Percona XtraDB Cluster (PXC) 101

Introduction to Percona XtraDB Cluster and some features.

Marcelo Henrique Gonçalves
Managed Services MySQL DBA @ Percona
Percona University São Paulo
27/04/2019

PERCONA

Percona XtraDB Cluster (PXC) 101

This presentation is based on a Webinar: Percona XtraDB Cluster (PXC) 101 by Krunal Bauskar.

- The PXC and the world
  - *Introduction to the software and it's way of replication*
  - *Limitations?*
  - *Is it usable for HA?*

- Exclusive features in PXC compared to Galera/MariaDB

© 2019 Percona

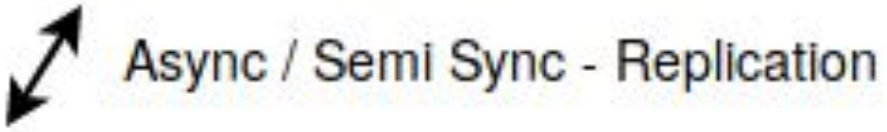
Agenda

- Common ways to replicate data
- Ways of replicating async
- Synchronous Replication
- PXC
- How PXC replicates
- Limitations of PXC
- Some Features (5.7)
- HA solution using PXC

© 2019 Percona

Common ways to replicate data

- Replication built-in in MySQL, MariaDB or Percona Server: Binlog (GTID or not).
- Either ASYNChronous or SEMI-SYNChronous (not counting with NDB Cluster Synchronous replication).
- Asynchronous is that writes will be done in a replica without impacting master and application commit only guarantees data in the Master, not replica.



© 2019 Percona

Common ways to replicate data

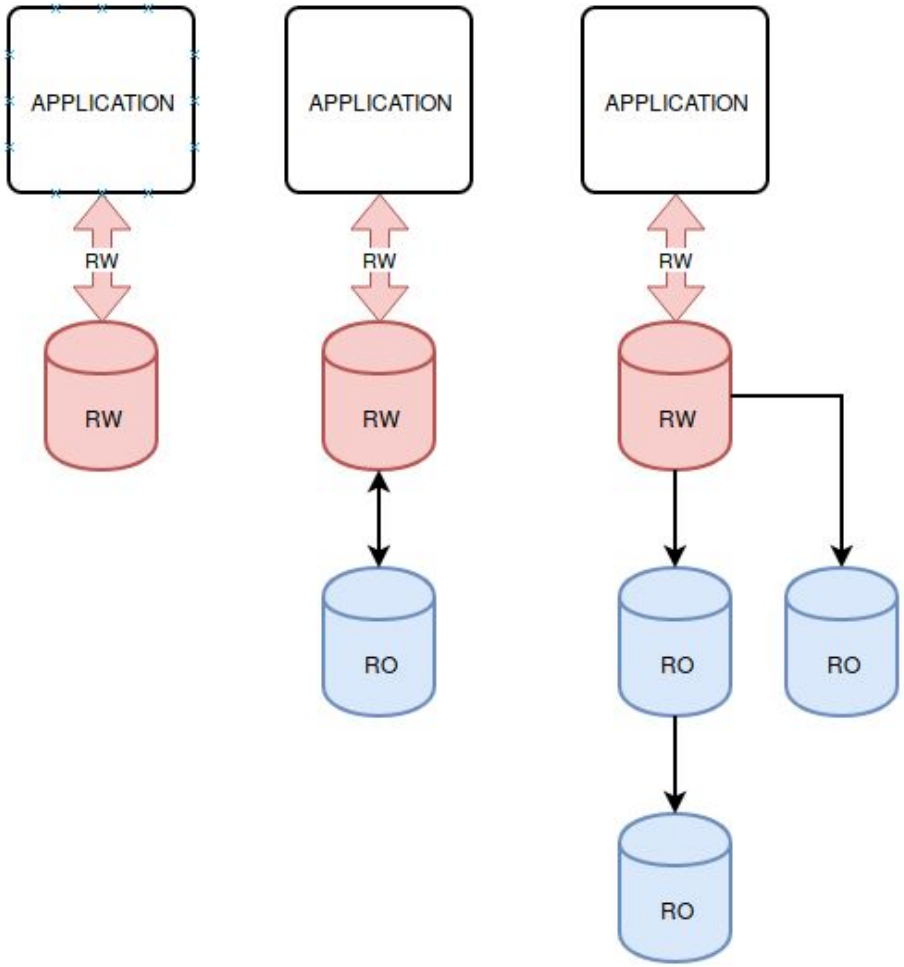
In my point of view:

