Deploying Galera Cluster in the real world

Colin Charles, Consultant, Galera Cluster
colin.charles@galeracluster.com | byte@bytebot.net
https://bytebot.net/blog | @bytebot on Twitter
FOSDEM, Brussels, Belgium
5 February 2023

whoami

- Consultant at Codership, makers of Galera Cluster
- Active in the MySQL ecosystem: Founding team of MariaDB Server (2009-2016), early at MySQL AB (pre-Sun exit), Percona.
- Past lives include Fedora Project (FESCO), OpenOffice.org
- MySQL Community Contributor of the Year Award winner 2014

<ad> Codership </ad>

- Codership are the original makers and engineers of Galera Cluster, a multi-master, virtually synchronous replication solution for the MySQL ecosystem
- If you use Percona XtraDB Cluster (PXC) or MariaDB Galera Cluster, you directly benefit from the work done by the team at Codership
- Remember that beyond engineering, you may also purchase **24/7 support, training, consulting, Galera Cluster Enterprise Edition (EE)** and a whole lot more!
- Codership sponsored my travel to FOSDEM 2023

What is Galera Cluster?

- Can be described as *virtually* synchronous replication
- High Availability with no data loss, and consistent data across all nodes — no Single Point of Failure (SPoF)
- Quorum based failure handling
- Optimistic concurrency control
- 100% multi-primary (multi-master) cluster (all nodes are equal in terms of having the same data, no lagging secondaries, 24/7 availability, etc.)
  - This is a core feature of the product by design, has automatic transaction conflict detection and management, and your application can issue any transaction to any Galera Cluster node. Works well in WAN/Clouds
  - You do not need automatic failovers via a framework, no need to designate single nodes for writes and the rest for reads, configuration is simple, easier handling of scheduled downtime
- Parallel replication
- Thousands of users in various industries: e-commerce, betting/gambling, telecoms, banking, insurance, gaming, healthcare, media, marketing, advertising, travel, education, SaaS, PaaS, IaaS, etc.

Virtually synchronous replication does come with trade-offs

It can never be as fast as asynchronous (just commit to primary) or semi-synchronous (primary and one secondary) replication. We can’t beat the laws of physics when you have to write to a 3-node minimum.

Picking a distribution

- Codership Galera Cluster (upstream)
  - Based on MySQL 5.7 (galera3) and MySQL 8.0 (galera4), now with CLONE SST
- MariaDB Galera Cluster
  - Base is MariaDB Server, and it is included since 5.5.29 (separate: Mar 2013), and part of MariaDB since 10.1.8 (GA: Oct 2015). All the features of MariaDB, e.g. Oracle support, SEQUENCES, system versioned tables, optimiser features, etc. First to get Galera 4 (10.4)!
- Percona XtraDB Cluster (PXC)
  - Base is Percona Server, comes with ProxySQL, “Strict Mode”, e.g. disallow MyISAM, tables without primary keys, ROW binlog_format, logging to a file (not tables), `innodb_autoinc_lock_mode` set to 2, etc. Automatic configuration of SSL encryption (`pxc-encrypt-cluster-traffic` variable)

Some feature highlights of Galera 3.x to 4

- Intelligent donor selection, preferring a donor that can do an Incremental State Transfer (IST)
- Cluster crash recovery, say, after a power failure via `pc.recovery=ON` so all nodes maintain cluster information persistently, not requiring a bootstrap
- GTID compatibility (whether in MariaDB Server or MySQL)
- Improved foreign key support
- New mysql.* tables: wsrep_cluster, wsrep_cluster_members, wsrep_streaming_log

A bit more of Galera 4

- Streaming replication, which replicates transactions of any size; transaction replicated in small increments (query configurable). Huge transaction support!
  - `wsrep_trx_fragment_unit` — unit metrics for fragmenting, options are bytes (writesets in bytes), rows (number of rows modified), statements (number of SQL statements issued)
  - `wsrep_trx_fragment_size` — threshold size in units when fragments will be replicated. 0 means no streaming.
- Better handling of poor networks — a node will always attempt to leave the cluster gracefully if it is not possible to recover from errors without sacrificing data consistency

