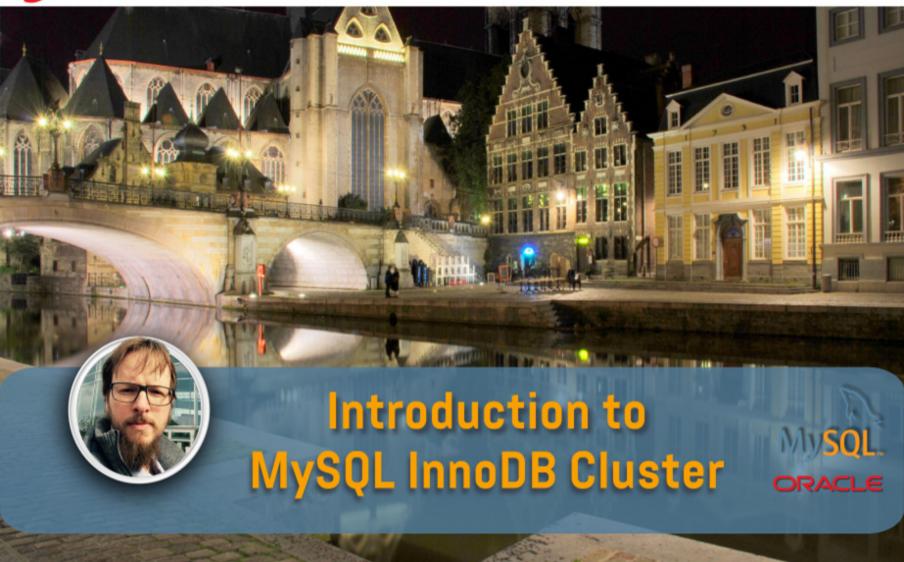




Co.Station Ghent Oktrooiplein 1, 9000 Ghent



ORACLE®



MySQL High Availability made easy

Percona University, Ghent, Belgium June 2017

Frédéric Descamps - MySQL Community Manager - Oracle

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purpose only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied up in making purchasing decisions. The development, release and timing of any features or functionality described for Oracle's product remains at the sole discretion of Oracle.



about.me/lefred

Who am 1?



Frédéric Descamps

- @lefred
- MySQL Evangelist
- Hacking MySQL since 3.23
- devops believer
- MySQL Community Manager since May
 2016
- living in Belgium **B E**