- **Semi-synchronous is just a nice name for Asynchronous.**

- **If the master loses connectivity with slave, it will keep allowing writes.**

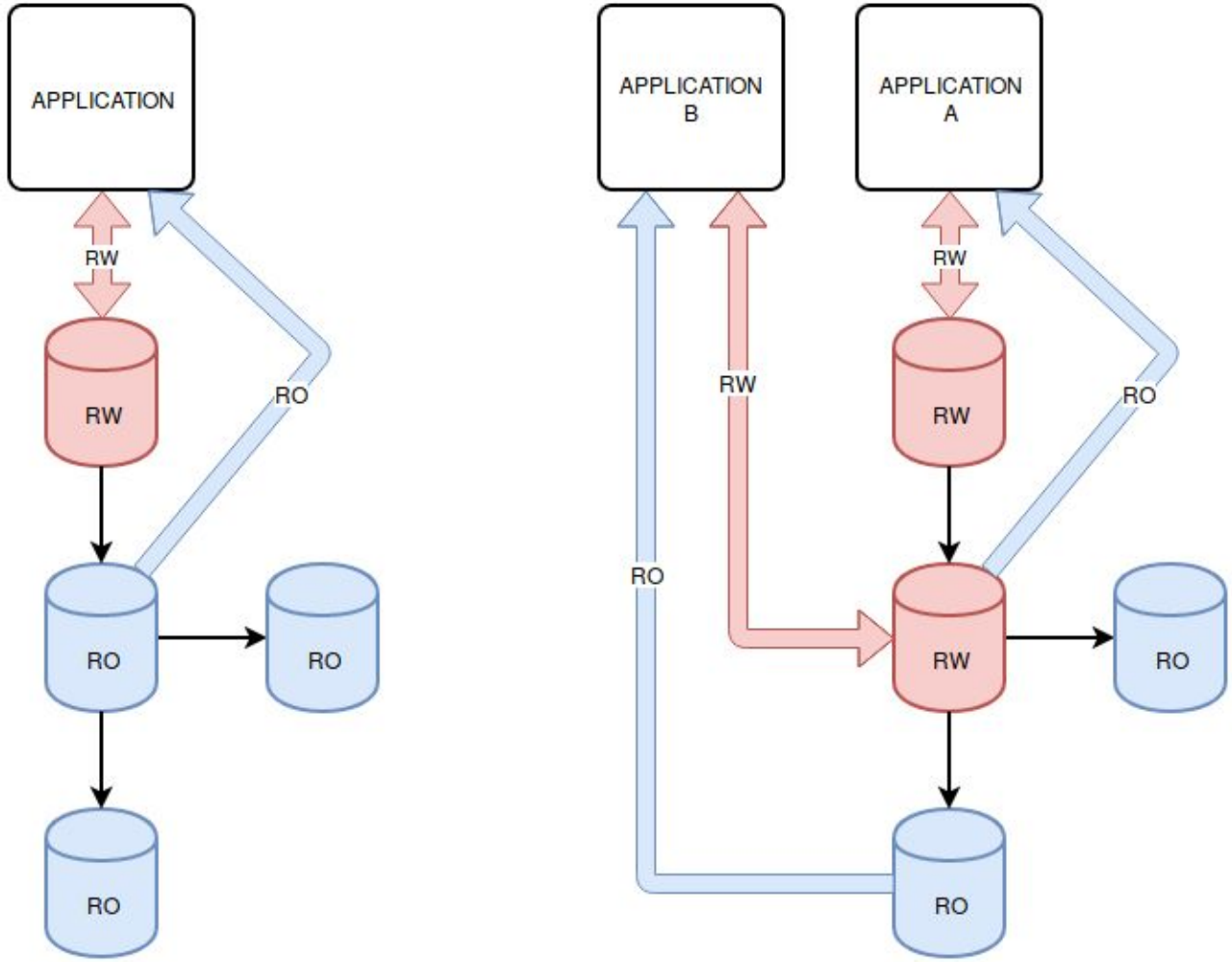
© 2019 Percona

Ways of replicating async



© 2019 Percona

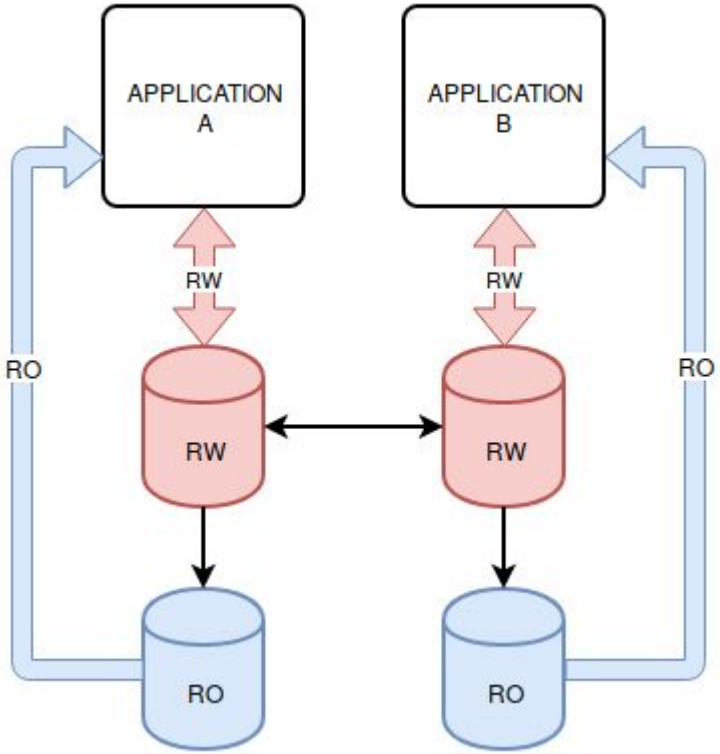
Ways of replicating async



© 2019 Percona

PERCONA

Ways of replicating async



+ Creativity
= Several Topologies

© 2019 Percona

Ways of replicating async

- Replication delay (possible loss of transactions). Not rarely, relaxed durability is used - sync_binlog.
- Ok read scalability - subject to high data latency.
- No horizontal write-scalability.
- Switchover interval, even being small.
- No protection against network failure.
