MySQL 8.0: The New Replication Features

Luís Soares
Software Development Director
MySQL Replication

ORACLE

Safe Harbor Statement

The following is intended to outline our general product direction. It is intended for information purposes 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 upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Agenda

Program Agenda

1. Introduction
2. Use Cases
3. Enhancements in MySQL 8 (and 5.7)
4. Roadmap
5. Conclusion

1 Introduction

Today...

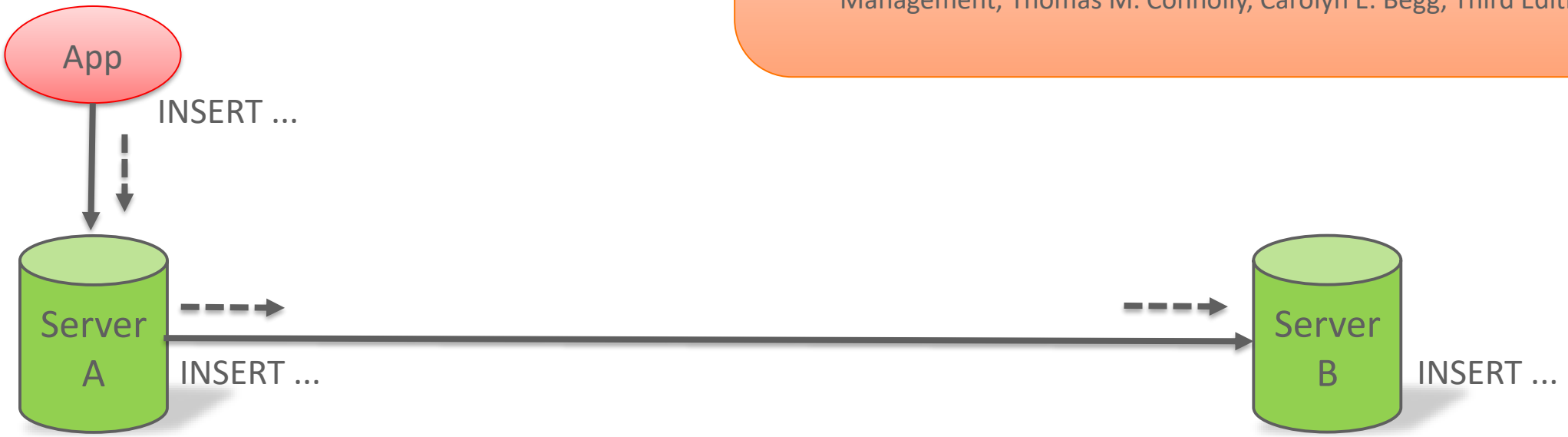
- Technology mesh.
- All things distributed.
- Large amounts of data to handle, transform, store.
- Offline periods are horribly expensive, simply unaffordable.
- Go green requires dynamic and adaptative behavior.
- Much more data to store – e.g. social media, “Look at all of my pictures!”; Monitoring – Keeping logs for N years! ; IoT – and much more.
- Moving, transforming and processing data quicker than anyone else means having an edge over competitors.
- It is a zoo. Distributed coordination and monitoring is key.

Database Replication

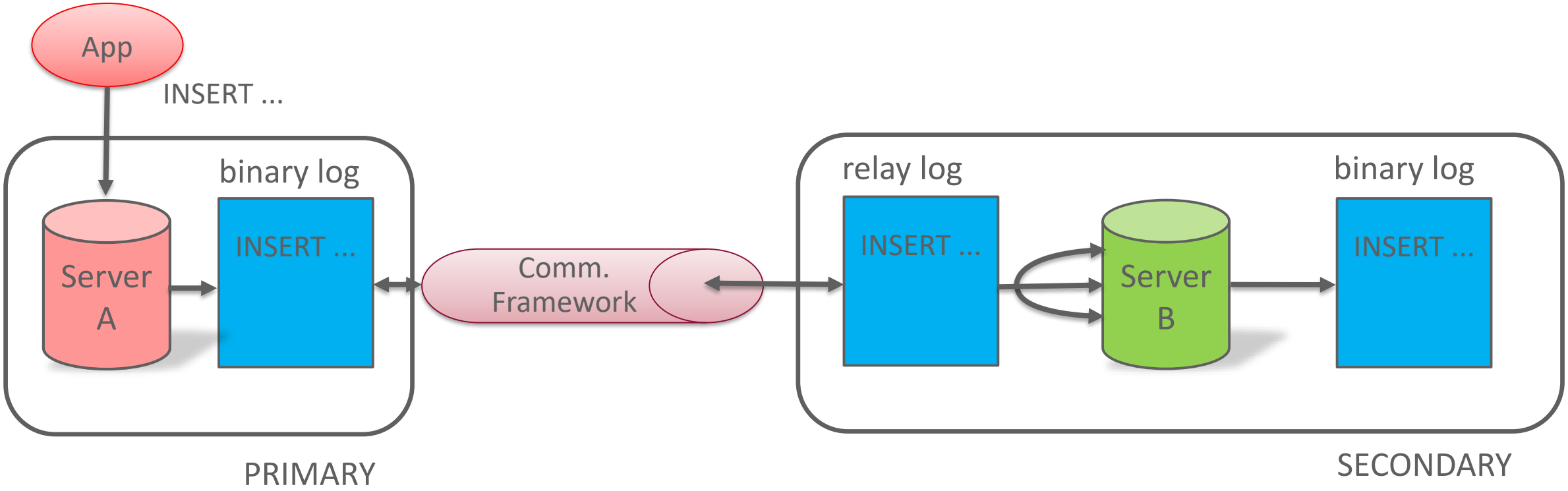
Replication

"The process of generating and reproducing multiple copies of data at one or more sites.",

Database Systems: A Practical Approach to Design, Implementation, and Management, Thomas M. Connolly, Carolyn E. Begg, Third Edition, 2002.



MySQL Database Replication: Overview



MySQL Database Replication: Some Notes

Binary Log

- Logical replication log recording master changes (binary log).
- Row or statement based format (may be intermixed).
- Each transaction is split into groups of events.
- Control events: Rotate, Format Description, Gtid, and more.

MySQL Database Replication: Some Notes

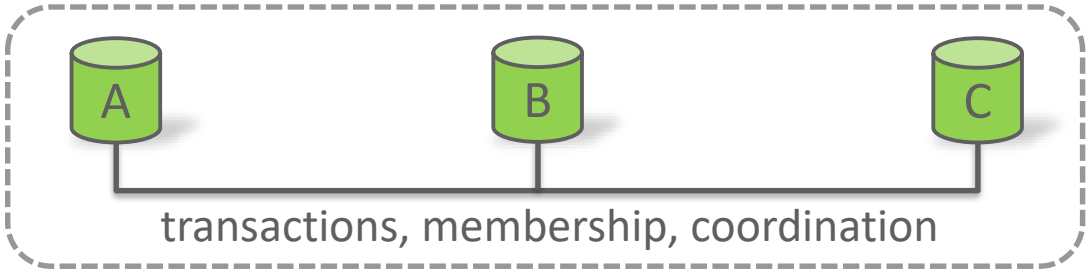
Coordination Between Servers

Since 3.23

asynchronous (native)

Since 5.5

semi-synchronous (plugin)



Since 5.7.17

And in MySQL 8 as of 8.0.1

group replication (plugin)

2 Use cases

Clustering Made Practical

Replicate

Automate

I ntegrate

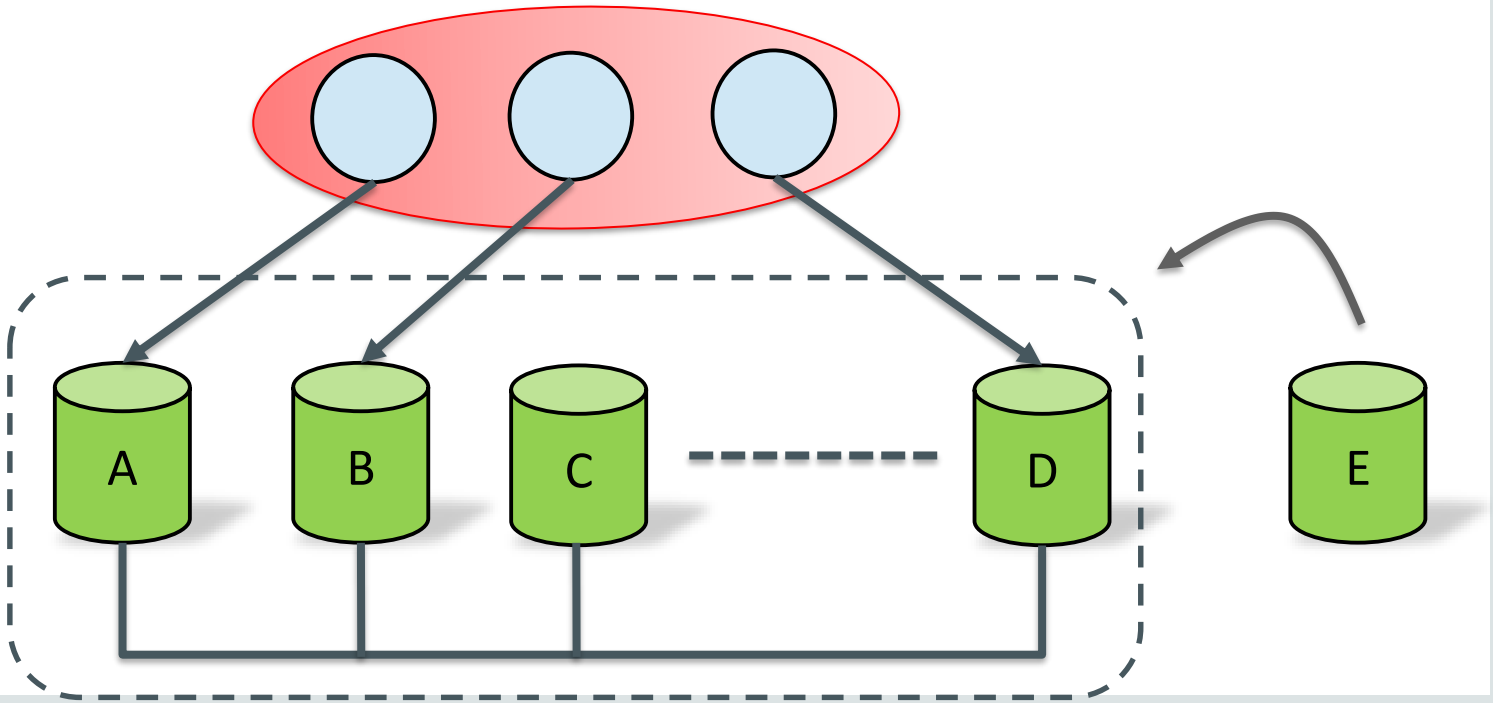
Scale

Enhance

Replicate

Group Replication

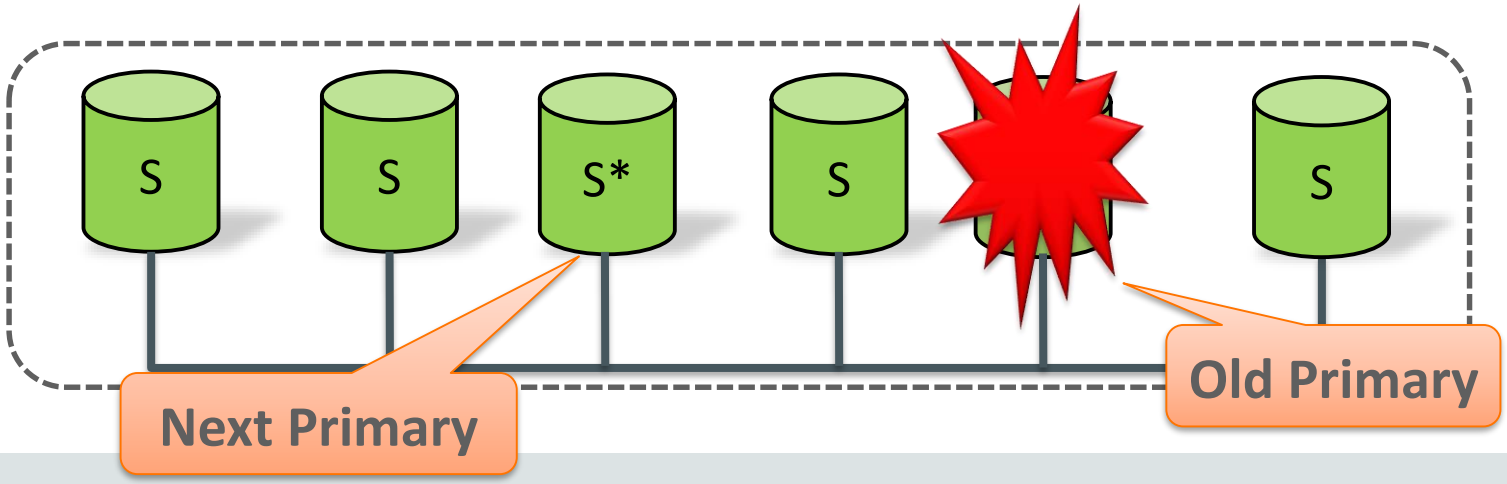
- For highly available infrastructures where:
  - the number of servers has to grow or shrink dynamically;
  - with as little pain as possible.



Automate

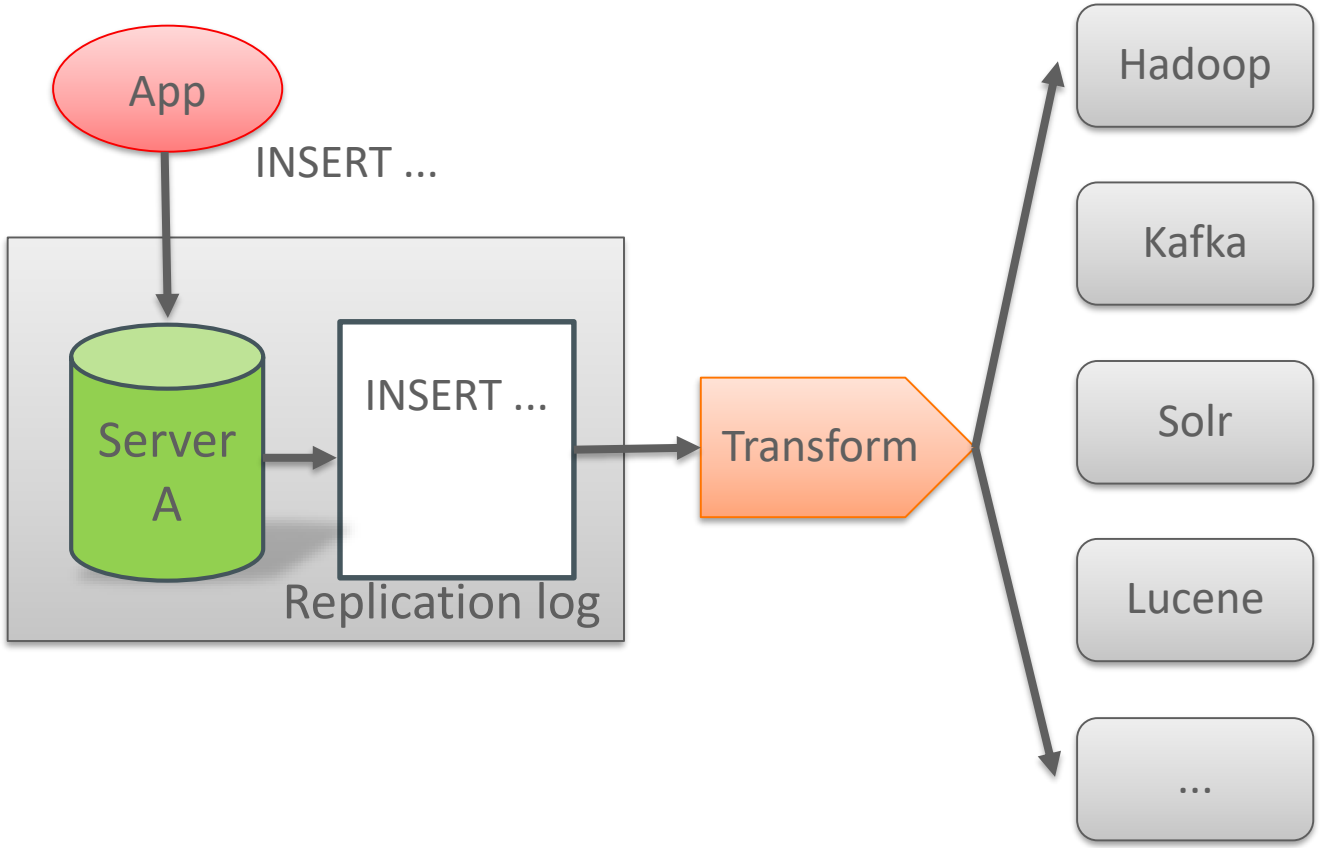
Group Replication

- **Single-primary mode**
  - Automatic PRIMARY/SECONDARY role assignment
  - Automatic new PRIMARY election on PRIMARY failures
  - Automatic setup of read/write modes on PRIMARY and SECONDARIES
  - Automatic global consistent view of which server is the PRIMARY