MySQL InnoDB Cluster

Easy MySQL High Availability



Ease-of-Use

Built-in HA



Out-of-Box Solution

Everything Integrated

Extreme Scale-Out

High Performance

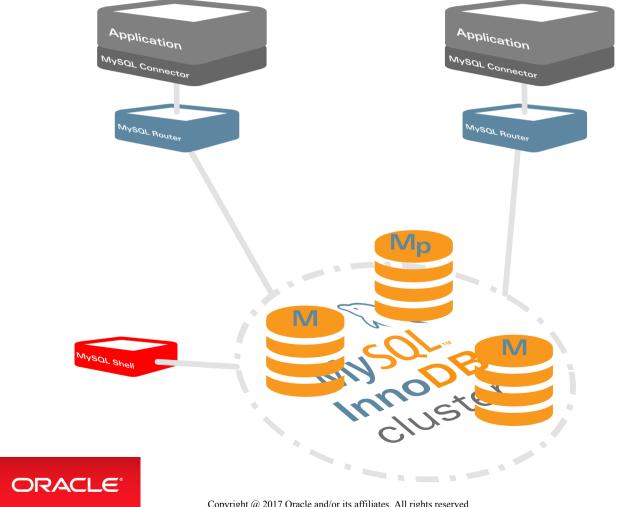


Our vision in 4 steps

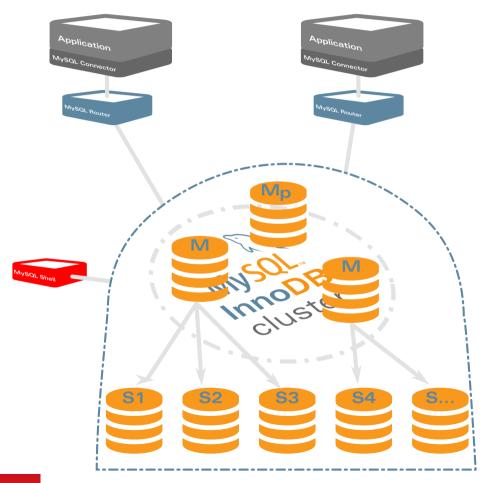




Step 2's Architecture

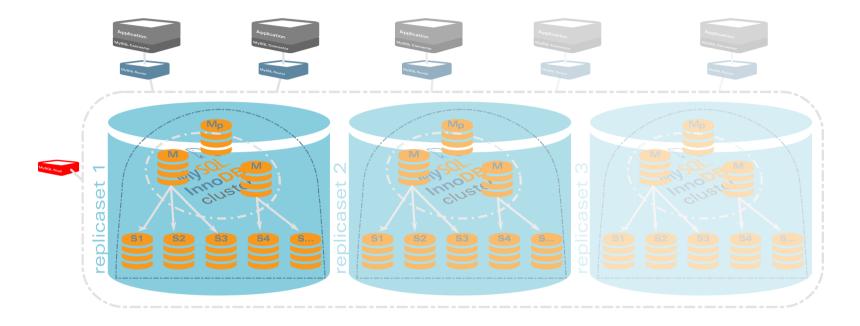


Step 3's Architecture



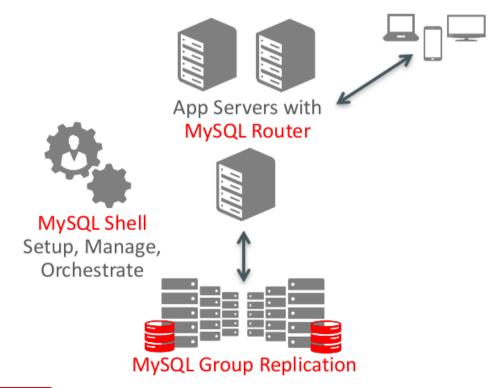


Step 4's Architecture



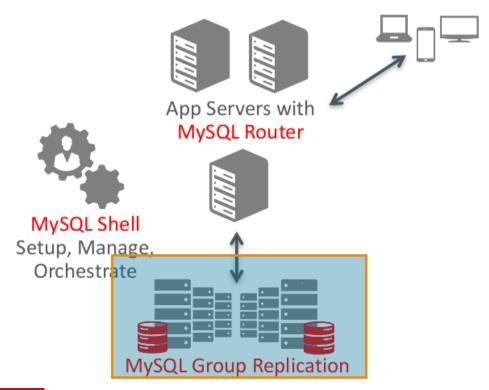


Group Replication: heart of MySQL InnoDB Cluster





Group Replication: heart of MySQL InnoDB Cluster





But before going further...





What is

High Availability?



High Availability

High availability is a characteristic of a system, which aims to ensure an agreed level of operational performance, usually uptime, for a higher than normal period.

There are three principles of systems design in reliability engineering which can help achieve high availability:

- Elimination of single points of failure. This means adding redundancy to the system so that failure of a component does not mean failure of the entire system.
- Reliable crossover. In redundant systems, the crossover point itself tends to become a single point of failure. Reliable systems must provide for reliable crossover.
- Detection of failures as they occur. If the two principles above are observed, then a user may never see a failure. But the maintenance activity must.



Source: Wikipedia

How to achieve it?



How to achieve it?

Multiple solutions exist, some worse than the others ;-)



How to achieve it?

Multiple solutions exist, some worse than the others ;-)

• use of share storage



How to achieve it?

Multiple solutions exist, some worse than the others ;-)

- use of share storage
- use of share blocs by network (drbd)



How to achieve it?

Multiple solutions exist, some worse than the others ;-)

- use of share storage
- use of share blocs by network (drbd)
- use of MySQL replication



How to achieve it?

Multiple solutions exist, some worse than the others ;-)

- use of share storage
- use of share blocs by network (drbd)
- use of MySQL replication

Of course the last one is the technique most spread, more flexible and more reliable.



There are multiple types of MySQL replication:



There are multiple types of MySQL replication:

asynchronous replication (async)



There are multiple types of MySQL replication:

- asynchronous replication (async)
- semi-synchronous replication (semi-sync)



There are multiple types of MySQL replication:

- asynchronous replication (async)
- semi-synchronous replication (semi-sync)
- group replication since (5.7.17)!!



• replicas (also called slaves) stream the replication logs from a unique master *(multi-source replication is possible since 5.7.6)*



- replicas (also called slaves) stream the replication logs from a unique master (multisource replication is possible since 5.7.6)
- slaves can be laging (having delay), there is no delivery guarantee *(unless for semi-sync)*



- replicas (also called slaves) stream the replication logs from a unique master (multisource replication is possible since 5.7.6)
- slaves can be laging (having delay), there is no delivery guarantee *(unless for semi-sync)*
- replicas provisioning is manual



- replicas (also called slaves) stream the replication logs from a unique master (multisource replication is possible since 5.7.6)
- slaves can be laging (having delay), there is no delivery guarantee *(unless for semi-sync)*
- replicas provisioning is manual
- data consistency on replicas is assumed and needs a manual verification



- replicas (also called slaves) stream the replication logs from a unique master (multisource replication is possible since 5.7.6)
- slaves can be laging (having delay), there is no delivery guarantee *(unless for semi-sync)*
- replicas provisioning is manual
- data consistency on replicas is assumed and needs a manual verification
- it's even possible to create circular replication (highly not recommended)



- replicas (also called slaves) stream the replication logs from a unique master (multisource replication is possible since 5.7.6)
- slaves can be laging (having delay), there is no delivery guarantee *(unless for semi-sync)*
- replicas provisioning is manual
- data consistency on replicas is assumed and needs a manual verification
- it's even possible to create circular replication (highly not recommended)
- write operations are safe only from a unique global master