- Master + Master = Can be dangerous

© 2019 Percona

Synchronous Replication

We can’t get rid of CAP theorem BUT we can have:

- All nodes open for read/write (1)
- Great scalability in read
- No single point of failure - Tightly coupled
- Easy to maintain/deploy - Automatic Node Provisioning
- Protect against network outages (2)
- All nodes as an active master (3)

1 - Multithreaded replication limited to slower node and network latency.
2 - Partial network outages, subject to Quorum
3 - Avoid same table write in multiple nodes

© 2019 Percona

PXC

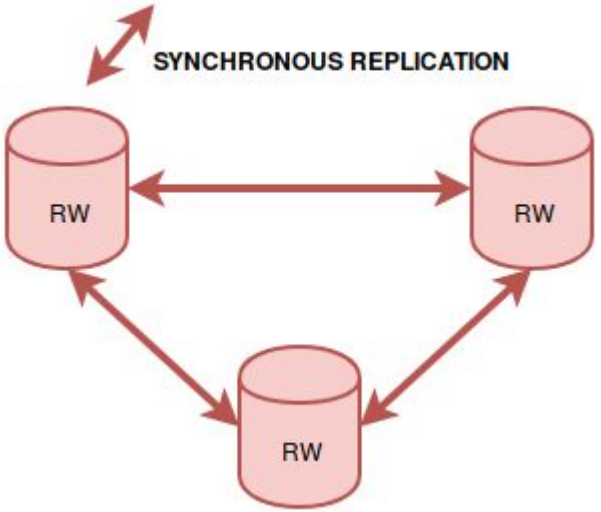
- PXC - multi-master solution with Percona Server
- Galera plugin - from Codeship
- Percona product -> independent and has complete life cycle.