Integrate

Binary Log

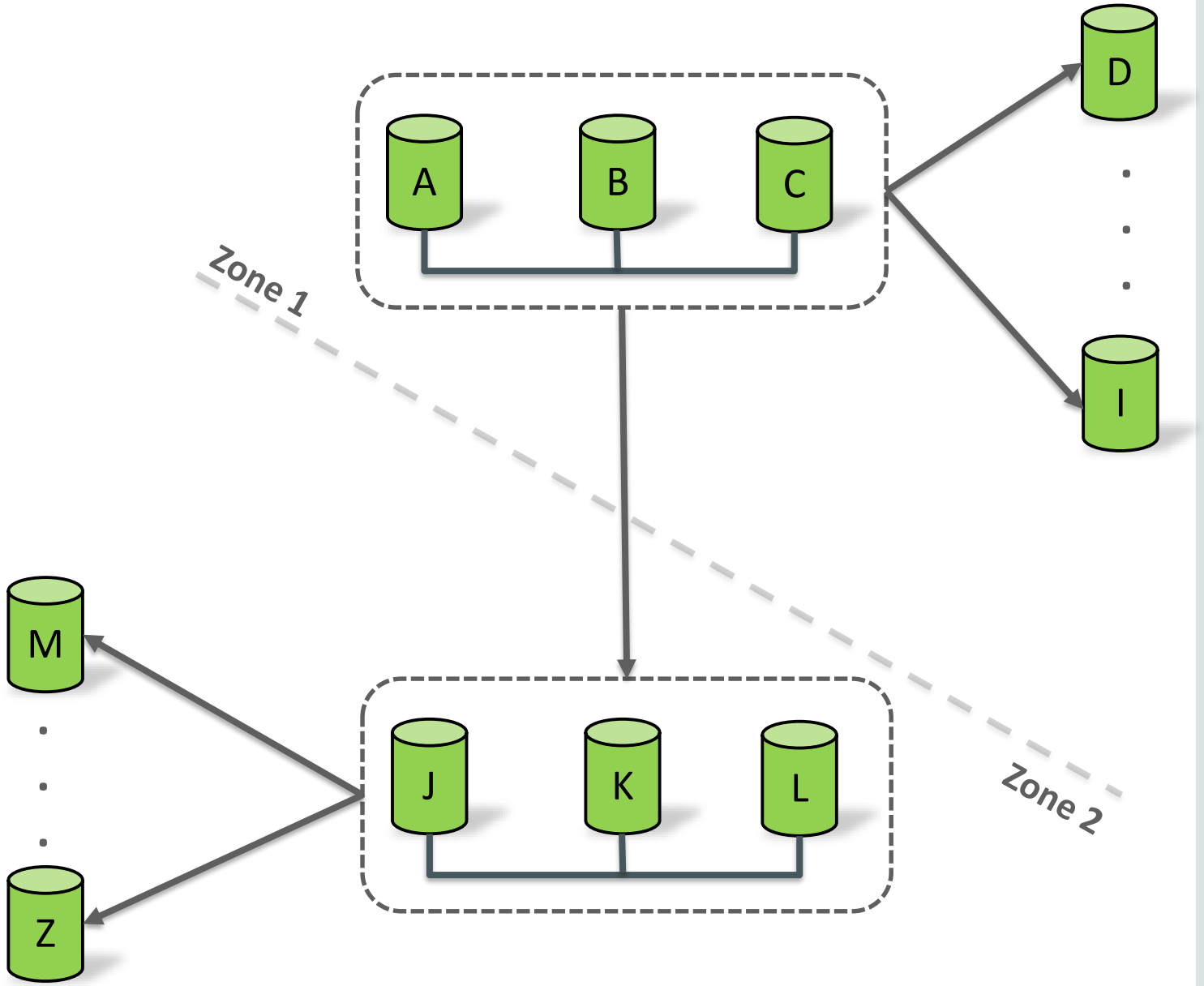


- **Logical replication log**
  - Extract, transform and load.
  - MySQL fits nicely with other technologies.

Scale

Asynchronous Replication

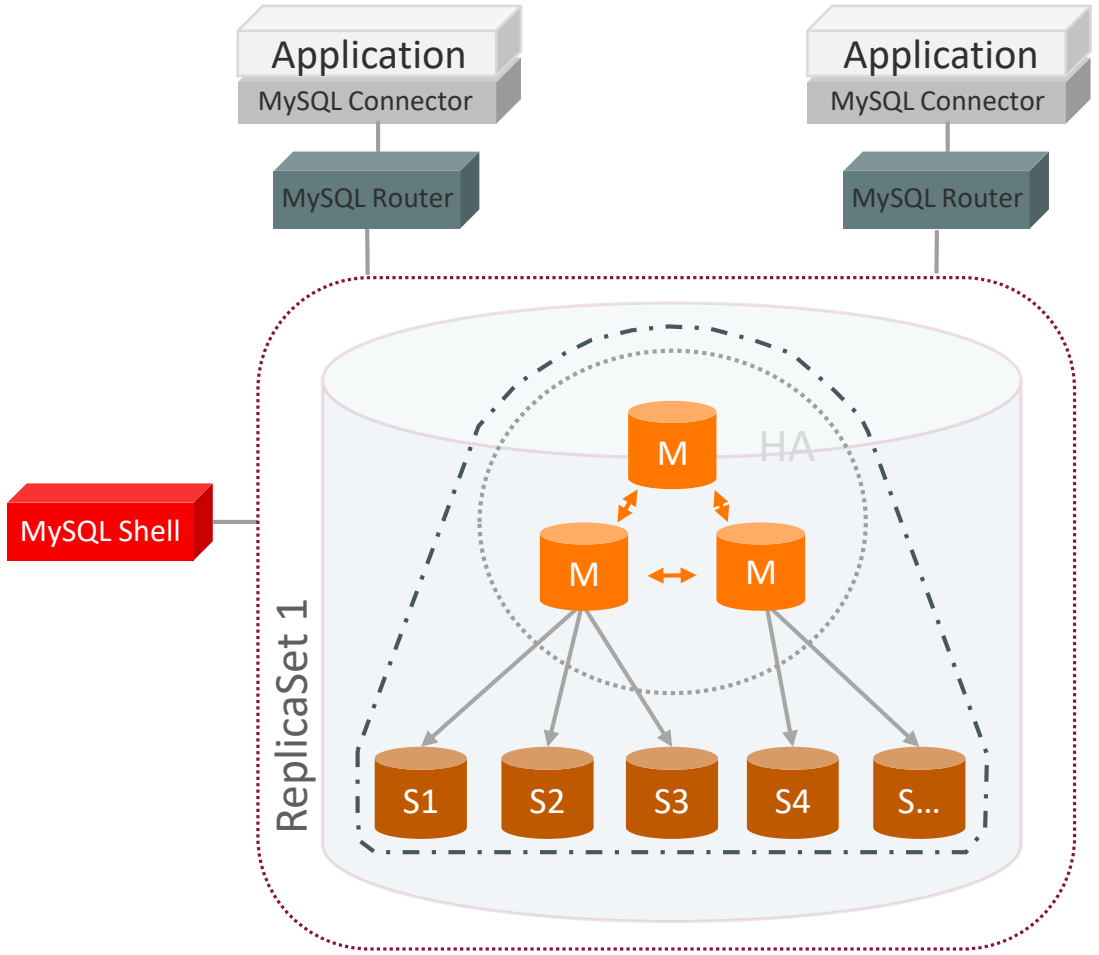
- **Replicate between clusters**
  - For disaster recovery
- **Read Replicas**
  - For read-scale out. Deploy asynchronous read replicas connected to the cluster



Enhance

InnoDB Cluster

- **InnoDB Cluster Integrated Solution**
  - Group Replication for high availability.
  - Asynchronous Replication for Read Scale-out.
  - One-stop shell to deploy and manage the cluster.
  - Seamlessly and automatically route the workload to the proper database server in the cluster.
  - Hide failures from the application.



3 Enhancements in MySQL 8 (and 5.7)

- 3.1 Consistency
- 3.2 Operations
- 3.3 Monitoring
- 3.4 Performance
- 3.5 Security
- 3.6 Other

New in 8.0.14

Consistency Levels

- Eventual Consistency (default)
  - Transaction does not wait at all.
  - Executes on the current snapshot of the data on that member.
- Before Consistency (Synchronize on Reads)
  - Transaction waits for all preceding transactions to complete.
  - Executes on the most up to date snapshot of the data in the group.
- After Consistency (Synchronize on Writes)
  - Transaction waits until all members have executed it.
  - Executes on the current snapshot of the data on that member.

New in 8.0.14

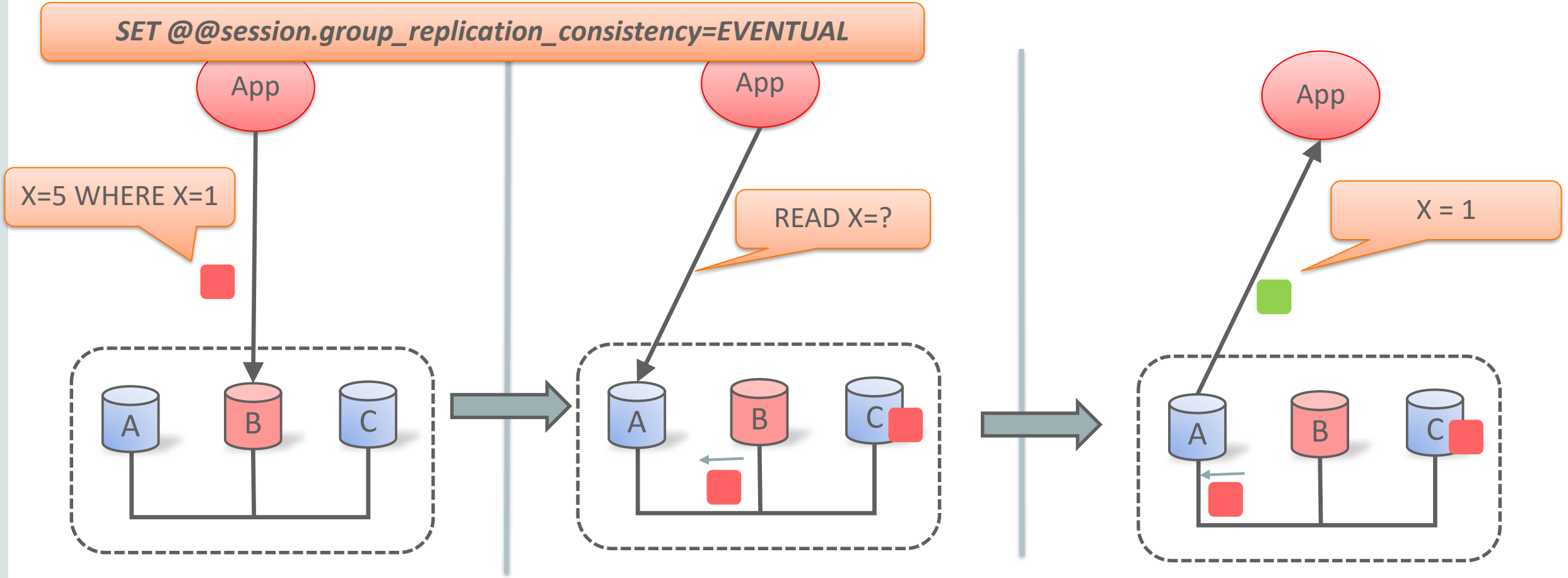
Consistency Levels

- Before and After (Yes, you can **combine** both)
  - Transaction waits for all preceding transactions and for all members to execute it.
  - Executes on the most up to date snapshot of the data in the group and updates everywhere before returning to the application.
- Before On Primary Fail-over
  - Transaction waits for all transactions in the new primary's replication backlog to be executed.
  - Executes on the snapshot of the data that the old primary was in when it stepped down (or crashed).

New in 8.0.14

Accessing Data: Synchronize Execution **Eventually**

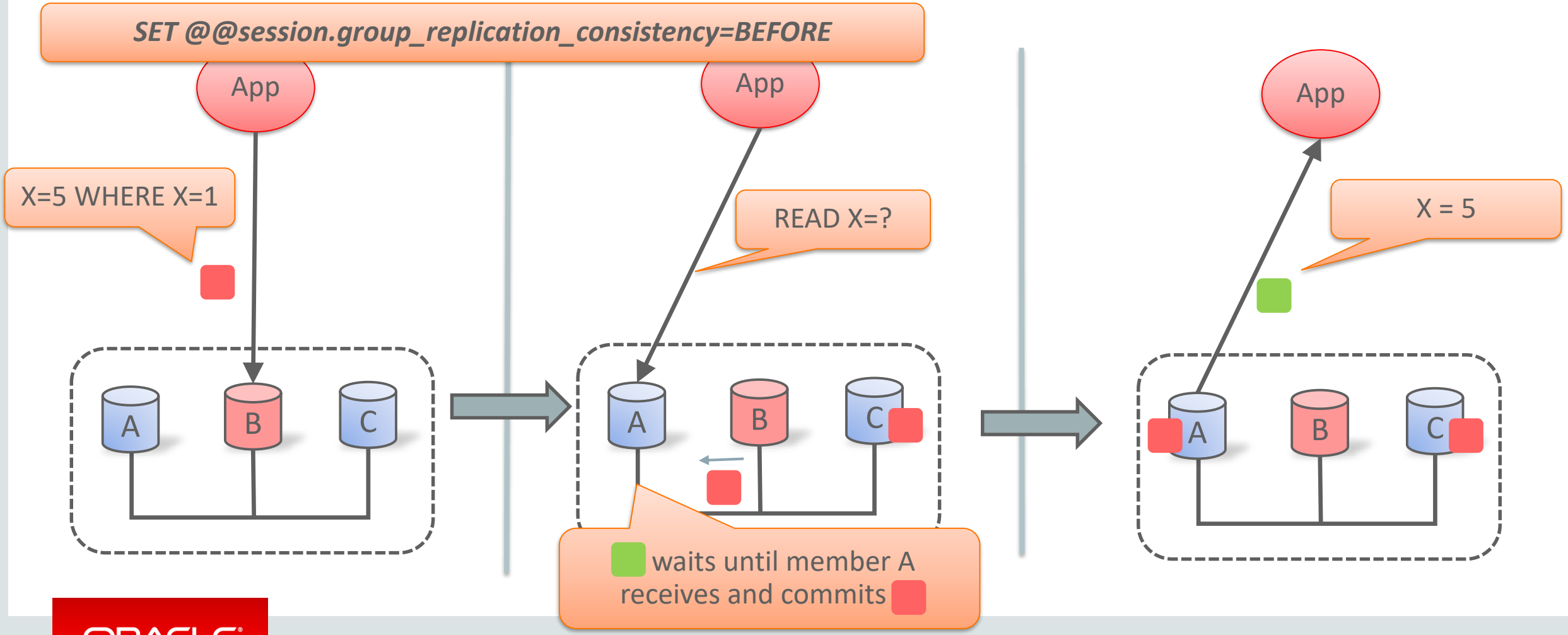
Transaction 🟩 Does Not Wait for 🟥 to Complete Before it Executes On Member A.