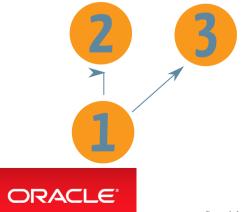
• complex topologies can be built

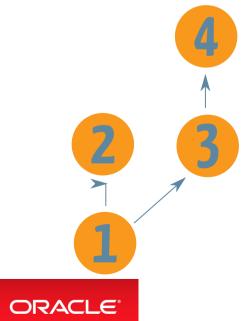


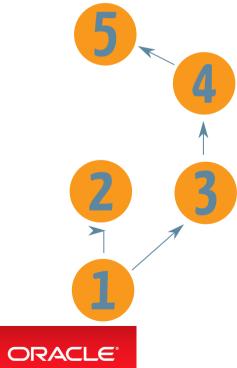
• complex topologies can be built

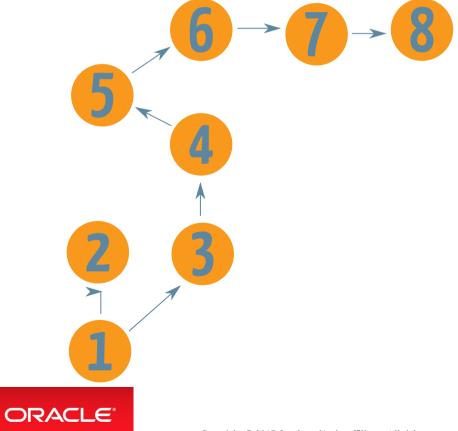


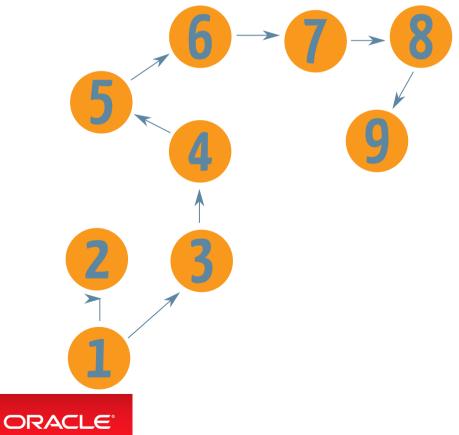
complex topologies can be built

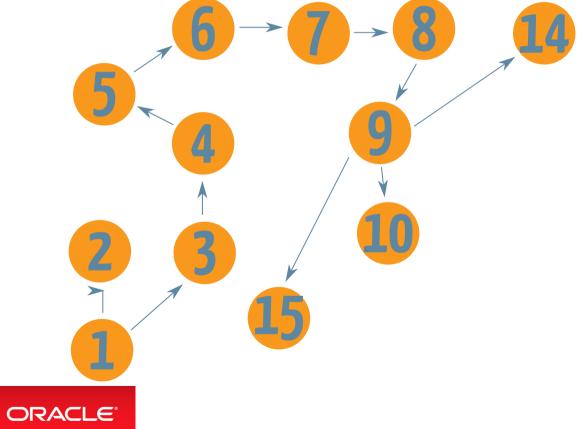


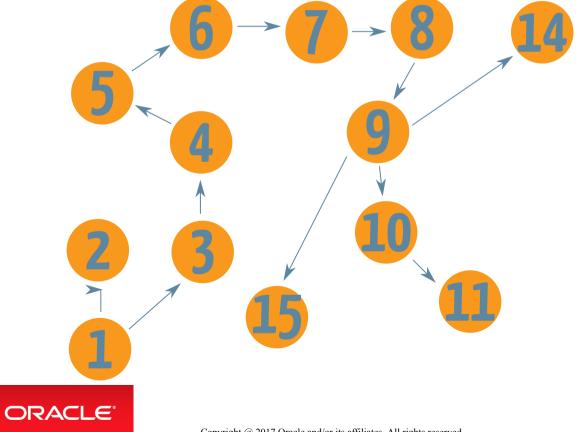


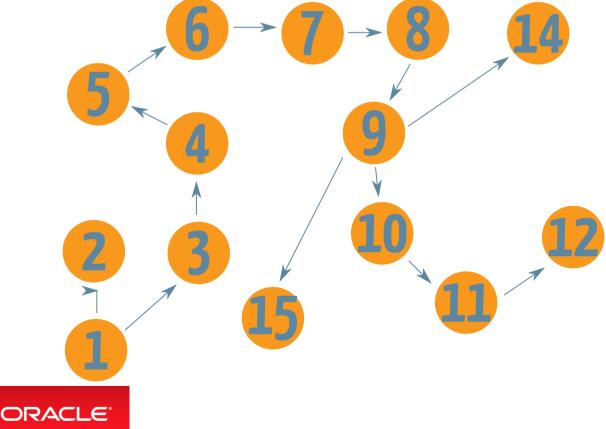


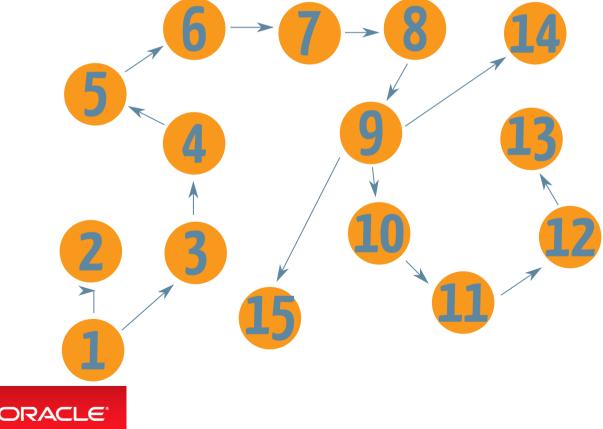


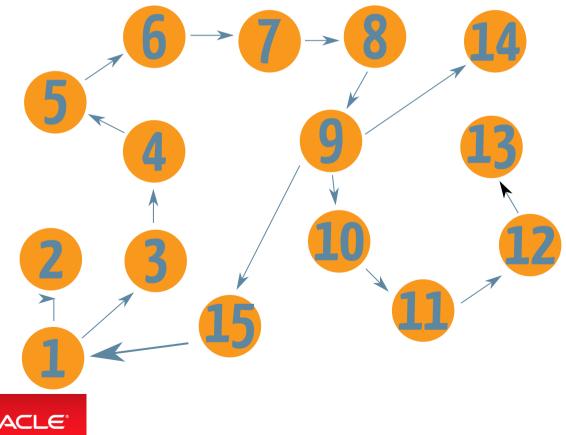












It exists a lot of possible architectures and as much tools to enforce them.



It exists a lot of possible architectures and as much tools to enforce them.

They are all based on the same principle:



It exists a lot of possible architectures and as much tools to enforce them.

They are all based on the same principle:

1 master



It exists a lot of possible architectures and as much tools to enforce them.

They are all based on the same principle:

- 1 master
- 1 or many replicas (slaves)



It exists a lot of possible architectures and as much tools to enforce them.

They are all based on the same principle:

- 1 master
- 1 or many replicas (slaves)

If the master as an issue, the most accurate slave needs to takeover the role and be promoted as master for all the remaining slaves still online.





And this is the most complicated part...



And this is the most complicated part...

It exists a large amount of external tools that can be used to achieve this task. Most of them are only compatible with GNU/Linux.



And this is the most complicated part...

It exists a large amount of external tools that can be used to achieve this task. Most of them are only compatible with GNU/Linux.

Their problem is that they create complexity in the architecture's design. You need to be DBA & sysadmin to manage those solutions:



And this is the most complicated part...