1. Consistency
2. Availability
3. Partition Tolerance

Percona

XTRADB CLUSTER

Percona XtraDB Cluster is an active/active high availability and high scalability open source solution for MySQL clustering.

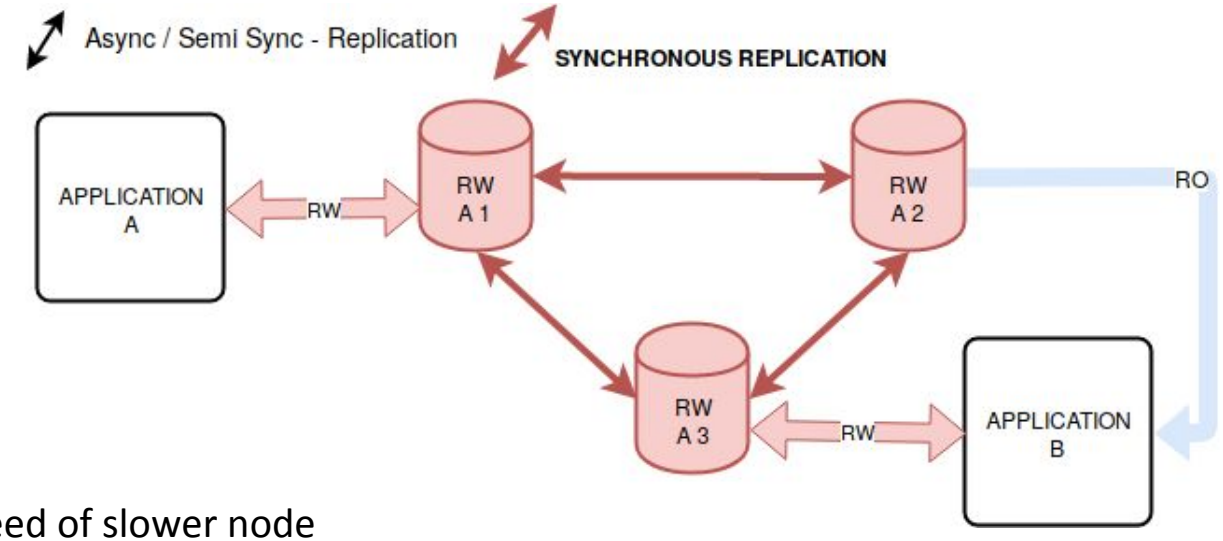


© 2019 Percona

PERCONA

PXC