New in 8.0.14

Accessing Data: Synchronize **Before** Execution

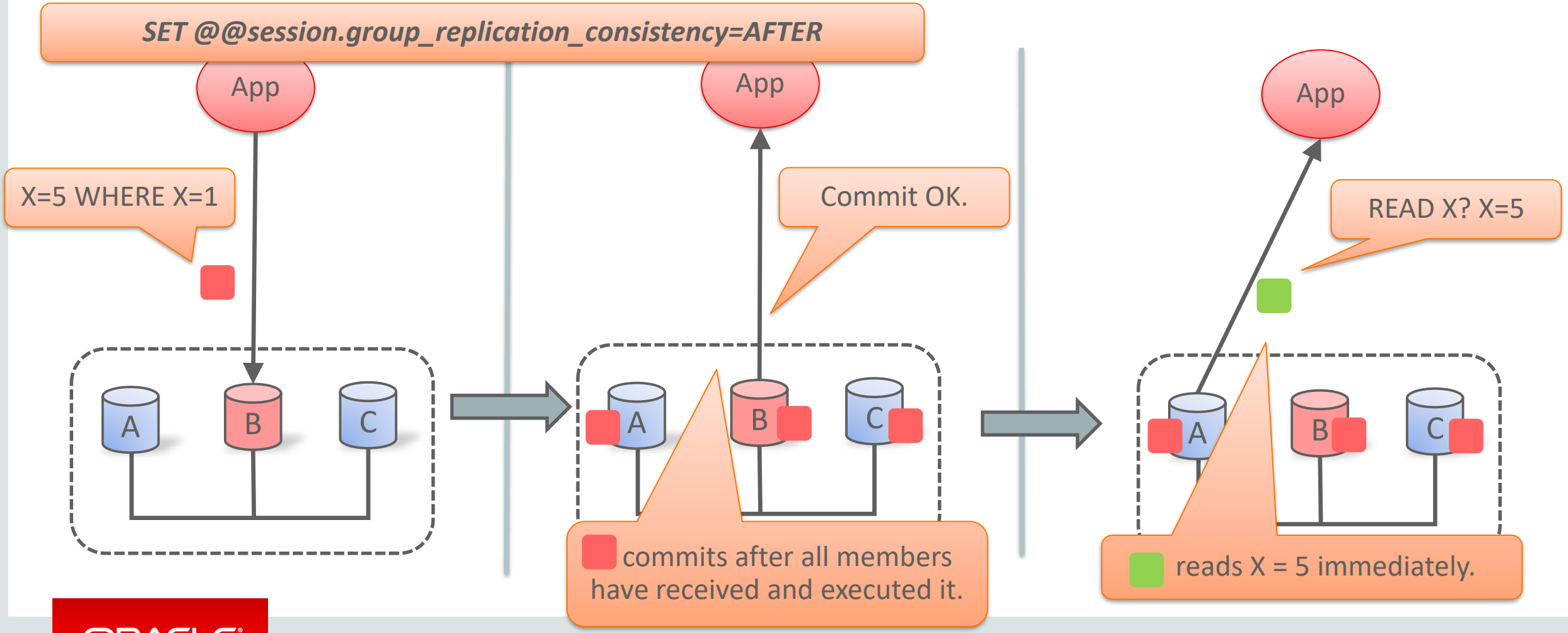
Transaction 🟩 Waits for 🟥 to Complete Before it Executes On Member A.



New in 8.0.14

Accessing Data: Synchronize **After** Execution

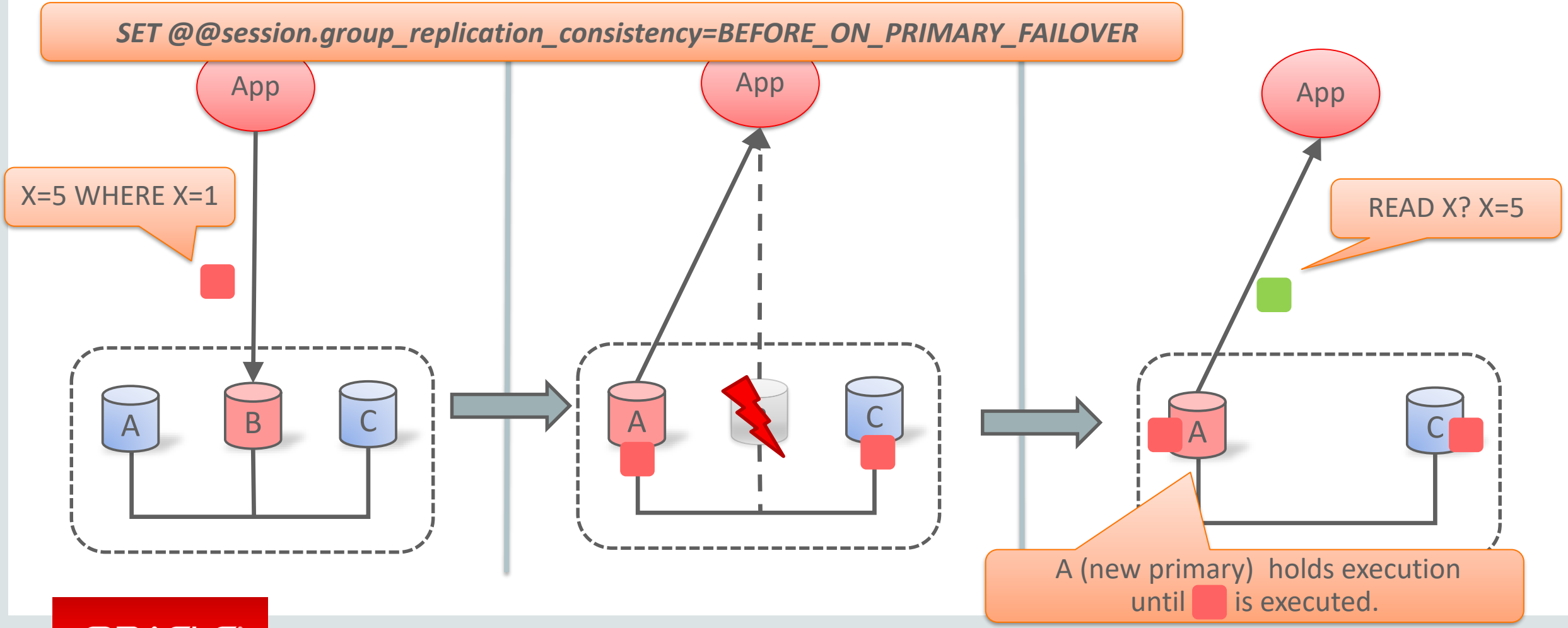
Waits For All Members to Execute. **Reads Updated Value Without Waiting.**



New in 8.0.14

Accessing Data: Synchronize **Before on Primary Fail-over**

Waits For All Members to Execute. **Reads Updated Value Without Waiting.**



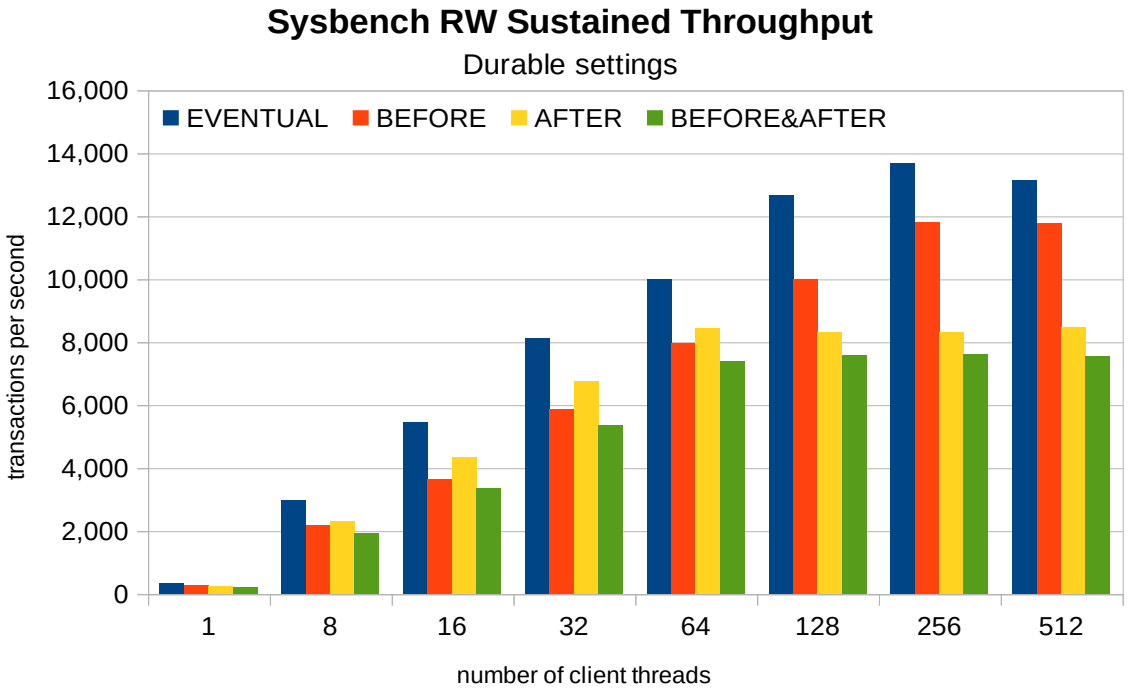
New in 8.0.14

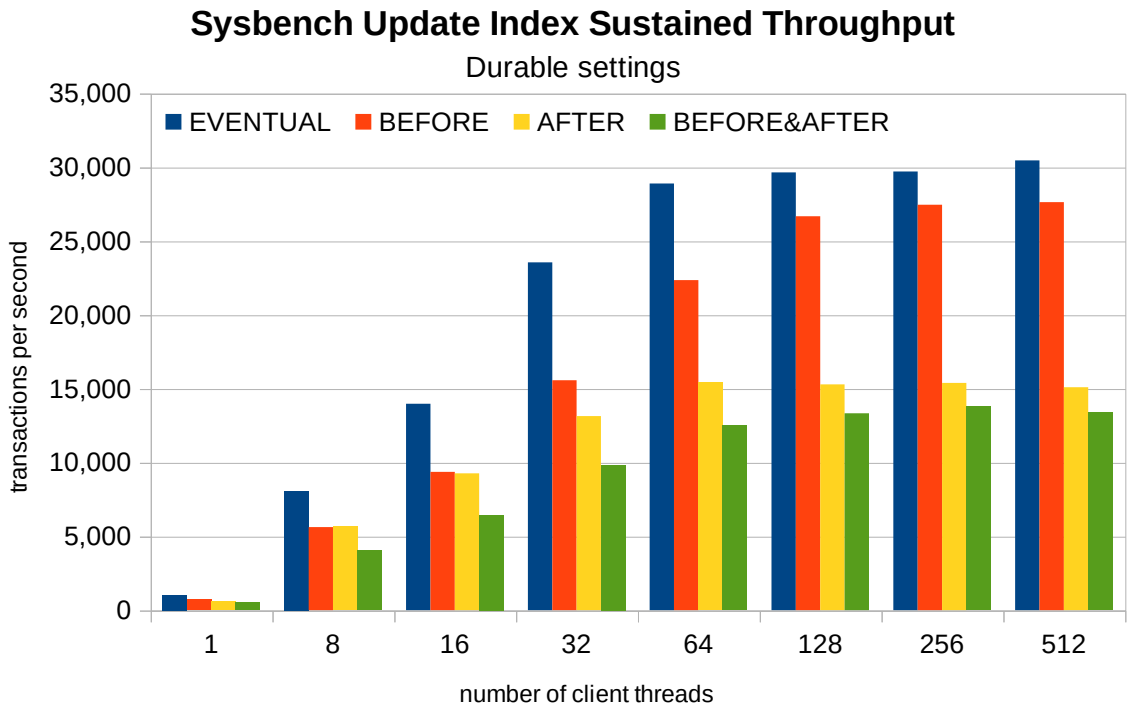
Consistency Levels – User Interface

- System Variable Controls the Behavior: ***group_replication_consistency.***
- *Global and Session scope*
  - ***Can be set per transaction.***
- Values:
  - ***EVENTUAL***
  - ***BEFORE_ON_PRIMARY_FAILOVER***
  - ***BEFORE***
  - ***AFTER***
  - ***BEFORE_AND_AFTER***

Consistency Levels

- Consistency has an impact on throughput.
- Usually not all transactions are executed under strong consistency requirements...





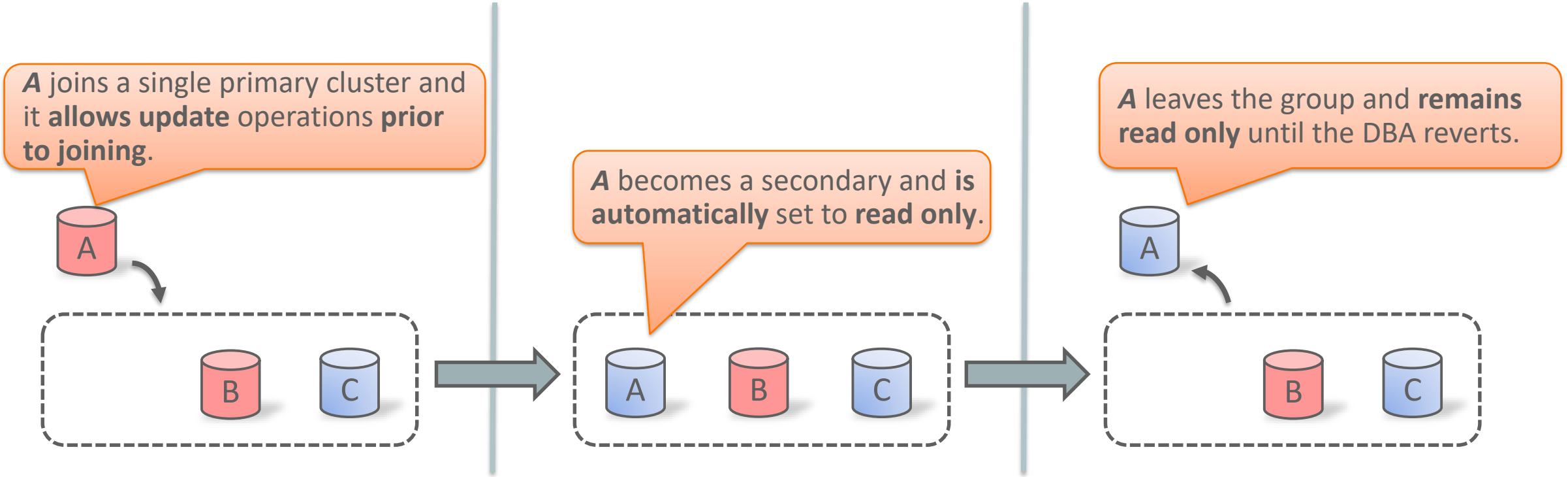
3 Enhancements in MySQL 8 (and 5.7)

- 3.1 Consistency
- 3.2 Operations
- 3.3 Monitoring
- 3.4 Performance
- 3.5 Security
- 3.6 Other