It exists a large amount of external tools that can be used to achieve this task. Most of them are only compatible with GNU/Linux.

Their problem is that they create complexity in the architecture's design. You need to be DBA & sysadmin to manage those solutions:

MySQL-Utilities (mysqlrplcheck, mysqlrpladmin)



And this is the most complicated part...

It exists a large amount of external tools that can be used to achieve this task. Most of them are only compatible with GNU/Linux.

Their problem is that they create complexity in the architecture's design. You need to be DBA & sysadmin to manage those solutions:

- MySQL-Utilities (mysqlrplcheck, mysqlrpladmin)
- Pacemaker with MySQL dedicated OCF



And this is the most complicated part...

It exists a large amount of external tools that can be used to achieve this task. Most of them are only compatible with GNU/Linux.

Their problem is that they create complexity in the architecture's design. You need to be DBA & sysadmin to manage those solutions:

- MySQL-Utilities (mysqlrplcheck, mysqlrpladmin)
- Pacemaker with MySQL dedicated OCF
- custom solutions...





but what is it ?!?

• GR is a plugin for MySQL, made by MySQL and packaged with MySQL



- GR is a plugin for MySQL, made by MySQL and packaged with MySQL
- GR is an implementation of Replicated Database State Machine theory



- GR is a plugin for MySQL, made by MySQL and packaged with MySQL
- GR is an implementation of Replicated Database State Machine theory
- GR uses a Paxos based protocol



- GR is a plugin for MySQL, made by MySQL and packaged with MySQL
- GR is an implementation of Replicated Database State Machine theory
- GR uses a Paxos based protocol
- GR allows to write on all Group Members (cluster nodes) simultaneously while retaining consistency



- GR is a plugin for MySQL, made by MySQL and packaged with MySQL
- GR is an implementation of Replicated Database State Machine theory
- GR uses a Paxos based protocol
- GR allows to write on all Group Members (cluster nodes) simultaneously while retaining consistency
- GR implements conflict detection and resolution



- GR is a plugin for MySQL, made by MySQL and packaged with MySQL
- GR is an implementation of Replicated Database State Machine theory
- GR uses a Paxos based protocol
- GR allows to write on all Group Members (cluster nodes) simultaneously while retaining consistency
- GR implements conflict detection and resolution
- GR allows automatic distributed recovery



- GR is a plugin for MySQL, made by MySQL and packaged with MySQL
- GR is an implementation of Replicated Database State Machine theory
- GR uses a Paxos based protocol
- GR allows to write on all Group Members (cluster nodes) simultaneously while retaining consistency
- GR implements conflict detection and resolution
- GR allows automatic distributed recovery
- Supported on all MySQL platforms!!
 - Linux, Windows, Solaris, OSX, FreeBSD





• no longer necessary to handle server fail-over manually or with a complicated script



- no longer necessary to handle server fail-over manually or with a complicated script
- GR provides fault tolerance



- no longer necessary to handle server fail-over manually or with a complicated script
- GR provides fault tolerance
- GR enables update-everywhere setups



- no longer necessary to handle server fail-over manually or with a complicated script
- GR provides fault tolerance
- GR enables update-everywhere setups
- GR handles crashes, failures, re-connetcs automatically



- no longer necessary to handle server fail-over manually or with a complicated script
- GR provides fault tolerance
- GR enables update-everywhere setups
- GR handles crashes, failures, re-connetcs automatically
- Allows an easy setup of a MySQL service high available!



OK, but how does it work?



it's just ...



it's just ...





it's just ...



... no, in fact the writesets replication is **synchronous** and then certification and apply of the changes are local to each nodes and asynchronous.



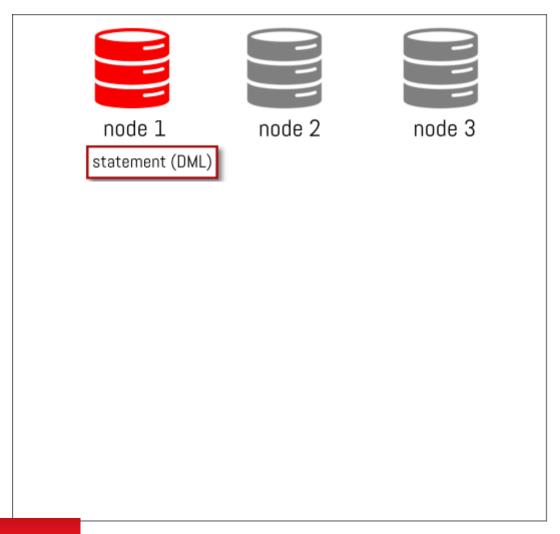
it's just ...

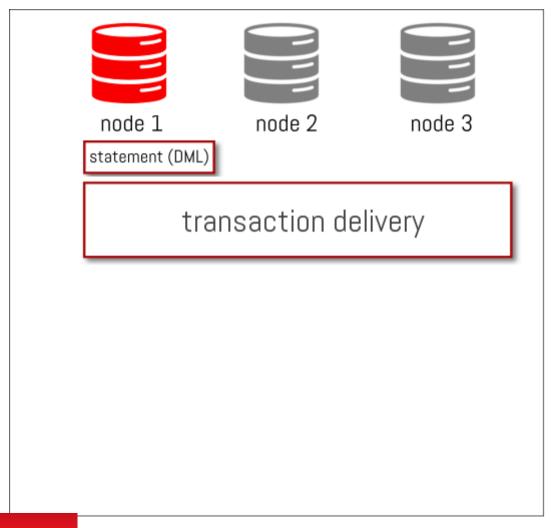


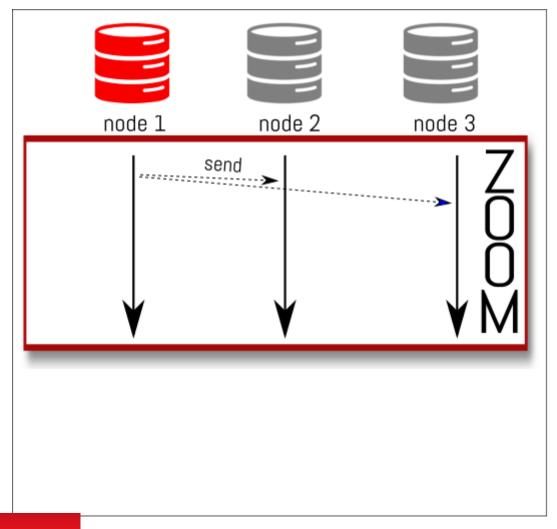
... no, in fact the writesets replication is **synchronous** and then certification and apply of the changes are local to each nodes and asynchronous.