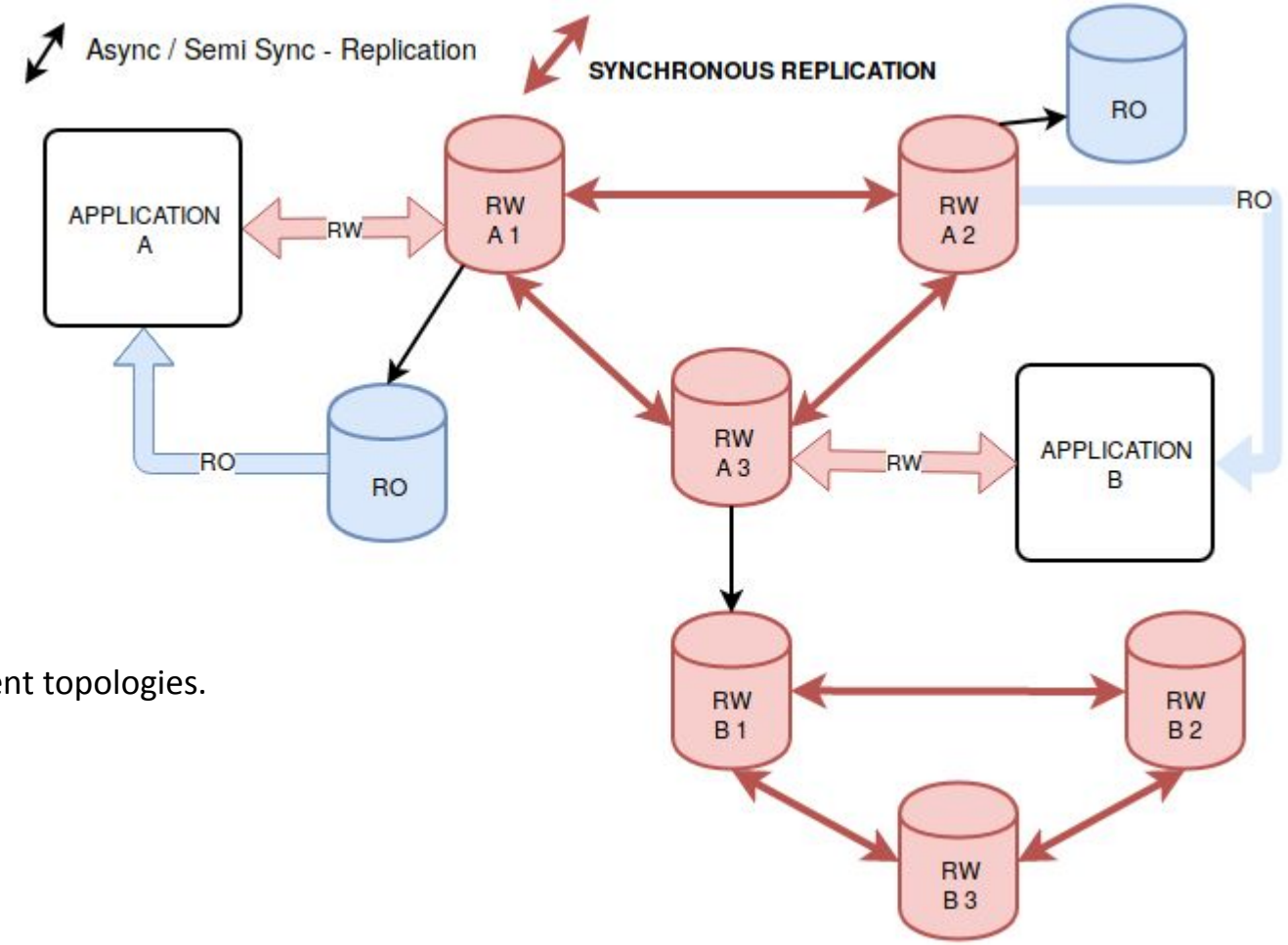
- **Write to any node(1)**
- **Read from any node**
- **CURRENT(2) AND CONSISTENT**



1 Subject to deadlocks, first committer wins, speed of slower node
2 Subject to <fc_limit> number of write sets.

© 2019 Percona

PXC



Can be used in many different topologies.