Backported to 5.7.20

Preventing (Involuntary) Writes on Stale Members

Automatically Setting Server to Read-Only

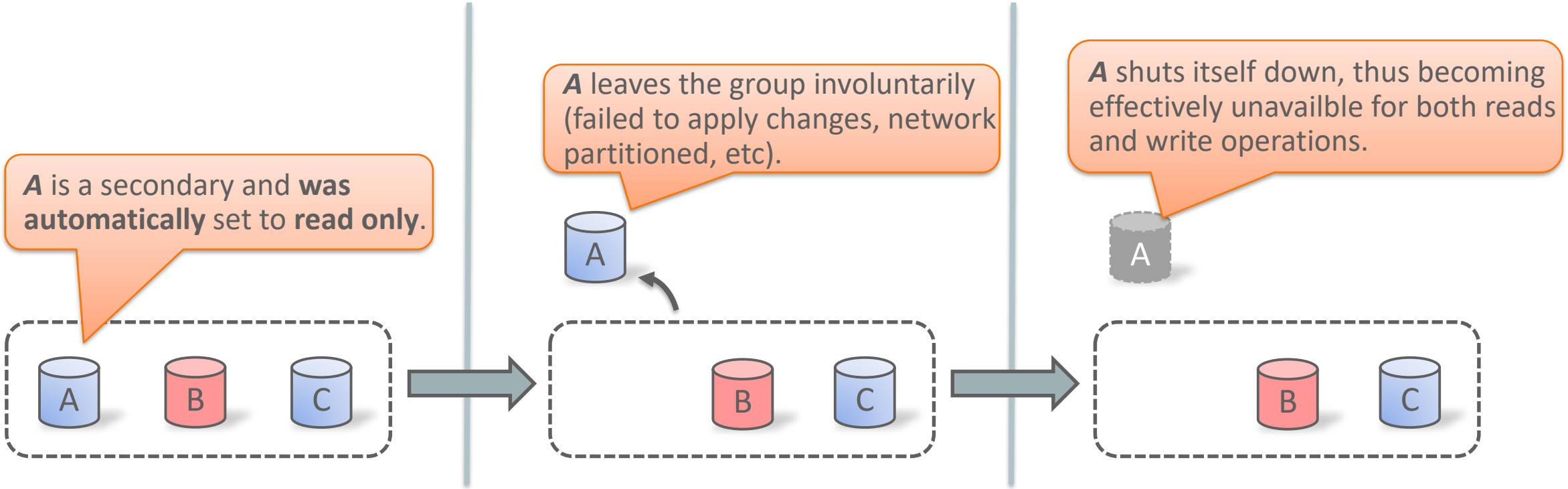


New in 8.0.12

Backported to 5.7.24

Preventing (Involuntary) Reads/Writes on Stale Members

Automatically Shoot Member in the Head (ST*NITH)

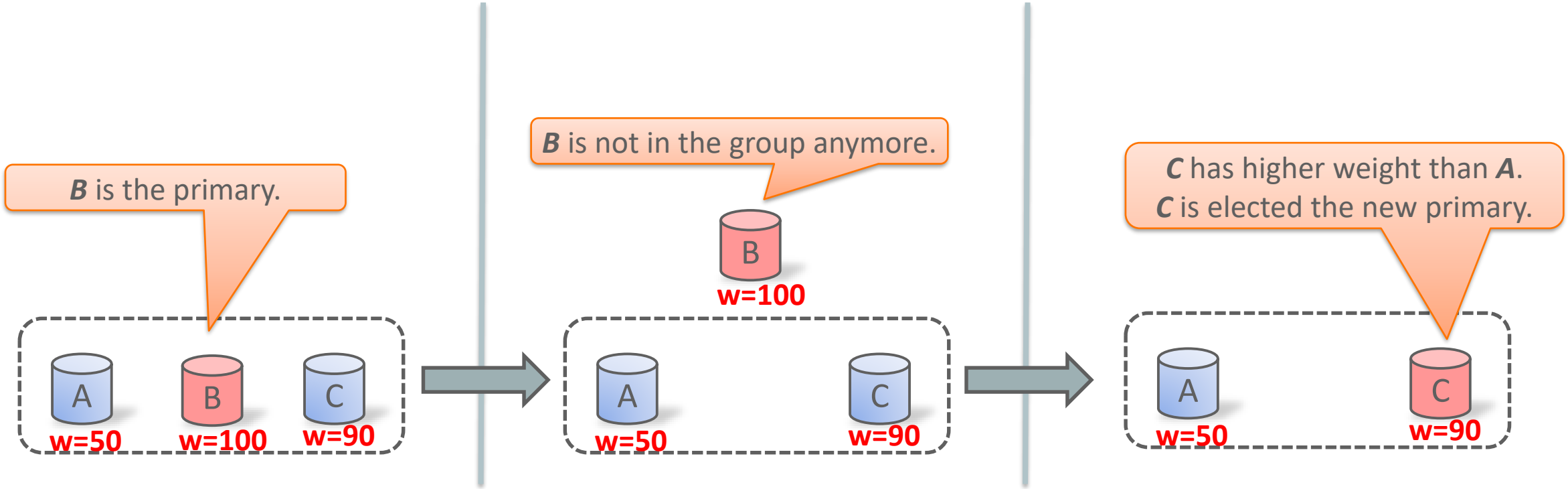


@@group_replication_exit_state_action={ READ_ONLY | ABORT_SERVER }

Backported to 5.7.20

Control Primary Promotion: Priorities

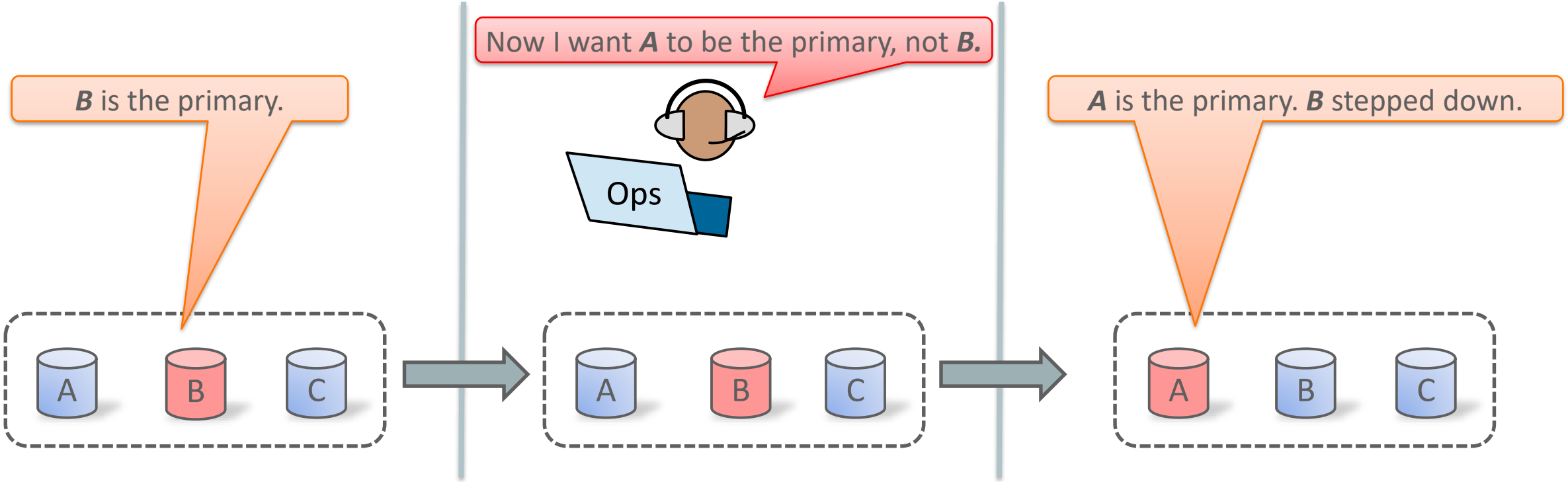
Choose next primary by assigning election weights to the candidates.



New in 8.0.13

Control Primary Election: Choose Your Primary.

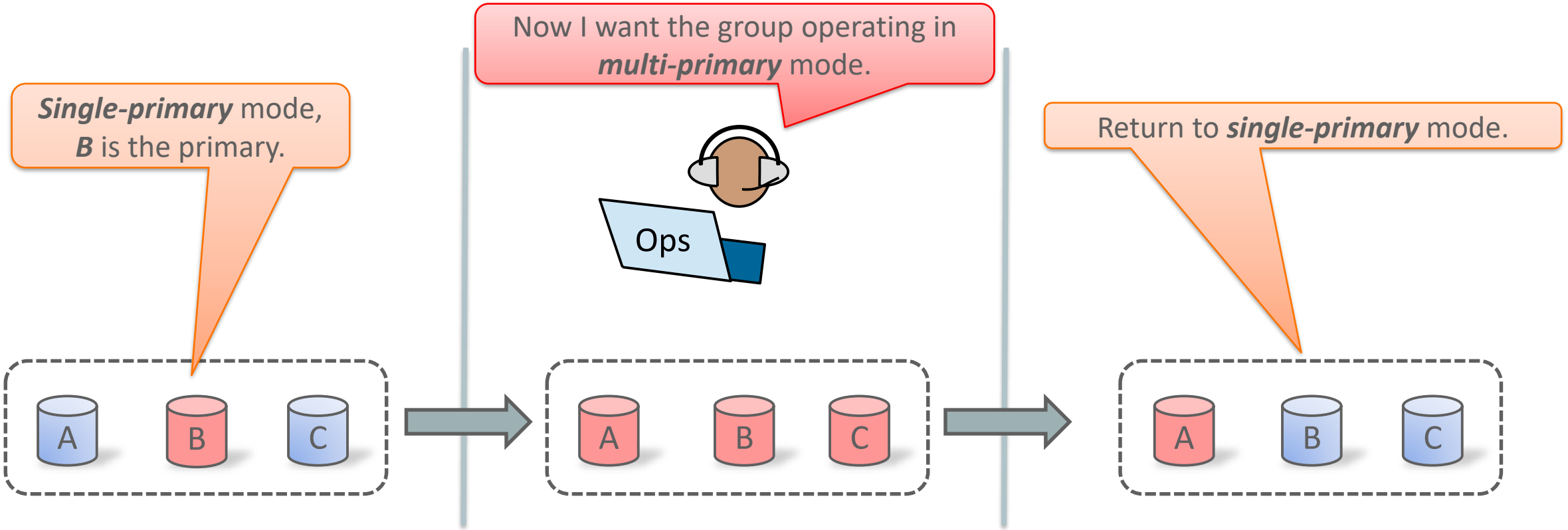
User tells current primary to give up its role and assign it to another server.



New in 8.0.13

Single-Primary to Multi-Primary and Back Online

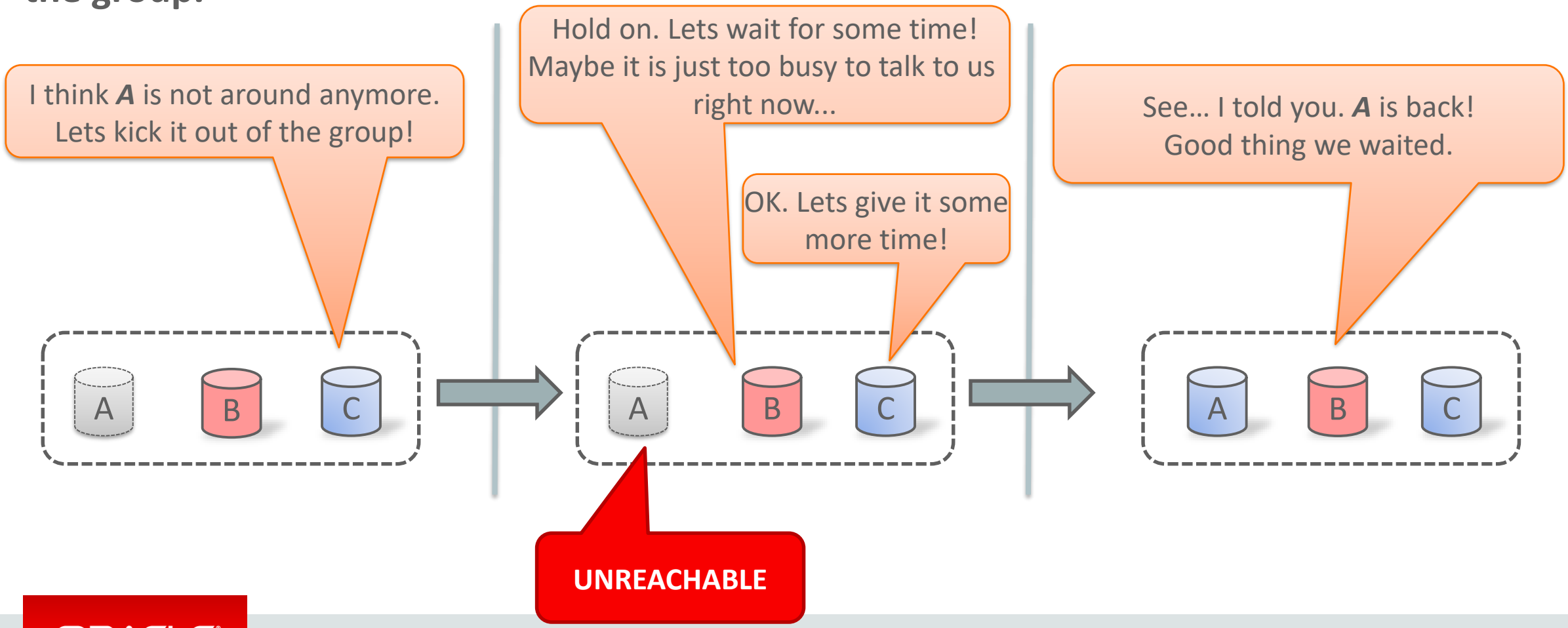
User can specify, online, on which mode the group operates.



New in 8.0.13

Dealing With Unreliable Networks: Relaxed Member Eviction

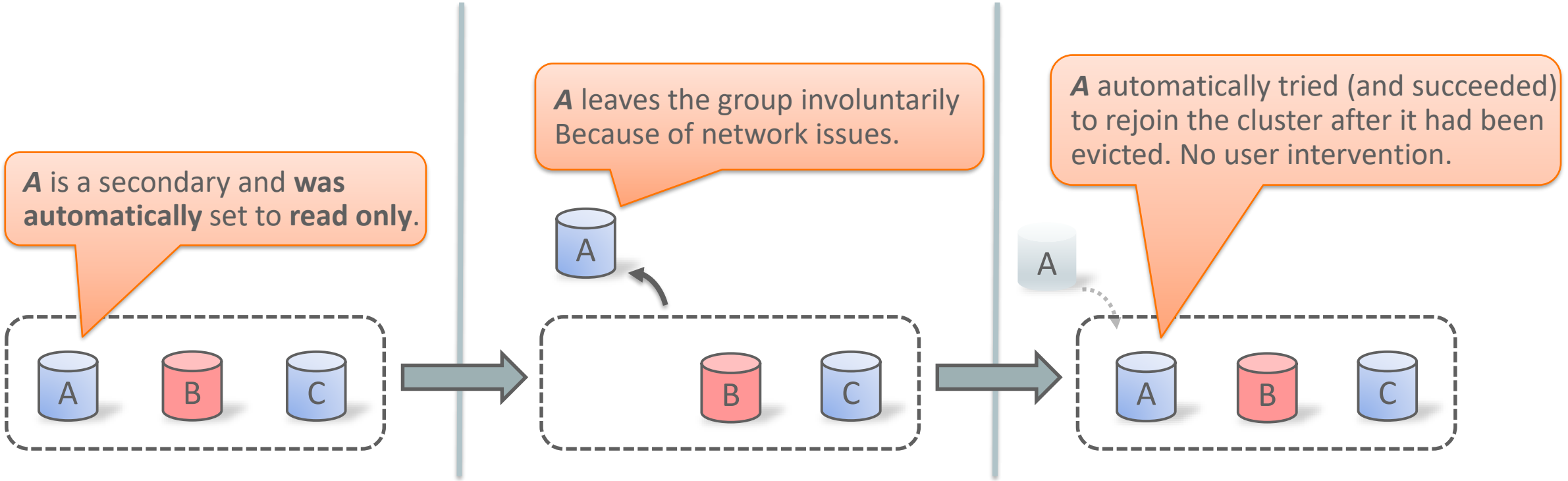
User controls the amount of time to wait until others decide to evict a member from the group.



New in 8.0.16

Automatic Cluster-Rejoin on Partitions

Member tries to rejoin automatically in case it gets evicted.