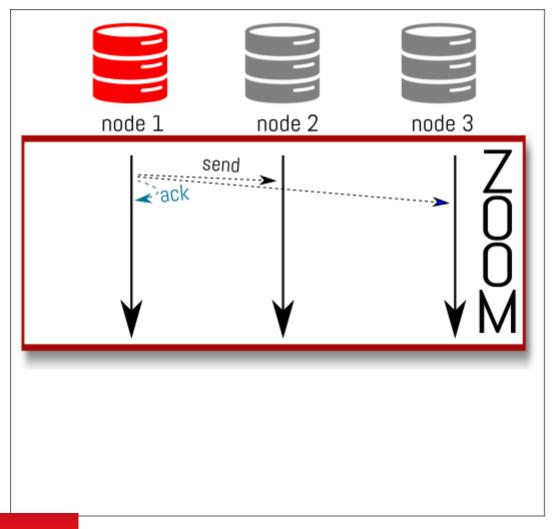
not that easy to understand... right? As a picture is worth a 1000 words, let's illustrate this...

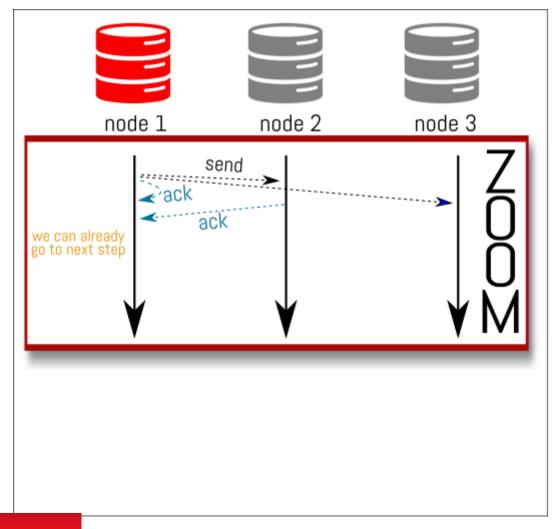


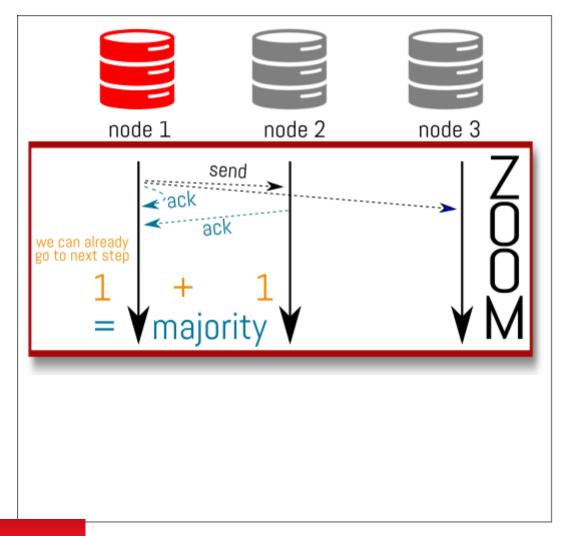


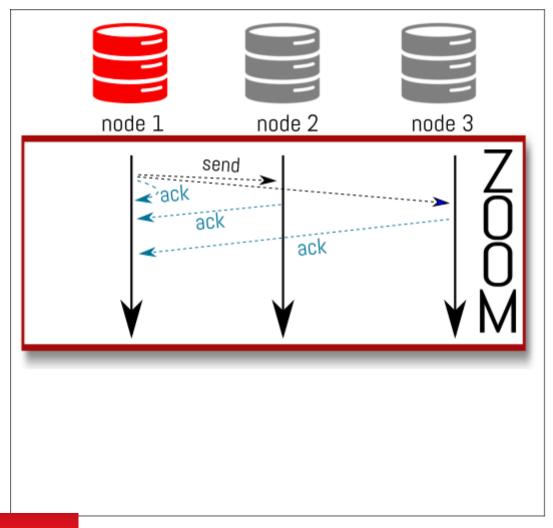


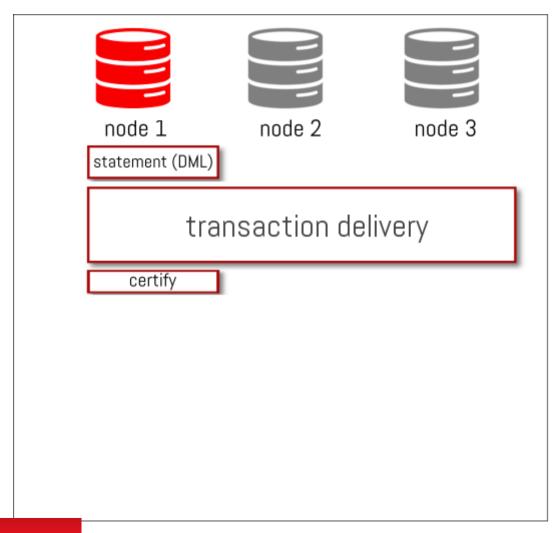




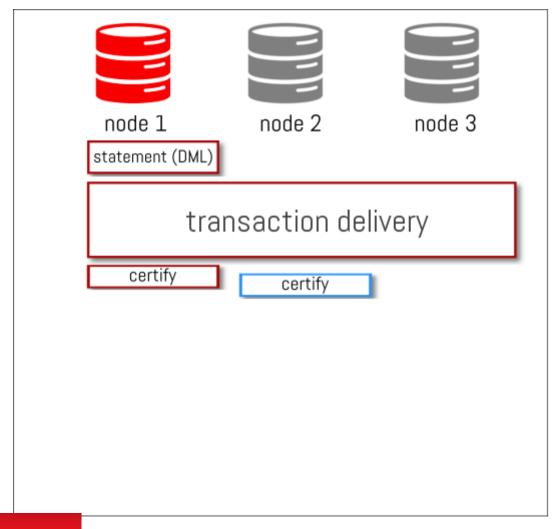


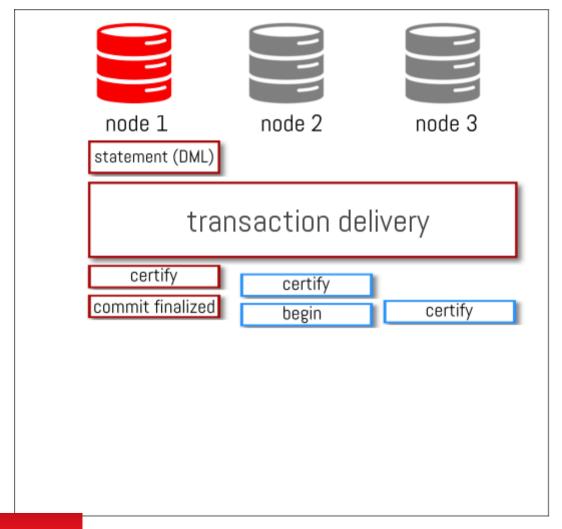


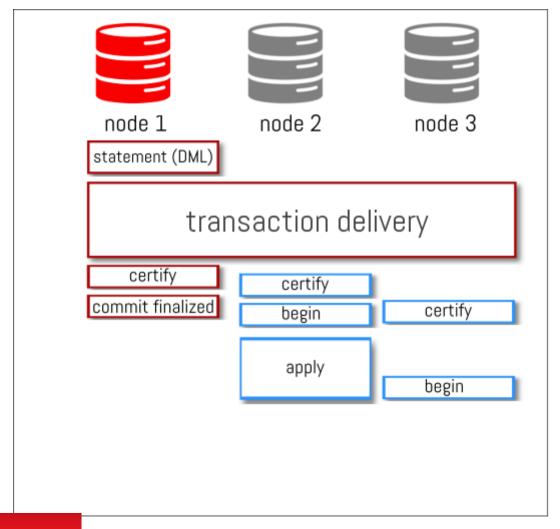


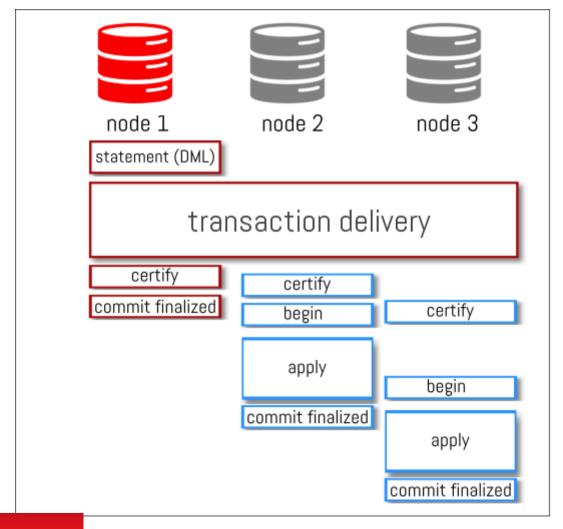














Certification is the process that only needs to answer the following unique question:

• can the write (transaction) be applied?



- can the write (transaction) be applied?
 - based on unapplied earlier transactions



- can the write (transaction) be applied?
 - based on unapplied earlier transactions
 - such conflicts must come for other members/nodes



- can the write (transaction) be applied?
 - based on unapplied earlier transactions
 - such conflicts must come for other members/nodes
- happens on every member/node



- can the write (transaction) be applied?
 - based on unapplied earlier transactions
 - such conflicts must come for other members/nodes
- happens on every member/node
- should be deterministic on every node



- can the write (transaction) be applied?
 - based on unapplied earlier transactions
 - such conflicts must come for other members/nodes
- happens on every member/node
- should be deterministic on every node
- results are not reported to the group



- can the write (transaction) be applied?
 - based on unapplied earlier transactions
 - such conflicts must come for other members/nodes
- happens on every member/node
- should be deterministic on every node
- results are not reported to the group
 - pass: enter in the apply queue



- can the write (transaction) be applied?
 - based on unapplied earlier transactions
 - such conflicts must come for other members/nodes
- happens on every member/node
- should be deterministic on every node
- results are not reported to the group
 - pass: enter in the apply queue
 - fail: drop the transaction



- can the write (transaction) be applied?
 - based on unapplied earlier transactions
 - such conflicts must come for other members/nodes
- happens on every member/node
- should be deterministic on every node
- results are not reported to the group
 - pass: enter in the apply queue
 - fail: drop the transaction
- serialized by the total order in GCS/XCOM + GTID



- can the write (transaction) be applied?
 - based on unapplied earlier transactions
 - such conflicts must come for other members/nodes
- happens on every member/node
- should be deterministic on every node
- results are not reported to the group
 - pass: enter in the apply queue
 - fail: drop the transaction
- serialized by the total order in GCS/XCOM + GTID
- cost is based on trx size (# rows & # keys)



GTID

GTIDs are the same as those used by asynchronous replication.