Enterprise only features (sorry, FOSDEM)

MariaDB, Codership Galera Cluster, some open in Percona

- Enterprise only features: non-blocking DDL
  - `wsrep_osu_method: TOI | RSU | NBO`
- `gcache` encryption to ensure a fully encrypted data directory
- XA transaction support
- Black Box (buffered error logging in Percona Server)

Biggest hurdle to upgrade to Galera 4: “We don’t want to migrate to MySQL 8”

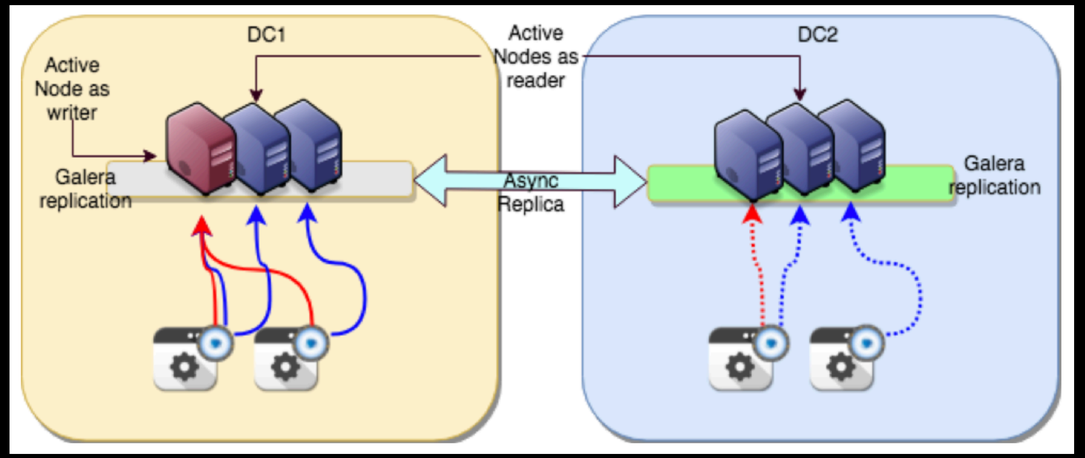
Well, get real, MySQL 5.7 will EOL in October 2023

Common setups

- 3 Galera Cluster nodes, in one data centre
- 9 Galera Cluster nodes, in three data centres (recommended*)
  - all database operations are local (segmented — `gmcast.segment`). Flow control fully configurable. Latency penalty as minimal as possible (until COMMIT). Encryption! Can also work with asynchronous replication.

* — geo-distributed Galera Clusters

- Marco Tusa — https://www.percona.com/blog/2018/11/15/how-not-to-do-mysql-high-availability-geographic-node-distribution-with-galera-based-replication-misuse/ AND https://www.percona.com/blog/2018/11/15/mysql-high-availability-on-premises-a-geographically-distributed-scenario/ AND http://www.tusacentral.net/joomla/index.php/mysql-blogs/238-260-thousands-thanks (check out the YouTube video: https://www.youtube.com/watch?v=3rGrFgbpW04)



Quorum Components

https://galeracluster.com/library/documentation/weighted-quorum.html

Realistic common setups we see…

Managing trade-off's

- 2 node Galera Cluster
  - https://galeracluster.com/library/kb/two-node-clusters.html
- 3-node Galera Cluster across 2 data centres
- 3-node Galera Cluster across 3 data centres
- 5-node cluster spread across 2 data centres
- 7-node cluster in one data centre, with 4 asynchronous secondaries hanging off one

my.cnf [galera]

```
binlog_format=ROW

default-storage-engine=innodb

innodb_autoinc_lock_mode=2

bind-address=0.0.0.0

wsrep_on=ON

wsrep_provider=/usr/lib64/galera-4/libgalera_smm.so

wsrep_cluster_name="galera"

wsrep_cluster_address="gcomm://
188.166.179.177,165.22.50.152,165.22.49.92,159.65.94.184,206.189.117.122,206.189.31.7,143.244.180.78,143.1
98.151.217,143.110.235.12"

wsrep_provider_options="gmcast.segment=1"

wsrep_sst_method=rsync

wsrep_node_address="165.22.49.92"
```