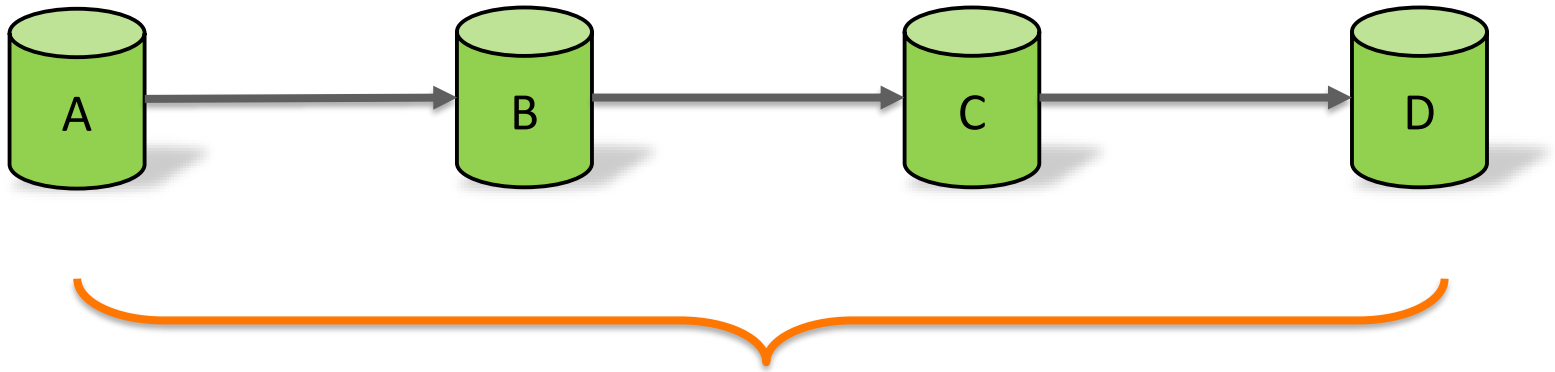


3 Enhancements in MySQL 8 (and 5.7)

- 3.1 Consistency
- 3.2 Operations
- 3.3 **Monitoring**
- 3.4 Performance
- 3.5 Security
- 3.6 Other

Monitor Lag With Microsecond Precision

Through the entire asynchronous topology



How much time does my data take to reach D coming from A?

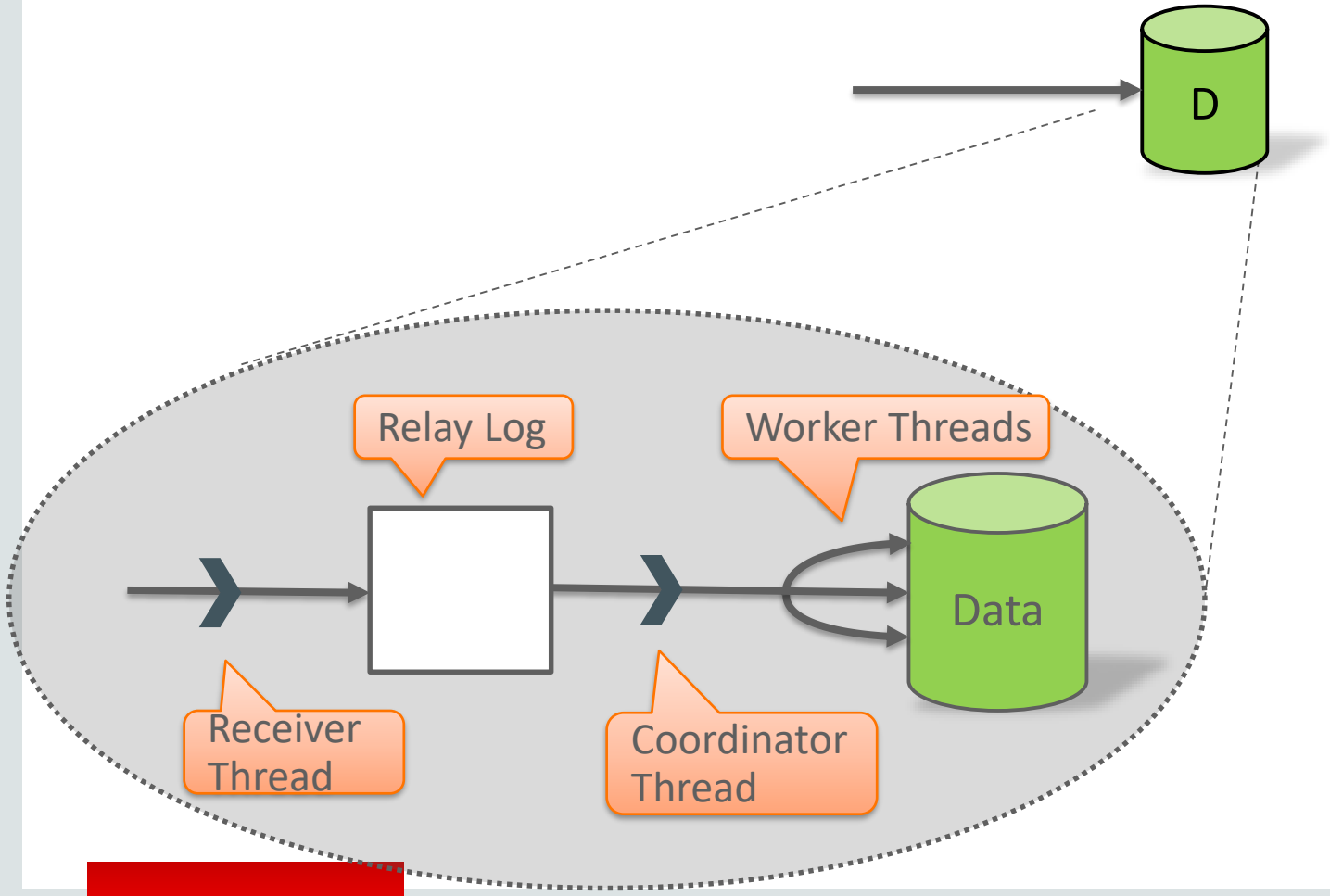
Monitor Lag With Microsecond Precision

From the immediate master

How much time does my data originated in A takes to flow from B to C?

Monitor Lag with Microsecond Precision

For each stage of the replication applier process

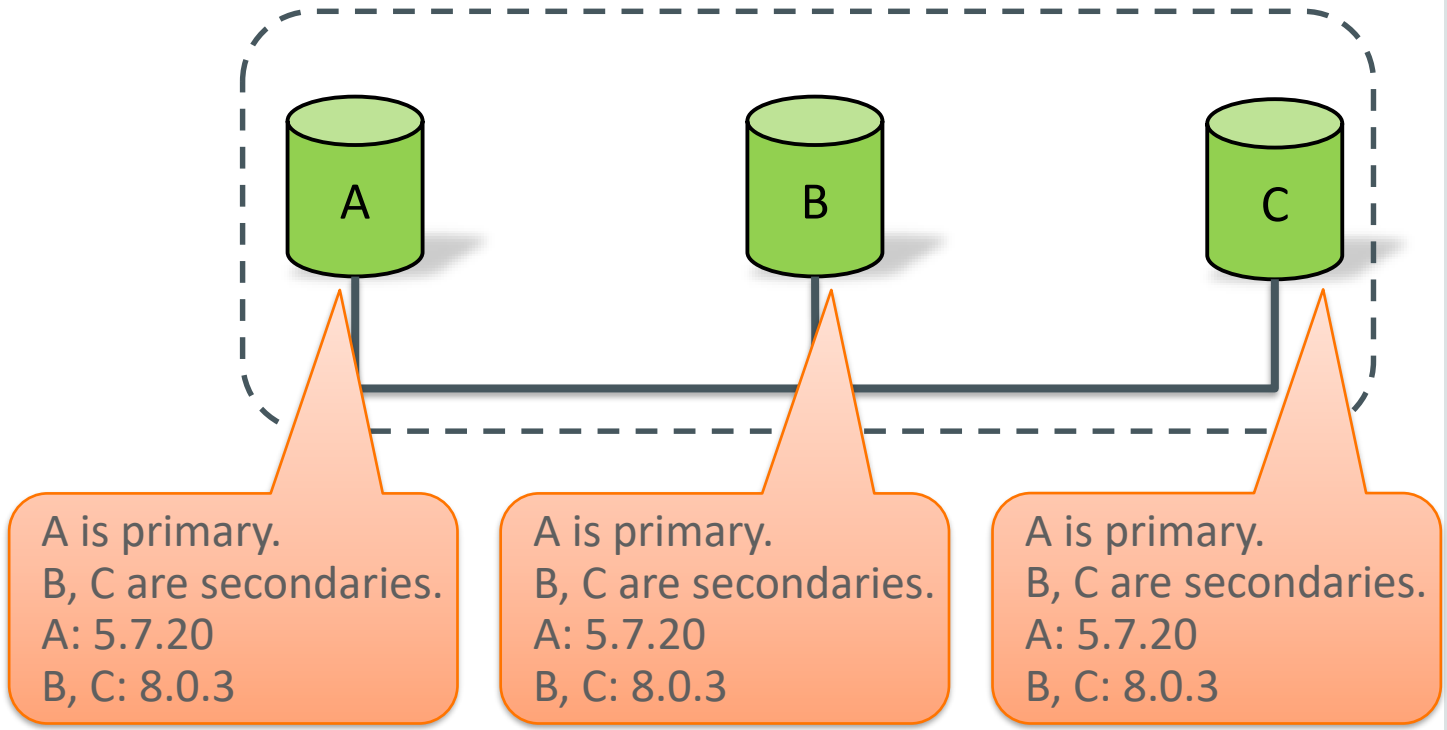


- **Per Stage Timestamps**
  - User can monitor how much time it takes for a specific transaction to traverse the pipeline.

Global Group Stats Available on Every Server

Version, Role and more

- **Query one Replica, Get status of all**
  - Every replica reports group-wide information about roles and versions of the members of the group.
  - Also available at any replica are group-wide status.



New in 8.0.12

Group Replication Message Cache Memory Usage

- GCS/XCom's Paxos message cache is instrumented.
- GCS/XCom's Paxos message cache memory usage is exposed in performance schema.

```
-- This is a session open on ServerA and the user is reading stats on GCS_Xcom message cache
ServerA> select * from memory_summary_global_by_event_name where event_name
like "%GCS_XCom%"\G
*************************** 1. row ***************************
                 EVENT_NAME: memory/group_rpl/GCS_XCom::xcom_cache
                COUNT_ALLOC: 28890317
                 COUNT_FREE: 28840318
   SUM_NUMBER_OF_BYTES_ALLOC: 24499151783
    SUM_NUMBER_OF_BYTES_FREE: 24470424555
              LOW_COUNT_USED: 0
          CURRENT_COUNT_USED: 49999
             HIGH_COUNT_USED: 50000
    LOW_NUMBER_OF_BYTES_USED: 0
CURRENT_NUMBER_OF_BYTES_USED: 28727228
   HIGH_NUMBER_OF_BYTES_USED: 135676530
1 row in set (0.01 sec)
```

3 Enhancements in MySQL 8 (and 5.7)

- 3.1 Consistency
- 3.2 Operations
- 3.3 Monitoring
- 3.4 **Performance**
- 3.5 Security
- 3.6 Other

Backported to 5.7.22

Highly Efficient Replication Applier

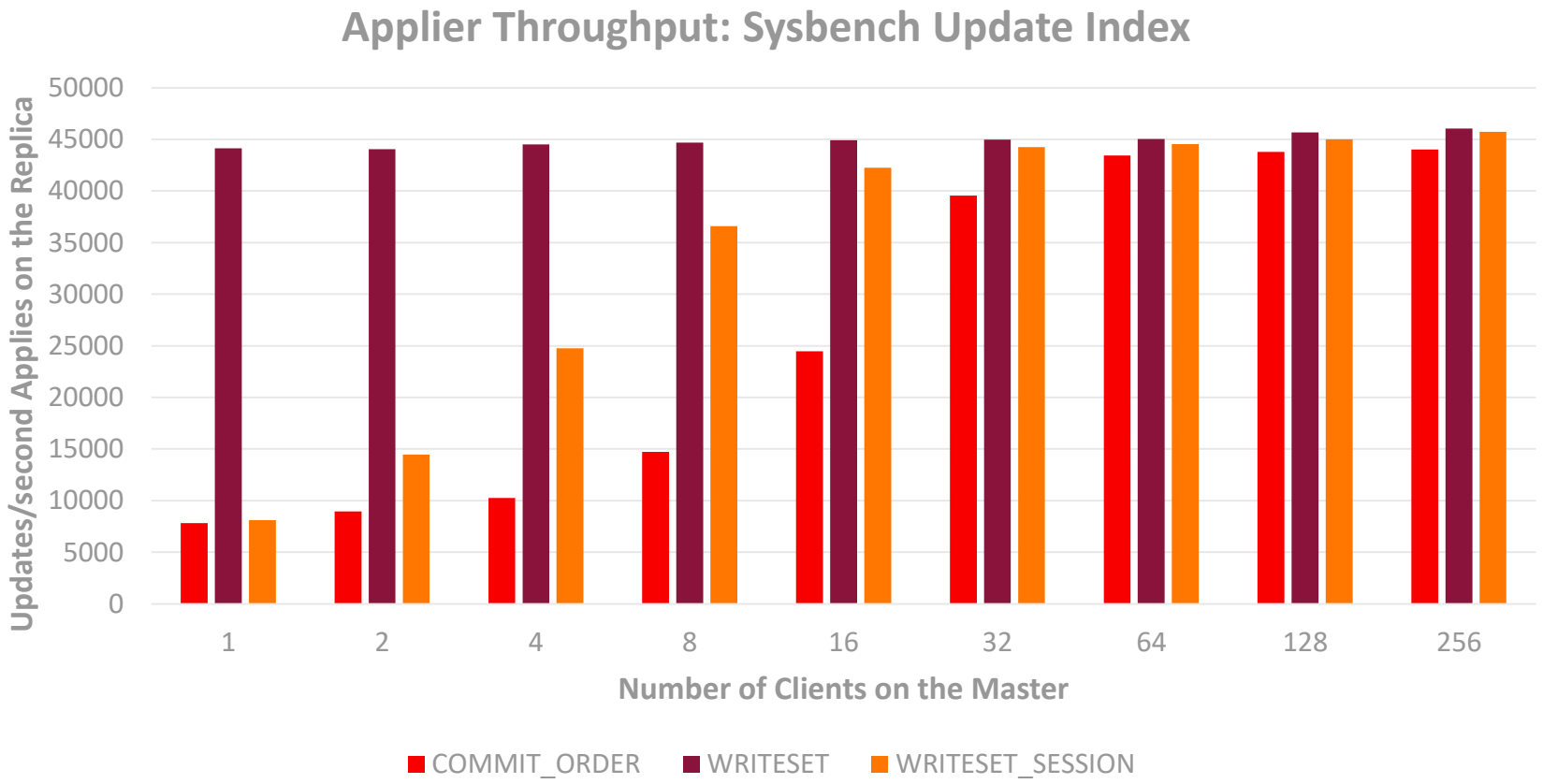
Write set parallelization

- Delivers the **best throughput** of the three dependency trackers, at **any** concurrency level.
- WRITESET dependency tracking allows **applying** a **single threaded** workload in **parallel**.
- Fast Group Replication recovery – time to catch up.