but transactions use the Group's GTID

exclusively works with InnoDB tables only



- exclusively works with InnoDB tables only
- every tables must have a PK defined



- exclusively works with InnoDB tables only
- every tables must have a PK defined
- only IPV4 is supported



- exclusively works with InnoDB tables only
- every tables must have a PK defined
- only IPV4 is supported
- a good network with low latency is important



- exclusively works with InnoDB tables only
- every tables must have a PK defined
- only IPV4 is supported
- a good network with low latency is important
- maximum of 9 members per group



- exclusively works with InnoDB tables only
- every tables must have a PK defined
- only IPV4 is supported
- a good network with low latency is important
- maximum of 9 members per group
- log-bin must be enabled and only ROW format is supported



Requirements (2)

• enable GTIDs



Requirements (2)

- enable GTIDs
- replication meta-data must be stored in system tables

```
--master-info-repository=TABLE
--relay-log-info-repository=TABLE
```



Requirements (2)

- enable GTIDs
- replication meta-data must be stored in system tables

```
--master-info-repository=TABLE
--relay-log-info-repository=TABLE
```

writesets extraction must be enabled

```
--transaction-write-set-extraction=XXHASH64
```



Requirements (2)

- enable GTIDs
- replication meta-data must be stored in system tables

```
--master-info-repository=TABLE
--relay-log-info-repository=TABLE
```

writesets extraction must be enabled

```
--transaction-write-set-extraction=XXHASH64
```

log-slave-updates must also be enabled



• binlog checksum is not supported (but we have checksum in the replication channel)

--binlog-checksum=NONE



• binlog checksum is not supported (but we have checksum in the replication channel)

```
--binlog-checksum=NONE
```

• savepoints **were** not supported before 5.7.19 & 8.0.1



• binlog checksum is not supported (but we have checksum in the replication channel)

```
--binlog-checksum=NONE
```

- savepoints were not supported before 5.7.19 & 8.0.1
- SERIALIZABLE is not supported as transaction isolation level



• binlog checksum is not supported (but we have checksum in the replication channel)

```
--binlog-checksum=NONE
```

- savepoints were not supported before 5.7.19 & 8.0.1
- SERIALIZABLE is not supported as transaction isolation level
- http://lefred.be/content/mysql-group-replication-limitations-savepoints/
- http://lefred.be/content/mysql-group-replication-and-table-design/



Default = Single Primary Mode

By default, MySQL InnoDB Cluster runs in Single Primary Mode.



Default = Single Primary Mode

By default, MySQL InnoDB Cluster runs in Single Primary Mode.



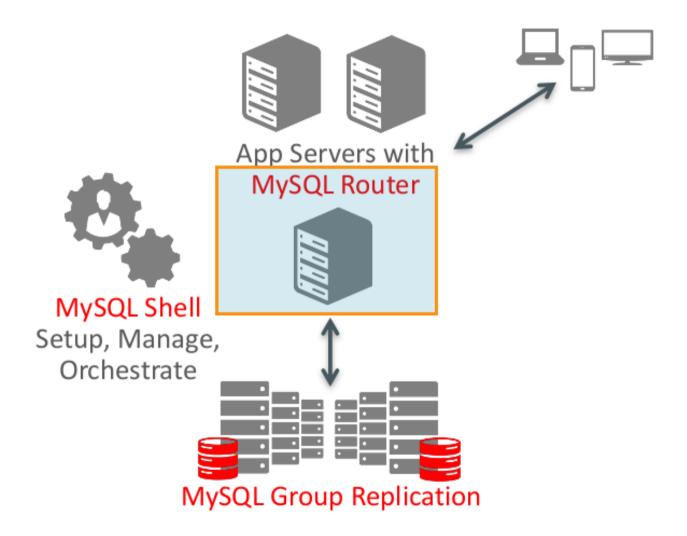
Default = Single Primary Mode

By default, MySQL InnoDB Cluster runs in Single Primary Mode.

In Single Primary Mode, a single member acts as the writable master (PRIMARY) and the rest of the members act as hot-standbys (SECONDARY).

The group itself coordinates and configures itself automatically to determine which member will act as the PRIMARY, through a leader election mechanism.







MySQL Router (GA!)

MySQL Router is lightweight middleware that provides transparent routing between your application and backend MySQL Servers. It can be used for a wide variety of use cases, such as providing high availability and scalability by effectively routing database traffic to appropriate backend MySQL Servers.



MySQL Router (GA!)

MySQL Router is lightweight middleware that provides transparent routing between your application and backend MySQL Servers. It can be used for a wide variety of use cases, such as providing high availability and scalability by effectively routing database traffic to appropriate backend MySQL Servers.

MySQL Router doesn't require any specific configuration. It configures itself automatically (bootstrap) using MySQL InnoDB Cluster's metadata.



MySQL Router (GA!)

MySQL Router is lightweight middleware that provides transparent routing between your application and backend MySQL Servers. It can be used for a wide variety of use cases, such as providing high availability and scalability by effectively routing database traffic to appropriate backend MySQL Servers.

MySQL Router doesn't require any specific configuration. It configures itself automatically (bootstrap) using MySQL InnoDB Cluster's metadata.

The MySQL Router development will be focusing on sharding.