Some configuration thoughts

- distinct `gmcast.segment` for each data centre
- increase replication windows
- increase timeouts above max RTT
- look at flow control
- you can also use a dedicated Galera Cluster network
- things to pay attention to: flow control (`gcs.fc_limit`, `gcs.fc_master_slave=yes`), `repl.causal_read_timeout=PT5S`, `evs.*`

More configuration thoughts

- Set your `gcache.size` — https://galeracluster.com/library/kb/customizing-gcache-size.html
- Retry: `wsrep_retry_autocommit=5`
- `wsrep_certify_nonPK=1` (really, use Primary Key's)
- `innodb-force-primary-key=1` (https://mariadb.com/docs/server/ref/mdb/cli/mariadbd/innodb-force-primary-key/)
- `wsrep_replicate_myisam=0`
  - The `wsrep_mode` system variable, for turning on WSREP features which are not part of default behavior (including the experimental Aria replication) https://mariadb.com/kb/en/galera-cluster-system-variables/#wsrep_mode

Galera Cluster nodes, with a Galera Arbitrator (garbd) node

- Galera Arbitrator is a member of a cluster that participates in voting, but not in the actual replication
- If you have access to a 3rd data centre, or put a one-node `garbd` in your DR site, you could also have a 2-paired cluster in 2 DCs, thus bringing your node count to a mere 7 nodes (instead of 9)
- When you have an even number of nodes, `garbd` functions as an odd node, to avoid split-brain situations. It can also request a consistent application state snapshot, which help with backups
- https://galeracluster.com/library/documentation/arbitrator.html

Proxies

- Galera Load Balancer (GLB)
- HAProxy
- ProxySQL
- MariaDB MaxScale

Backups + provisioning new nodes

- Take a backup from an asynchronous secondary
- Percona XtraBackup (`xtrabackup-v2`)
- Mariabackup
- CLONE SST

Common setup, runtime issues

- SELinux messing with your setup?
- Firewall ports not open?
- Can’t get IST and always getting SST? Check the ports!
- Consider avoiding long running queries, MySQL has `max_execution_time`, MariaDB has an enhanced `KILL`, etc.
- DNS got you down? Versus IP…
- Are you allowed a maintenance window for upgrades?
- Don’t start/restart 2 nodes at the same time :-)

wsrep functions for developers

- `WSREP_LAST_SEEN_GTID()` — returns GTID for last written transaction observed by client
- `WSREP_LAST_WRITTEN_GTID()` — returns GTID of the last write transaction made by the client
- `WSREP_SYNC_WAIT_UPTO_GTID()` — blocks the client until the node applies and commits the given transaction

Wide adoption

- Nextcloud
- PowerDNS
- OpenStack
- Plenty of Kubernetes operators
- Plenty of Docker images

Still things to improve on

- MariaDB 11.0 brings `wsrep_provider_options` to be split into options (good! better for automation) https://jira.mariadb.org/browse/MDEV-22570
- More granularity, e.g. monitoring why you see `gcache.page.000001` files (how is the RingBuffer utilisation?)
- Should encryption be turned on, out of the box, like Percona XtraDB Cluster (PXC) 8?
- Make schema changes, upgrades, easier. Focus on Galera Manager — https://galeracluster.com/galera-mgr/

Further reading

- https://galeracluster.com/library/documentation/managing-fc.html
- https://galeracluster.com/library/documentation/auto-eviction.html
- https://galeracluster.com/library/documentation/backup-cluster.html
- https://galeracluster.com/library/training/tutorials/geo-distributed-clusters.html
- https://galeracluster.com/2021/05/galera-streaming-replication-feature-deep-dive-video-and-blog/

Thank You!

Colin Charles
colin.charles@galeracluster.com / byte@bytebot.net
http://bytebot.net/blog | @bytebot on twitter