Backported to 5.7.22

Highly Efficient Replication Applier

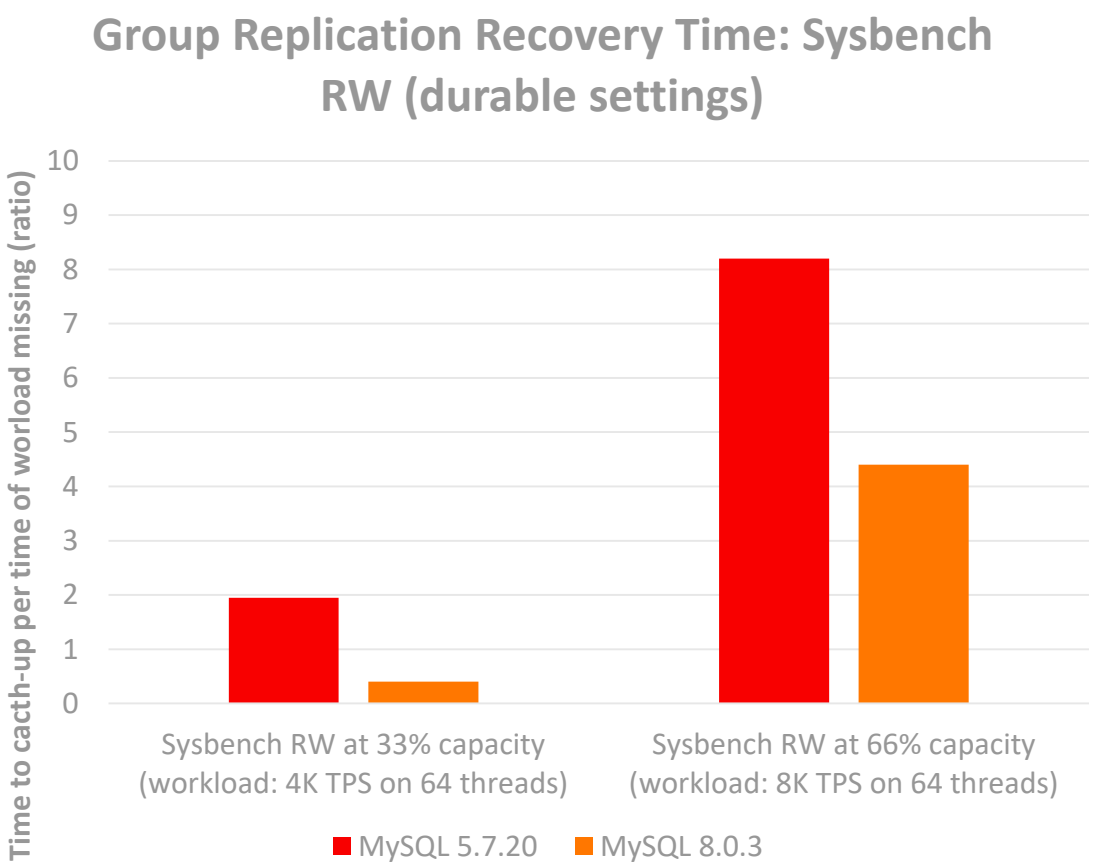
Write set parallelization

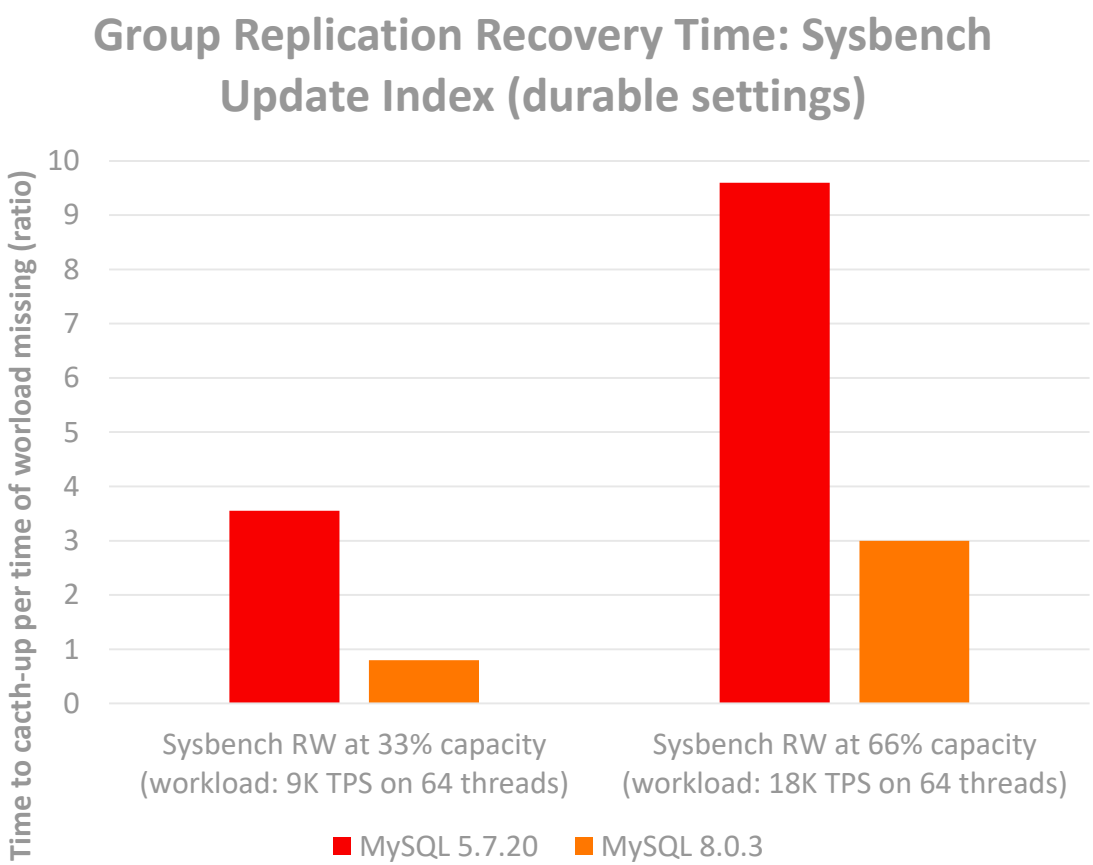


Backported to 5.7.22

Fast Group Replication Recovery

Replica quickly online by using WRITESET

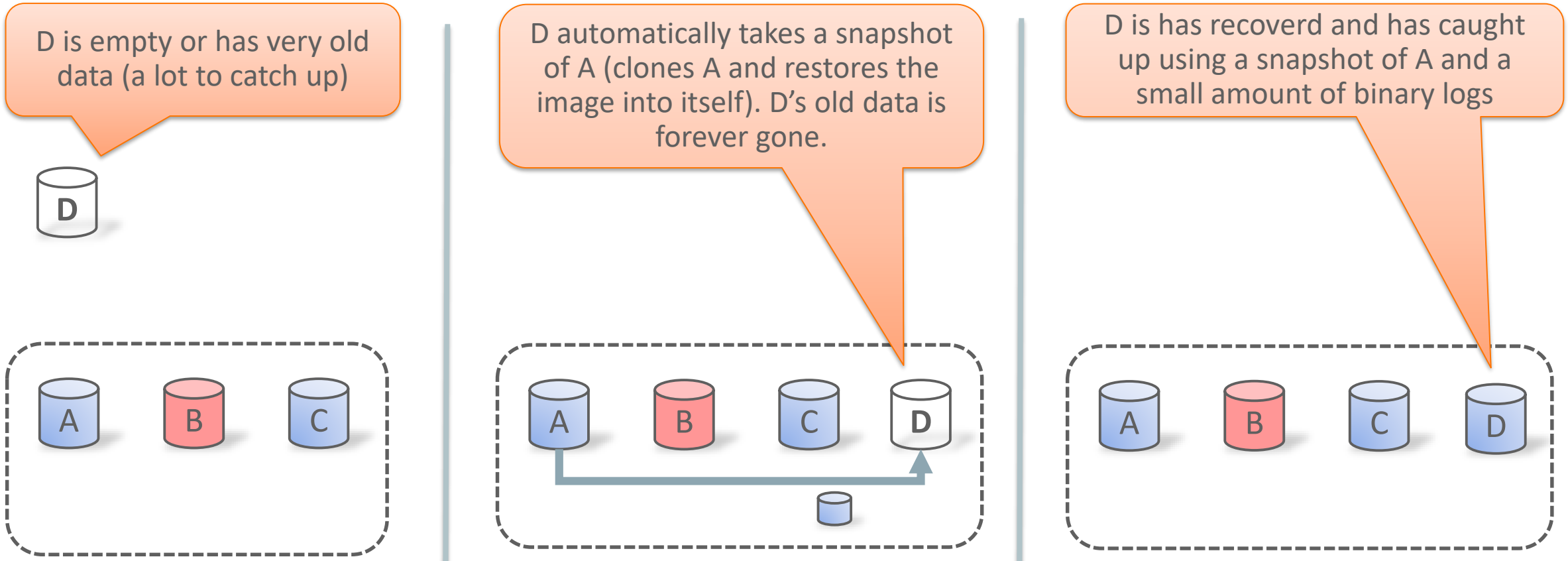




New in 8.0.17

Even Faster Group Replication Recovery: Clone Support

Empty or delayed replica quickly online by using Automatic Cloning and WRITESET



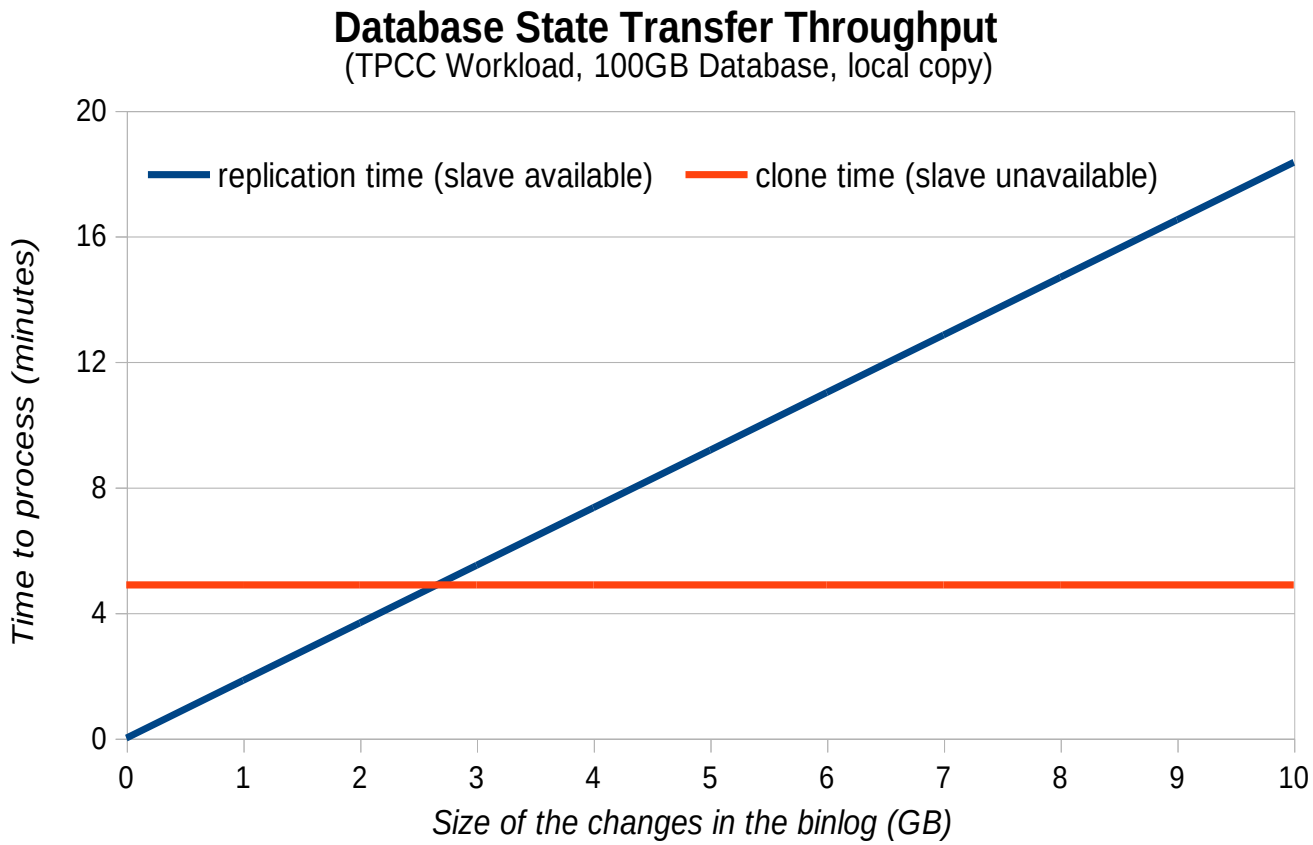
Requires that the **clone plugin** is **installed**. The clone plugin is shipped with MySQL 8.0.17.

New in 8.0.17

Even Faster Group Replication Recovery

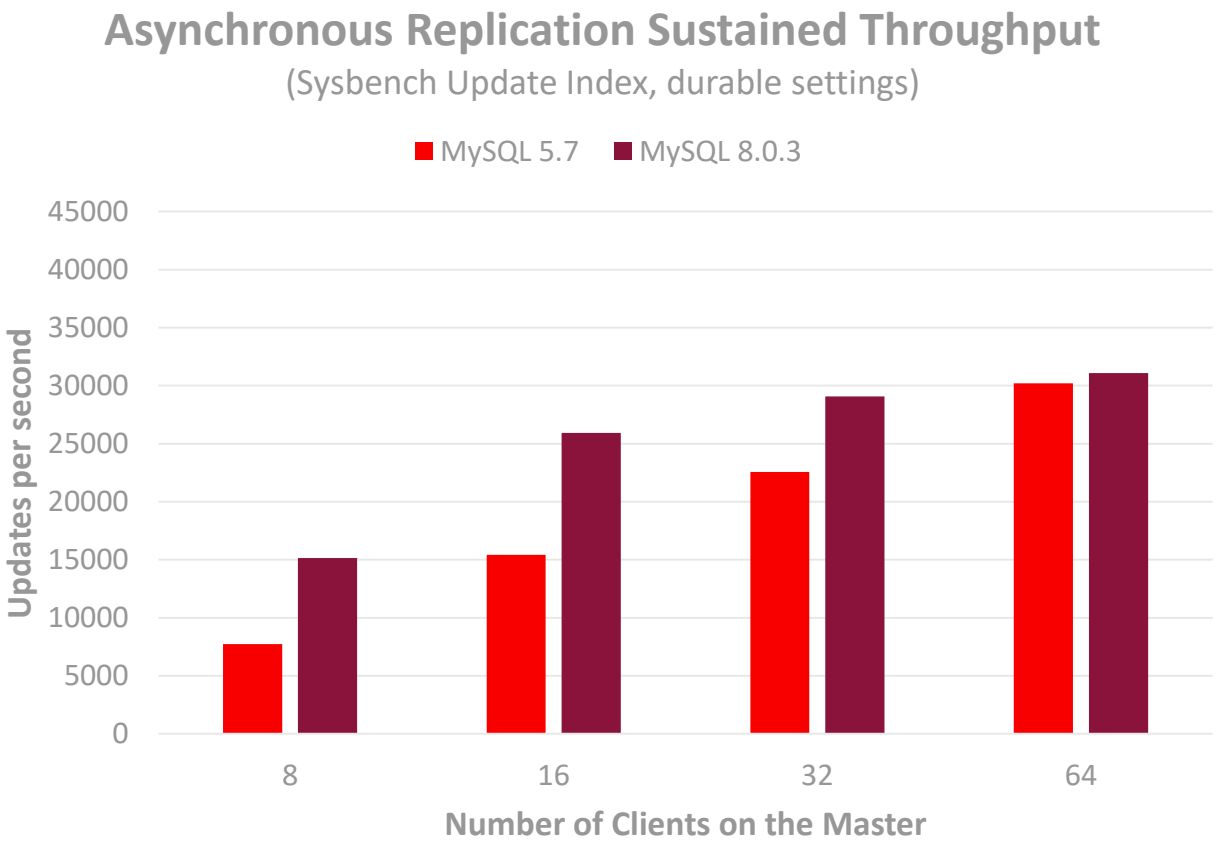
Empty or delayed replica quickly online by using Automatic Cloning and WRITESET

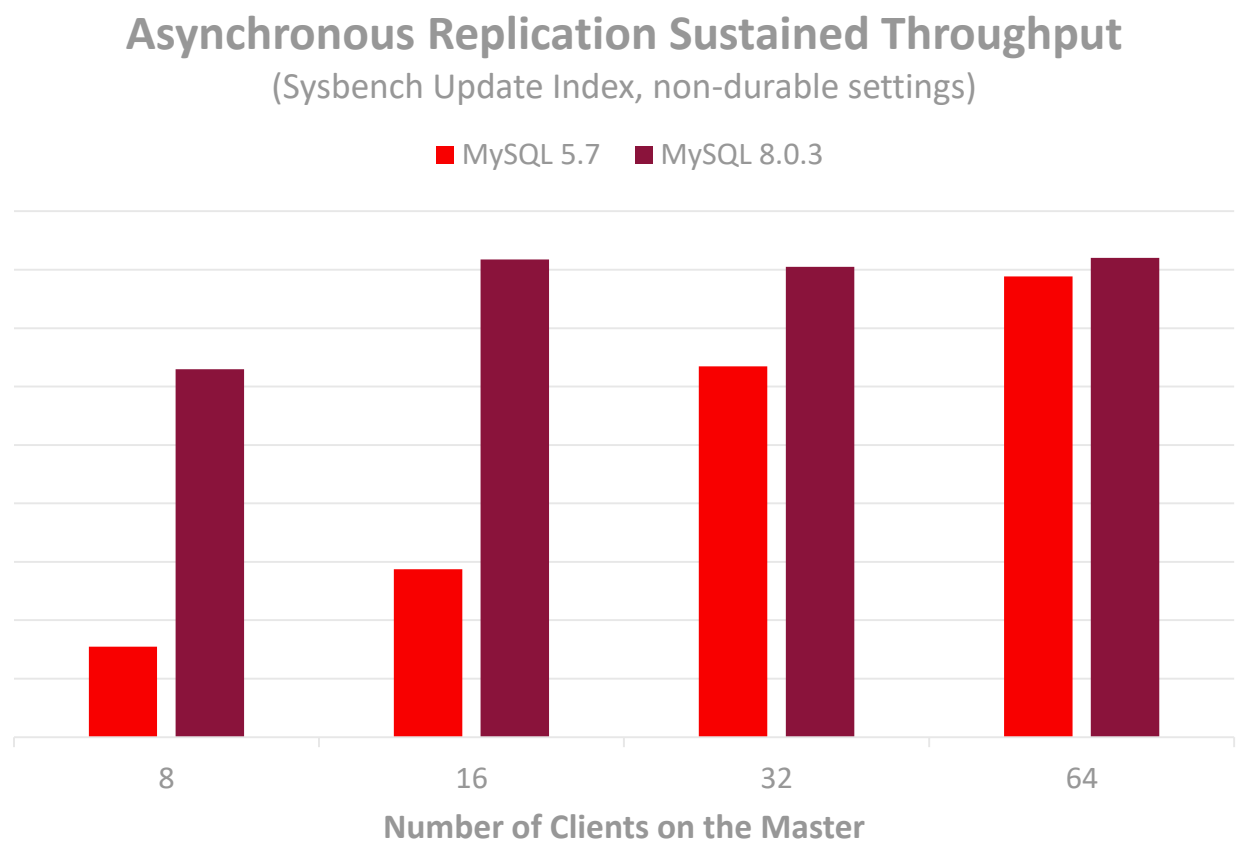
- Recovery using binary logs only vs recovery using clone and binary logs together.
  - There are cases binary logs are quicker and cases clone together with binary logs take less time.
- No Network involved.
  - Network adds latency
  - Network may not impact throughput (if it is not a bottleneck).