© 2019 Percona

PERCONA

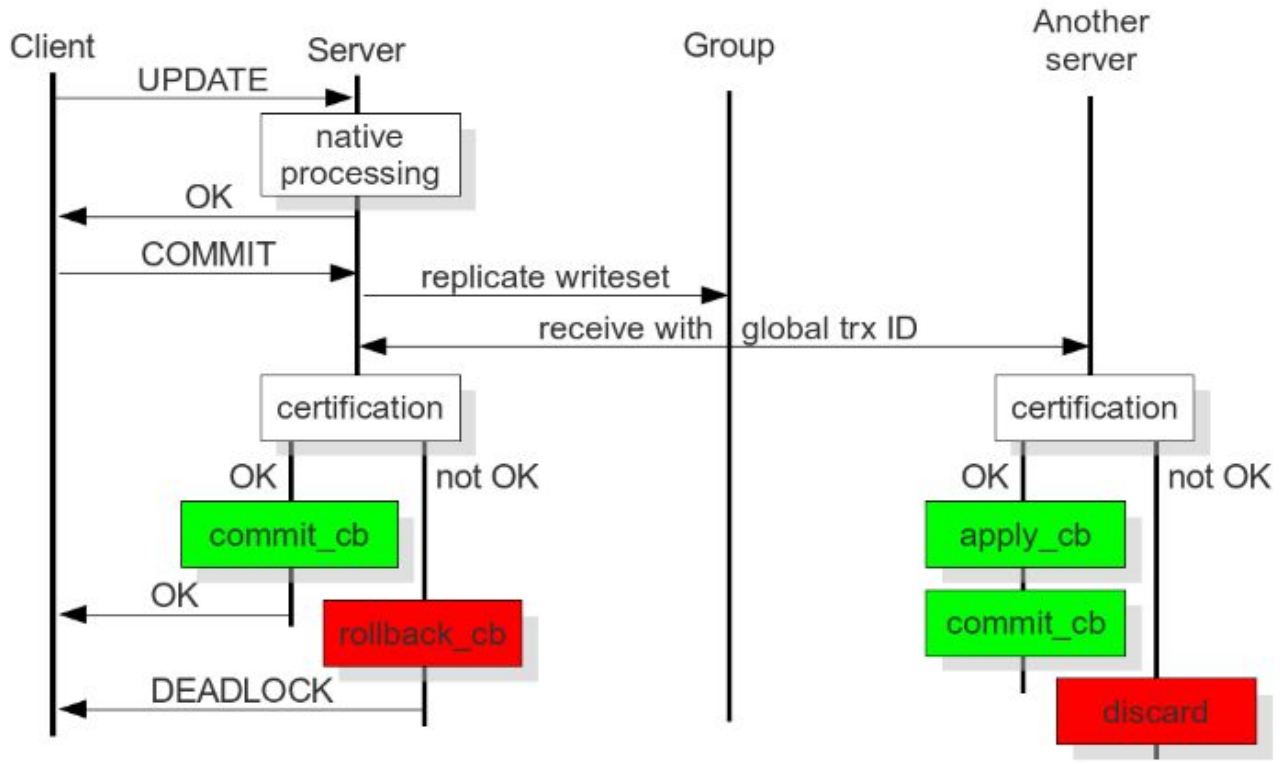
PXC - enterprise ready

- Automatic Node Provisioning
- Ability to handle conflicting workload
- Tunable flow control
- Parallel Processing
- Protection against network failure
- Can be Geo-distributed
- Cluster safe-mode
- Security
- Performance
- Trackability - easy troubleshoot

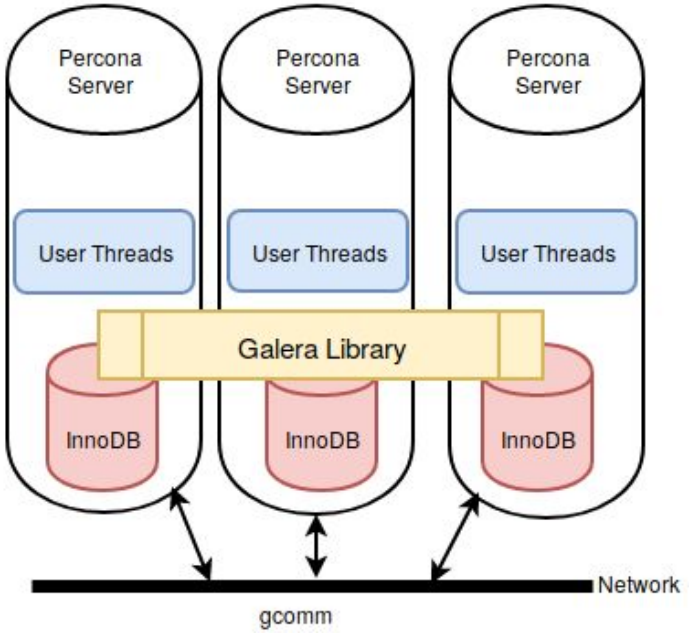
© 2019 Percona

How PXC replicates - Certification Based

Source: http://galeracluster.com/documentation-webpages/certificationbasedreplication.html



Copyright: 2014 Codership Ltd.



