
- Initiale Konfiguration (etwas mühselig)
- SST (DB Grösse, NW Bandbreite (WAN))
- Rolling Restart mit HW/OS/DB Upgrade!
- Deadlocks Und Hot Spots
- Forcieren eines SST: `rm grastate.dat`

Verschiedene Szenarien

- Initialer Cluster Start (neuer Cluster)
- Initialer Knoten Start (erfordert SST)
- Knoten Neustart (erfordert IST)
- Rolling Restart (z.B. für Upgrades)
- Cluster Neustart (bestehender Cluster)

Cluster Neustart

- Mit dieser Konfiguration:

```
wsrep_cluster_address = "gcomm://192.168.0.1,192.168.0.2?pc.wait_prim=no"
```

- Cluster verbleib im Status non-Primary:

```
SHOW GLOBAL STATUS LIKE 'wsrep_cluster_status';
+----------------------+-------------+
| Variable_name        | Value       |
+----------------------+-------------+
| wsrep_cluster_status | non-Primary |
+----------------------+-------------+
```

- Cluster in Primary Status versetzen:

```
SET GLOBAL wsrep_provider_options='pc.bootstrap=1';
```

Load Balancing

- In der Applikation (selber bauen!)
- Connectors
  - Connector/J
  - PHP: MySQLnd Replikations- und Load Balancing Plug-in
- SW Load Balancer
  - GLB, Pen, LVS, HAProxy, MySQL Proxy, SQL Relay,
- HW Load Balancer

Q & A

Fragen ?

Diskussion?

Wir haben Zeit für ein persönliches Gespräch...

- **FromDual bietet neutral und unabhängig:**
  - **Beratung**
  - **Remote-DBA**
  - **Support für MySQL, Galera, Percona Server und MariaDB**
  - **Schulung**

www.fromdual.com/presentations