High Cluster Throughput

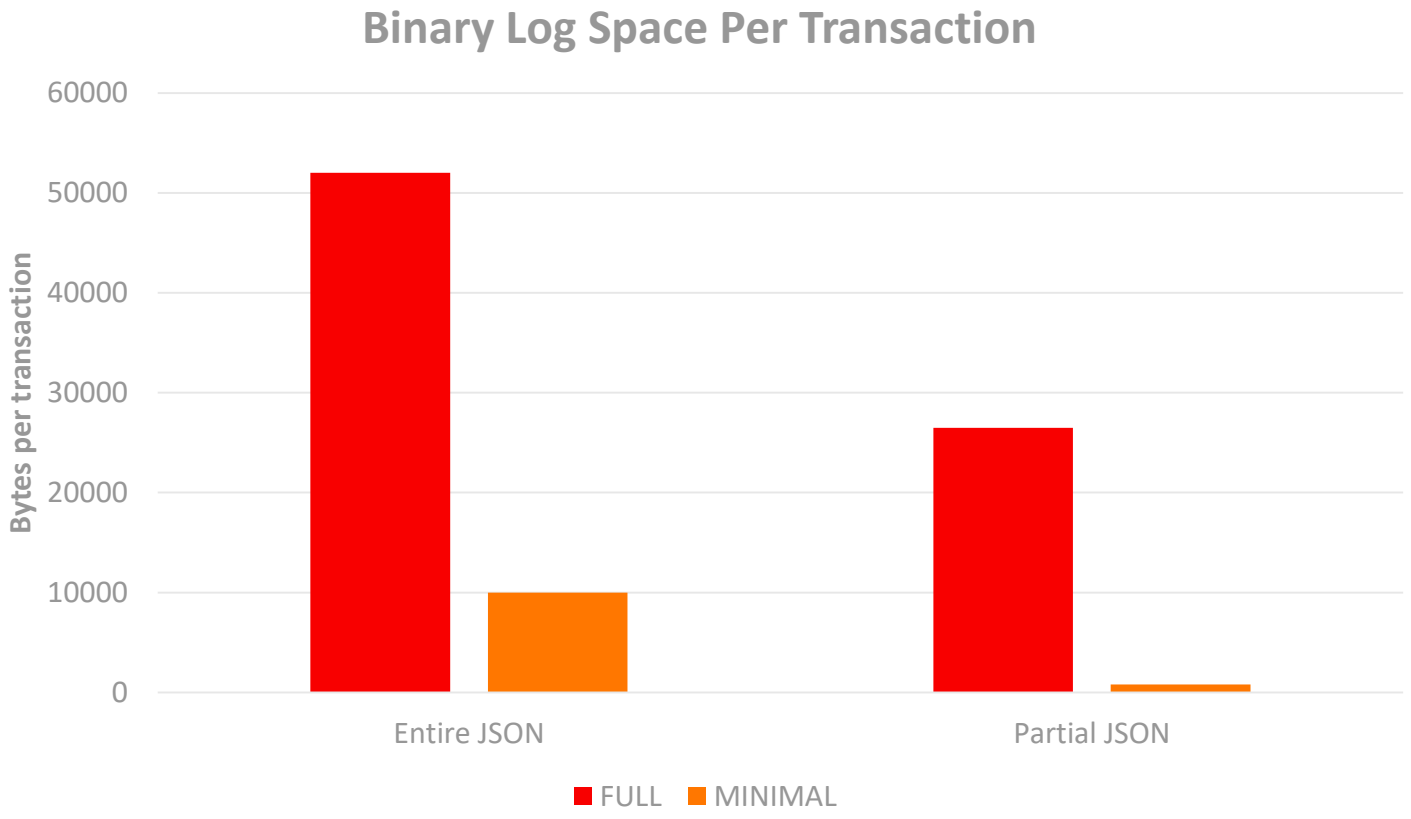
More transactions per second while sustaining zero lag on any replica





Efficient Replication of JSON Documents

Replicate only changed fields of documents (Partial JSON Updates)

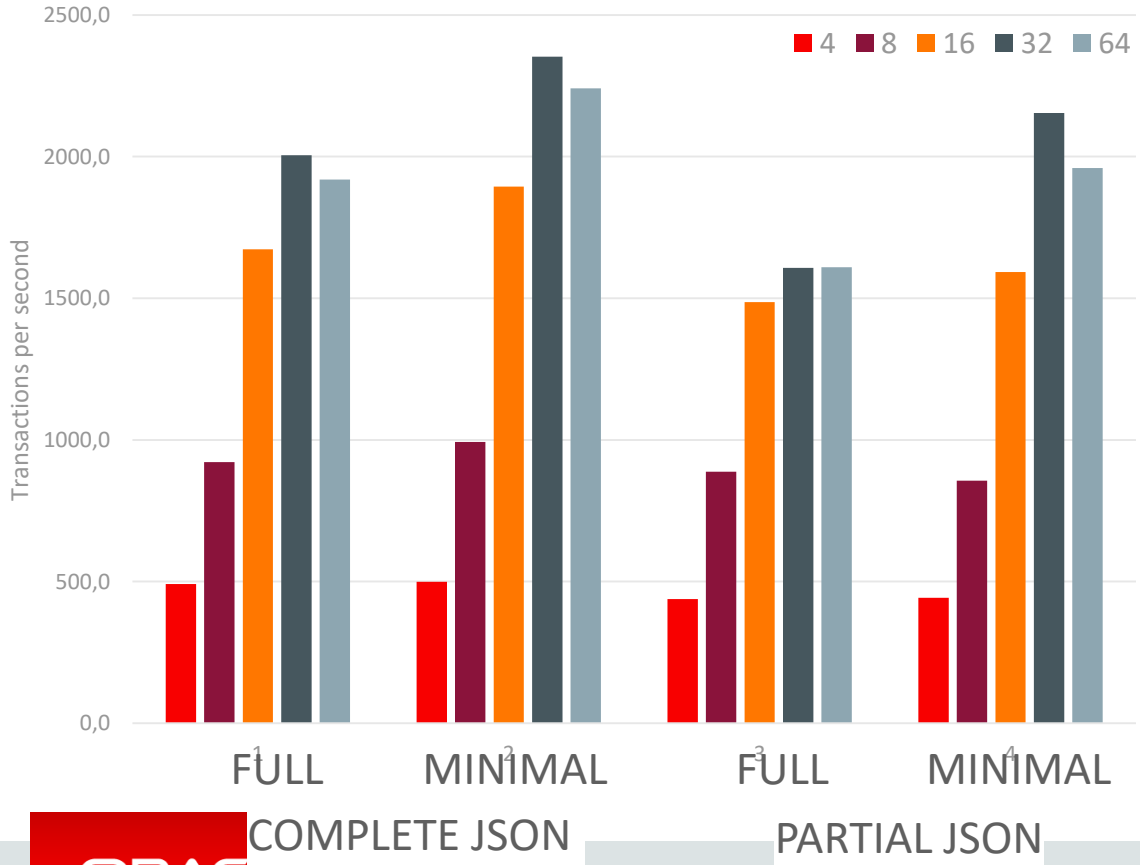


- Numbers are from a specially designed benchmark:
  - tables have 10 JSON fields,
  - each transaction modifies around 10% of the data

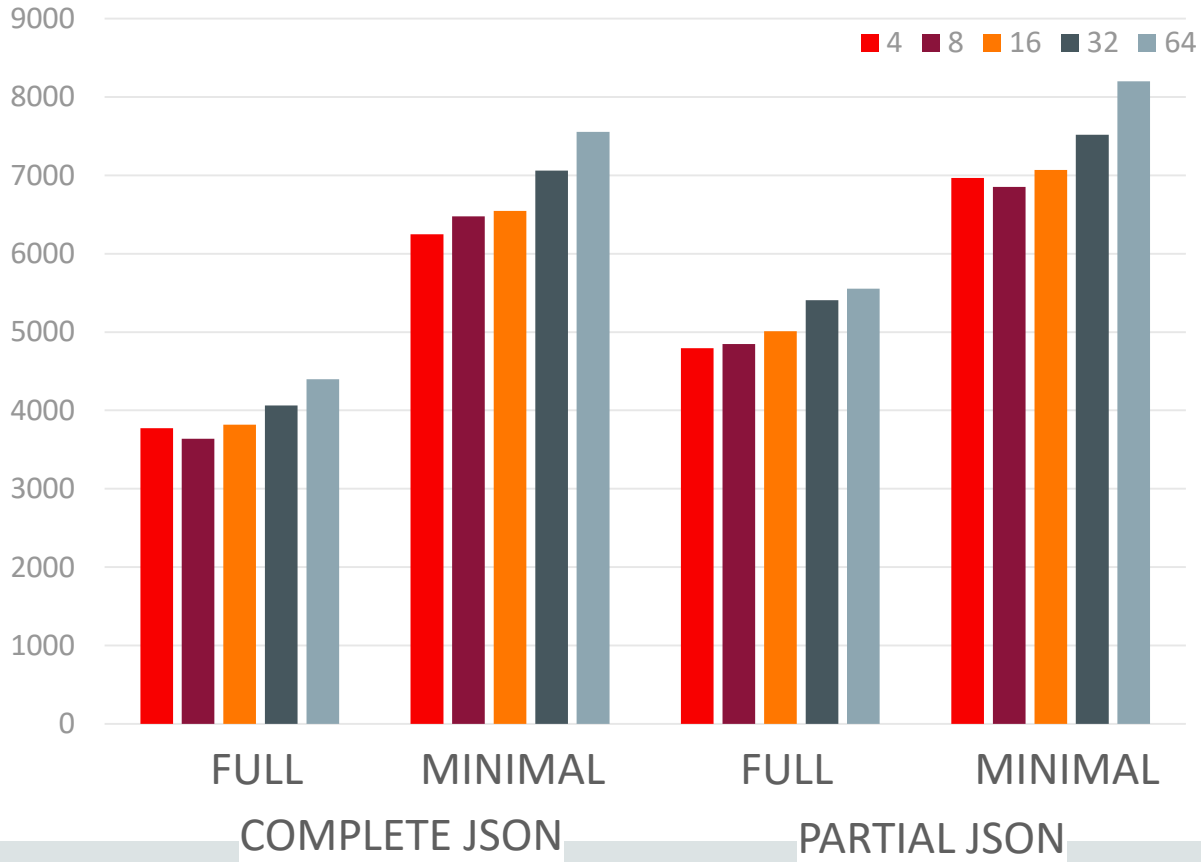
Efficient Replication of JSON Documents

Replicate only fields of the document that changed (Partial JSON Updates)

Throughput on the Master:
Partial JSON vs Complete JSON



Throughput on the Slave:
Partial JSON vs Complete JSON

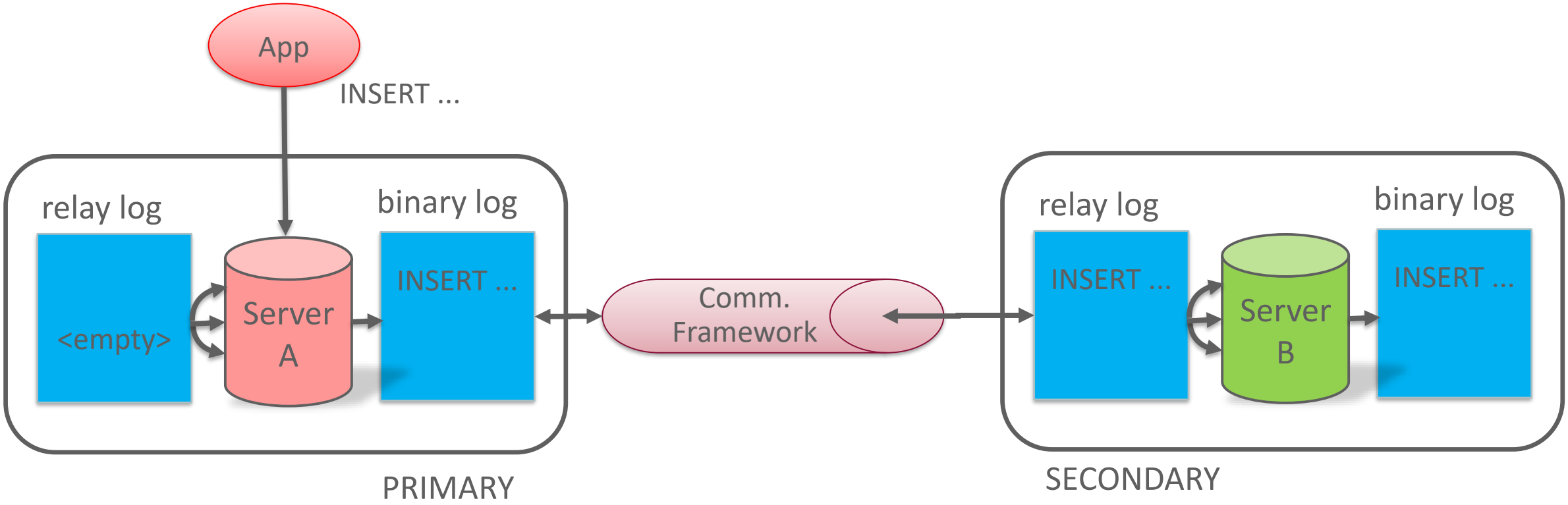


3 Enhancements in MySQL 8 (and 5.7)

- 3.1 Consistency
- 3.2 Operations
- 3.3 Monitoring
- 3.4 Performance
- 3.5 **Security**
- 3.6 Other