Layer 1: Percona Server (Write-set)
Layer 2: Galera Plugin
Layer 3: gcomm communication channel
Write-sets are written in GALERA CACHE (a file in each node)

© 2019 Percona

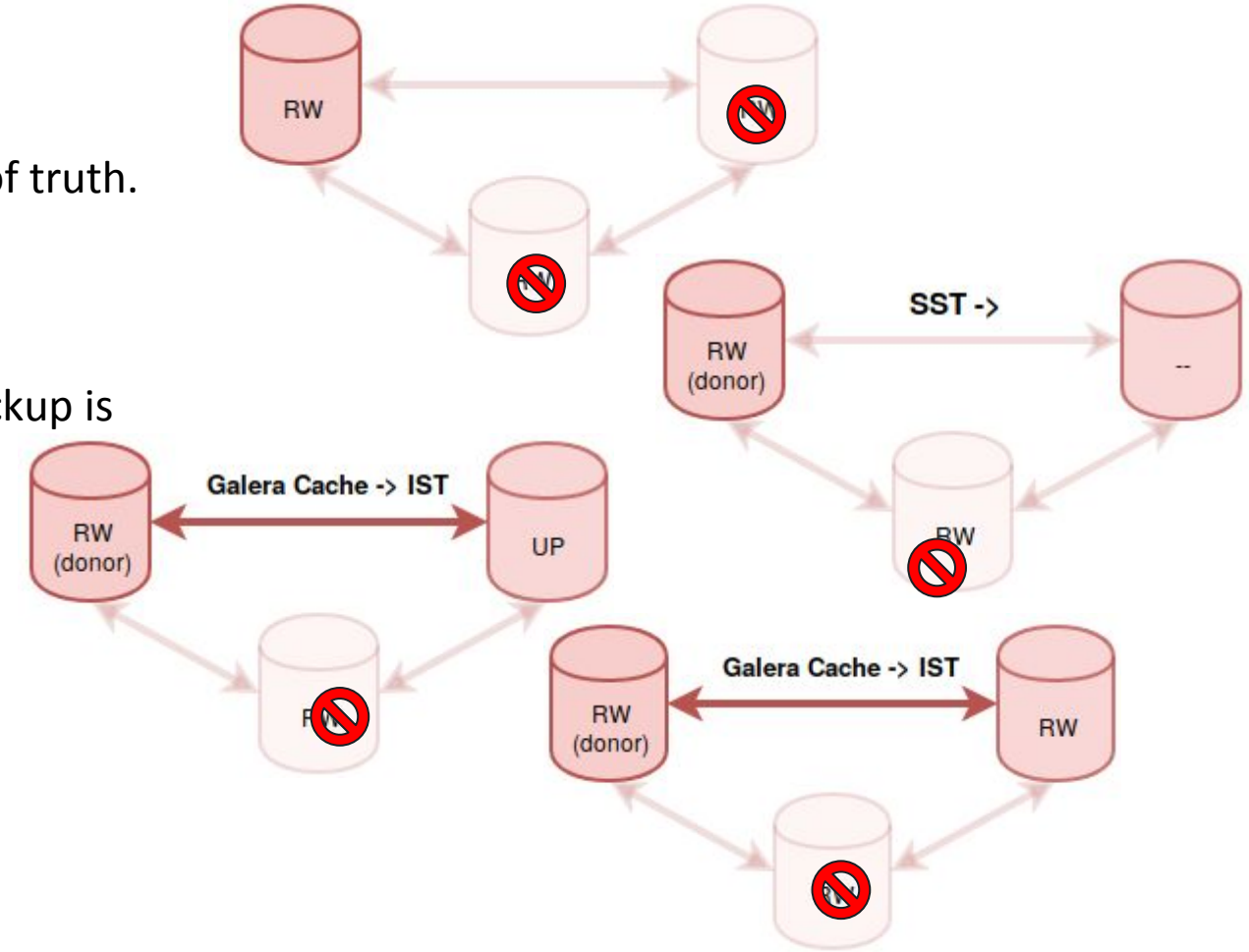
How PXC replicates - Steps SST + IST

- First node needs bootstrap `(systemctl start mysql@bootstrap)`
- The node that starts with the bootstrap is the source of truth.

- A new node will run SST - State Snapshot Transfer (mysqldump/rsync/xtrabackup)
- Any existing data in next node is erased, a new full backup is restored and it is started - joining the cluster.

- New node perform Incremental State Transfer - IST
- Reads galera cache from donor, and apply.

- New node is now synced with cluster.
- New node is Writable.