ProxySQL

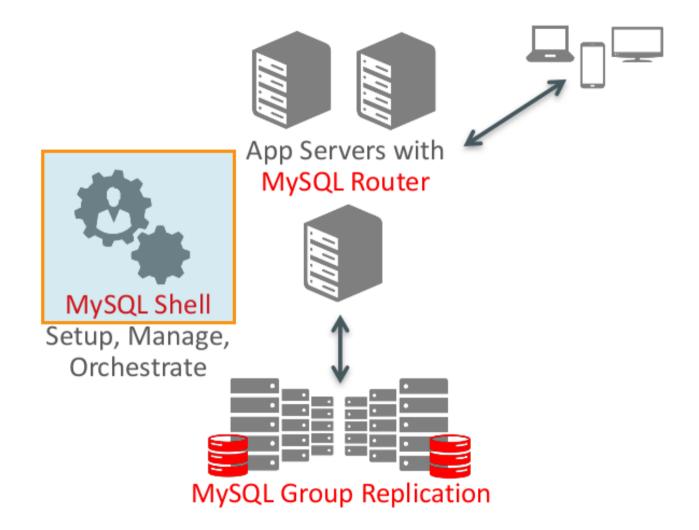
If you need some specific features that are not yet available in MySQL Router, like transparent R/W splitting, then you can use your software of choice.

We are also collaborating with ProxySQL.

ProxySQL has native support for Group Replication which makes it a good choice for advanced users.









MySQL Shell (GA!)

The MySQL Shell is an interactive Javascript, Python, or SQL interface supporting development and administration for the MySQL Server and is a component of the MySQL Server. You can use the MySQL Shell to perform data queries and updates as well as various administration operations.





The MySQL Shell provides:

Both Interactive and Batch operations



- Both Interactive and Batch operations
- Document and Relational Models



- Both Interactive and Batch operations
- Document and Relational Models
- CRUD Document and Relational APIs via scripting



- Both Interactive and Batch operations
- Document and Relational Models
- CRUD Document and Relational APIs via scripting
- Traditional Table, JSON, Tab Separated output results formats



- Both Interactive and Batch operations
- Document and Relational Models
- CRUD Document and Relational APIs via scripting
- Traditional Table, JSON, Tab Separated output results formats
- MySQL Standard and X Protocols



- Both Interactive and Batch operations
- Document and Relational Models
- CRUD Document and Relational APIs via scripting
- Traditional Table, JSON, Tab Separated output results formats
- MySQL Standard and X Protocols
- and more...



PREVIEW: the new Shell's prompt is nice and clear!

```
dba.checkInstanceConfiguration(i2)
Please provide the password for 'root@mysql2:3306':
Validating instance...
The instance 'mysql2:3306' is valid for Cluster usage
    "status": "ok"
              dba.checkInstanceConfiguration(i3)
Please provide the password for 'root@mysql3:3306':
```



don't forget, this is a HA solution!



Deploying a MySQL InnoDB Cluster using MySQL Shell's adminAPI:

```
mysql-js> var i1 = 'root@instance01:3306';
mysql-js> var i2 = 'root@instance02:3306';
mysql-js> var i3 = 'root@instance03:3306';

mysql-js> dba.checkInstanceConfiguration(i1);
mysql-js> dba.checkInstanceConfiguration(i2);
mysql-js> dba.checkInstanceConfiguration(i3);

mysql-js> shell.connect(i1);
mysql-js> var cluster = dba.createCluster('GhentCluster');

mysql-js> cluster.checkInstanceState(i2);
mysql-js> cluster.addInstance(i2);

mysql-js> cluster.addInstanceState(i3);
mysql-js> cluster.addInstance(i3);
```



DEMO!

https://youtu.be/y3WywG7Zhks



Thanks to the MySQL Shell we can automate everything ;-)

https://github.com/lefred/puppet-lefred-innodbcluster

Super easy using hiera:

common.yaml:

```
innodbcluster::mysql_root_password: StRongP4ssw0rD!
innodbcluster::mysql_bind_interface: eth1
innodbcluster::cluster_name: GhentCluster
innodbcluster::grant::user: fred
innodbcluster::grant::password: fred
innodbcluster::seed: instance01
```

instance01.yaml:

```
classes:
    - innodbcluster
innodbcluster::mysql_serverid: 1
```





example (2)

With hiera-eyaml you can even encrypt your password:

```
innodbcluster::mysql root password: >
    ENC[PKCS7,MIIBeQYJKoZIhvcNAQcDoIIBajCCAWYCAQAxggEhMIIBHQIBADAFMAACAQEw
    DQYJKoZIhvcNAQEBBQAEggEAhaKUTXZ/4L8/aL3XARMfDBEI+s5HPshPg9BI
    FLfovstrb8zmcbk5yb/KD0lDM8Elas0lrnpk8MxwNfKw+hB299JFp8ldAtUk
    ODIieTA8BgkqhkiG9w0BBwEwHQYJYIZIAWUDBAEqBBBJEeoyzHtW/WGpbiUz
    gcXTgBAnb1gGrBZAfAiv/ztwuZ9zl
innodbcluster::mysql bind interface:
innodbcluster::cluster name: GhentCluster
innodbcluster::grant::user: root
innodbcluster::grant::password: >
    ENC[PKCS7,MIIBeQYJKoZIhvcNAQcDoIIBajCCAWYCAQAxggEhMIIBHQIBADAFMAACAQEw
    DQYJKoZIhvcNAQEBBQAEggEAhaKUTXZ/4L8/aL3XARMfDBEI+s5HPshPg9BI
    FLfovstrb8zmcbk5yb/KD0lDM8Elas0lrnpk8MxwNfKw+hB299JFp8ldAtUk
    ODIieTA8BgkqhkiG9w0BBwEwHQYJYIZIAWUDBAEqBBBJEeoyzHtW/WGpbiUz
    gcXTgBAnb1gGrBZAfAiv/ztwuZ9z]
```



DEMO!

https://www.youtube.com/watch?v=skwlmBNE7ts





Resources

http://lefred.be/content/mysql-innodb-cluster-mysql-shell-starter-guide/

http://lefred.be/content/mysql-innodb-cluster-automated-installation-with-puppet/

http://lefred.be/content/category/mysql/group-replication/

https://www.slideshare.net/lefred.descamps/mysql-devops-webinar

http://mysqlserverteam.com/innodb-cluster-in-opc/



Thank you!

Kenny & Vadim for their tests, bugs and feature requests





Thank you!

Any Questions?