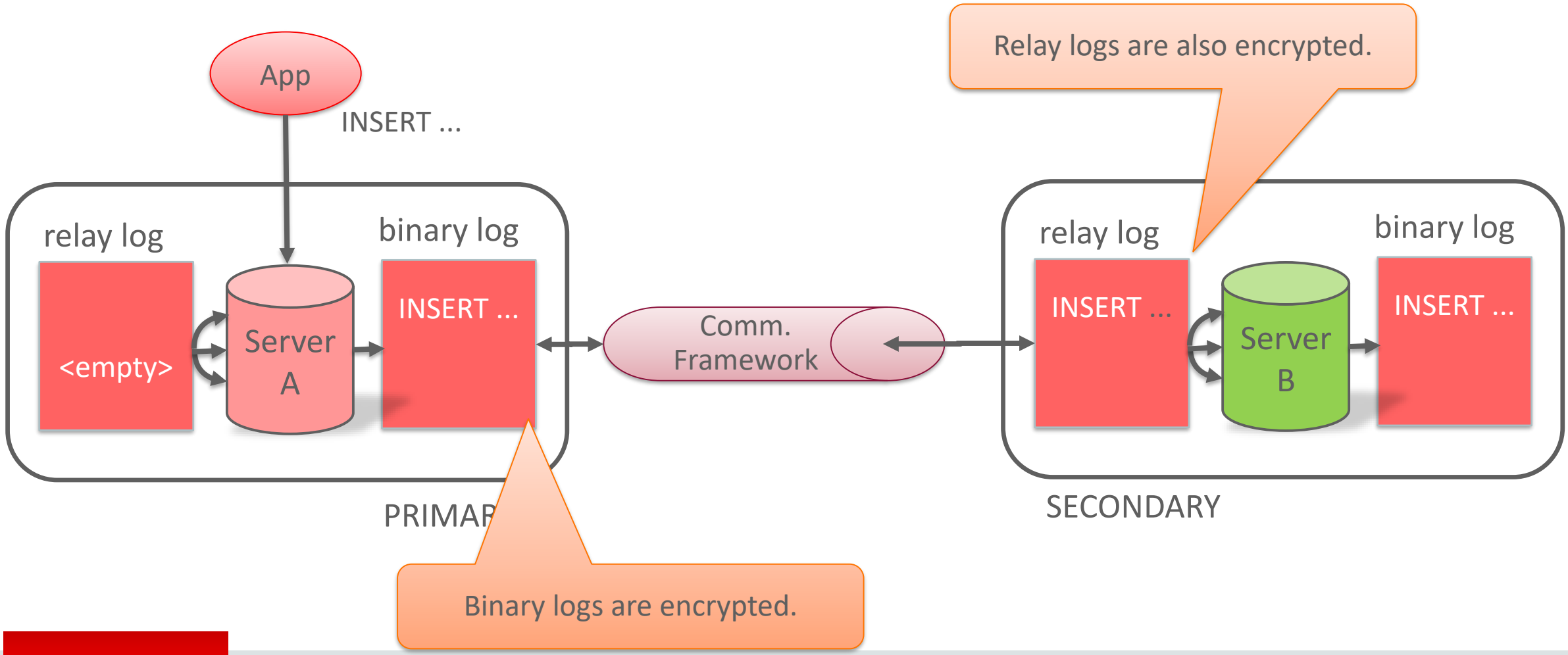
New in 8.0.14

Binary Log Encryption on Disk



New in 8.0.14

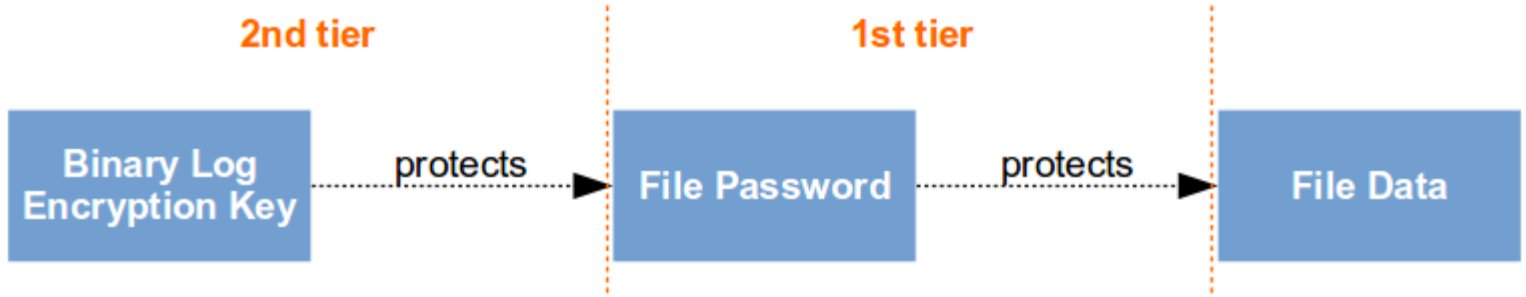
Binary Log Encryption on Disk



New in 8.0.14

Binary Log Encryption on Disk

- Protects Binary Log Data at rest.
- Controllable Using a System Variable: ***binlog_encryption***
- Two tier encryption, one master key and one key per file.

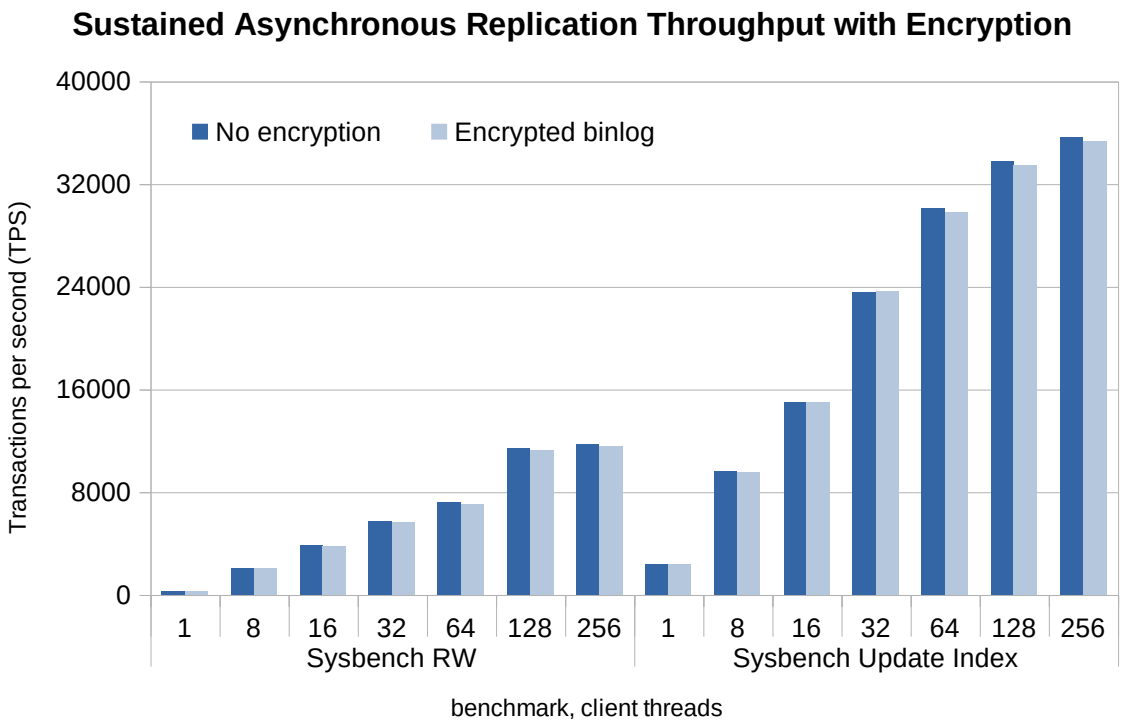


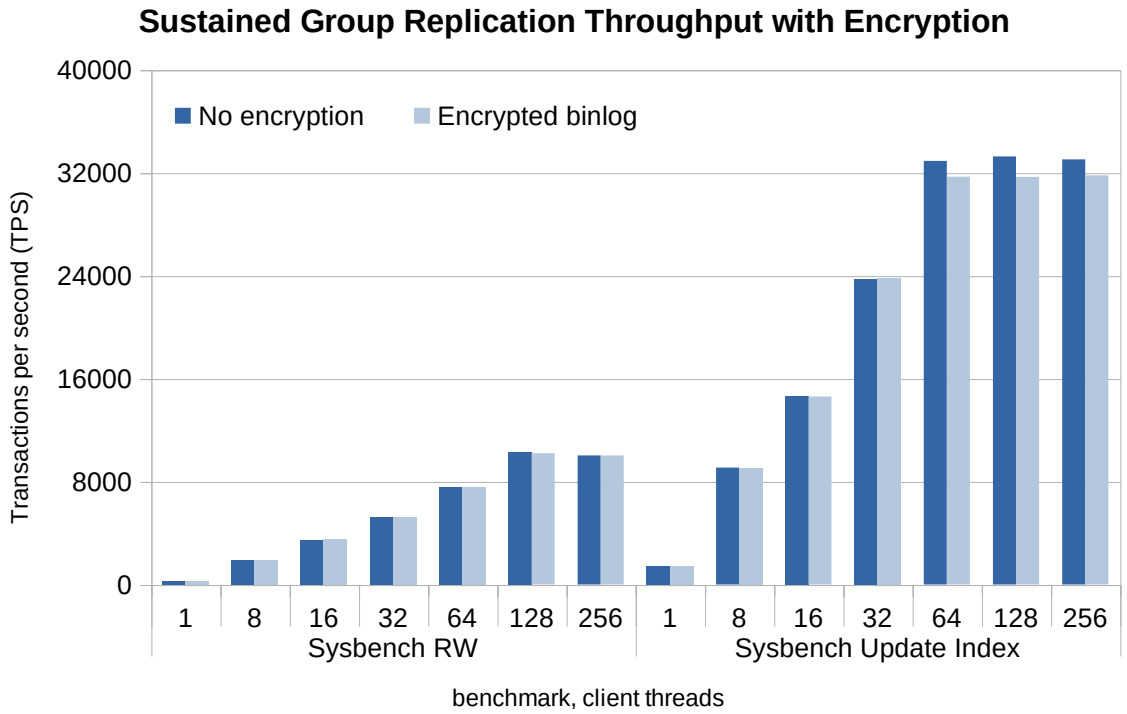
- Rotate the Binary Log Encryption Key:

ALTER INSTANCE ROTATE BINLOG KEY New in 8.0.16

Performance

- Marginal impact on throughput
  - More visible when the commit rate is higher.





3 Enhancements in MySQL 8 (and 5.7)

- 3.1 Binary Log Enhancements
- 3.2 Operations
- 3.3 Monitoring
- 3.4 Performance
- 3.5 Security
- 3.6 Other

Changes to defaults in MySQL 8

High performance replication enabled out-of-the-box

- Binary log is **on** by default.
- Logging of slave updates is **on** by default.
- Replication metadata is stored in **InnoDB tables** by default instead of files.
- Row-based applier falls back into **hash scans** to find rows instead of table scans.
- Transaction write-set extraction is **on** by default.
- Binary log expiration is set to **30 days** by default.
- Server-id is set to **1** by default instead of **0**.

Other MySQL 8 Group Replication Enhancements

- Monitoring: Group Replication threads instrumented and shown in performance schema
- Monitoring: Group Replication conditional variables and mutexes instrumented and shown in performance schema
- Operations: SAVEPOINT support when write sets are being extracted — Backported to 5.7.19
- Operations: Support hostnames in Group Replication whitelist — Backported to 5.7.21
- Operations: New options to fine tune the cluster automatic flow control.
- Operations: Cross-version replication policies for GR — New in 8.0.17

Other MySQL 8 Group Replication Enhancements

- Performance: More efficient code path between network layer and replication layer.
- Performance: Configurable communication pipeline size for group replication.
- Performance: Enhanced large transactions support for group replication.
- Performance: support for protocol compression in mysqlbinlog
- Troubleshooting: Dynamic and high performance debugging of group replication inter-node messaging
- Security: Encryption of transient replication files.
- Infrastruture: IPv6 support.

New in 8.0.13

New in 8.0.16

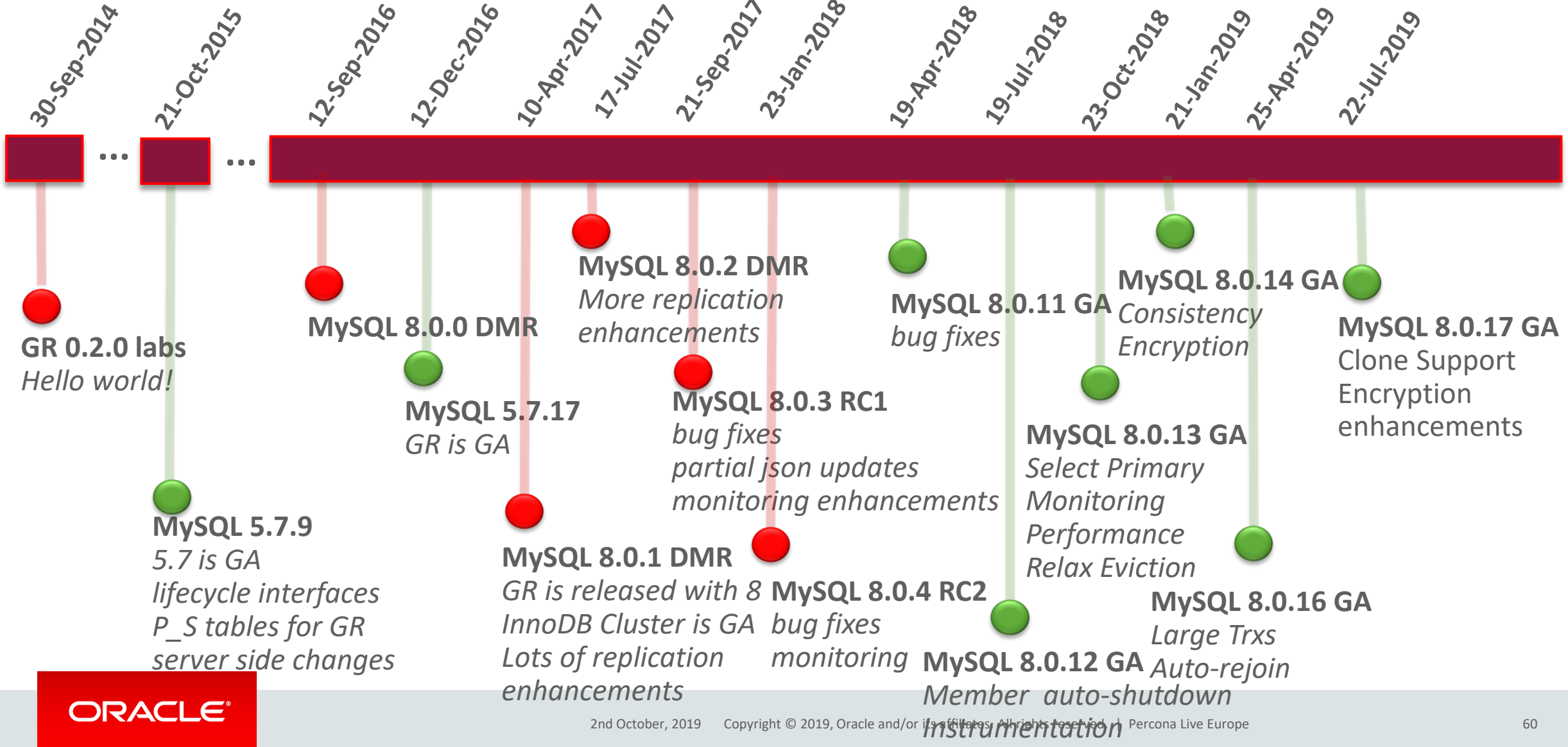
New in 8.0.17

New in 8.0.17

New in 8.0.14

4 Roadmap

The Road to MySQL 8 Group Replication and InnoDB Clusters



5 Conclusion

Conclusion

Latest MySQL 8 GA is out:

- Performance/efficiency improvements
  - Automatic Cloning of donors in Group Replication means one less provisioning step required from the user.
- More encryption features
  - Encrypt even transient replication data that touches the disk.
- Improved Operatinos and DBA experience
  - Mysqlbinlog supports protocol compression
  - Enhanced cross-version replication protection in Group Replication.
  - Enhanced distributed recovery by integrating the clone plugin. **Seamless and automatic snapshotting, provisioning and catch up**. Reduced operations overhead.

Where to go from here?

- Packages
  - http://www.mysql.com/downloads/
- Documentation
  - https://dev.mysql.com/doc/refman/8.0/en/
- Blogs from the Engineers (news, technical information, and much more)
  - http://mysqlhighavailability.com


ORACLE®