[Understanding how an IST donor is selected]

© 2019 Percona

How PXC replicates - Steps IST

Sample case:

1. Node 1 goes down - and starts again and join the cluster
2. Node 1 needs write-set starting from 14015 (already considering safe gap).
3. Node 2 galera cache has 14020, Node 3 galera cache has 13400
4. Node 1 SELECT node 2 as IST donor - Node 1 is YET not usable
5. Node 1 receives write-sets and apply - IST
6. Node 1 gets in sync and enabled Read Write again.

- If there is no node with galera cache old enough, a SST is done by Node 1.
- Can be avoided with gcache.freeze_purge_at_seqno

© 2019 Percona

How PXC replicates - workload

- Optimistic Lock - no distributed locks
- First committer win.
- Conflicting transaction forceful abort
- Certification Failure - can't allow 2 conflicting transactions.
- Flow control dynamically controlled.

* Transactions are queued - queue full triggers flow control

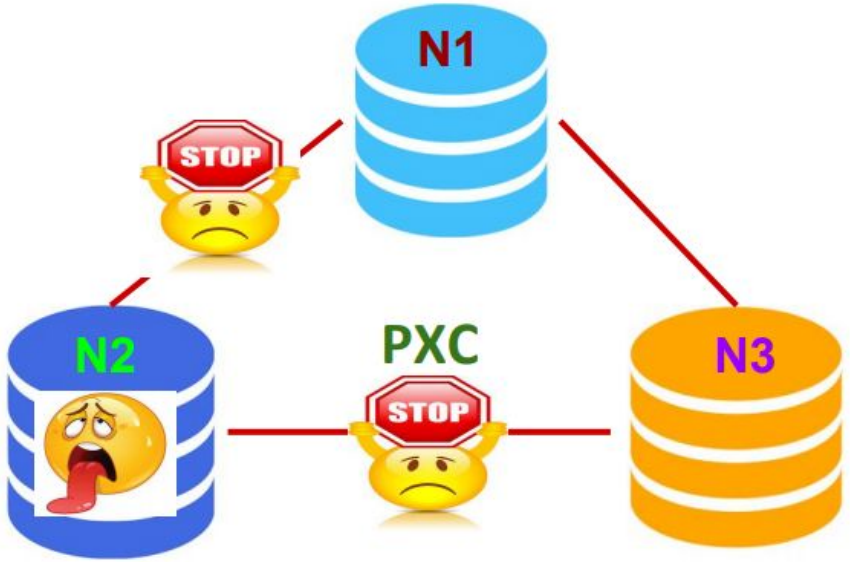


Image by **Krunal Bauskar**

© 2019 Percona

How PXC replicates - Parallel processing

- Parallel worker threads - works without collisions
- Coordinated to allow first commit win
- Online tunnable
- Low resources consumption
- Threads visible in processlist
- Only 1 rollback thread

© 2019 Percona

How PXC replicates - Quorum

- By default, each node votes 1 - Quorum is total of votes
- Votes online needs to be ½ +1 of total Cluster (talking to each other)

Cluster of 3 -> 2 up Cluster of 4 -> 3 up

Cluster of 5 -> 3 up Cluster of 6 -> 4 up

- Auto-recovery - Once nodes are back online -> join/sync
- Network variables configurable for different timeouts
- Protection against Split-Brain
- Geo distributed - with 'DC awareness' variable (gmcast.segment)
- Ability to have Arbitrator node - A node without data that only votes

© 2019 Percona

Limitations of PXC

- Works only with InnoDB
- Can’t use Lock tables, Unlock tables, GET_LOCK(), RELEASE_LOCK(), and similar
- Can’t log queries to table - log_output = FILE
- There is a limit for transactions size - LOAD DATA will commit every 10k rows
- Transaction can fail at commit stage
- XA transactions can’t be used
- Write throughput limited to slower node
- InnoDB fake changes feature not supported
- DELETE in tables without Primary Key not supported

Among other small ones in: https://www.percona.com/doc/percona-xtradb-cluster/LATEST/limitation.html

© 2019 Percona

Cluster Safe Mode

If workload that is not cluster safe: pxc_strict_mode acts:

ENFORCING

MASTER

PERMISSIVE

DISABLED

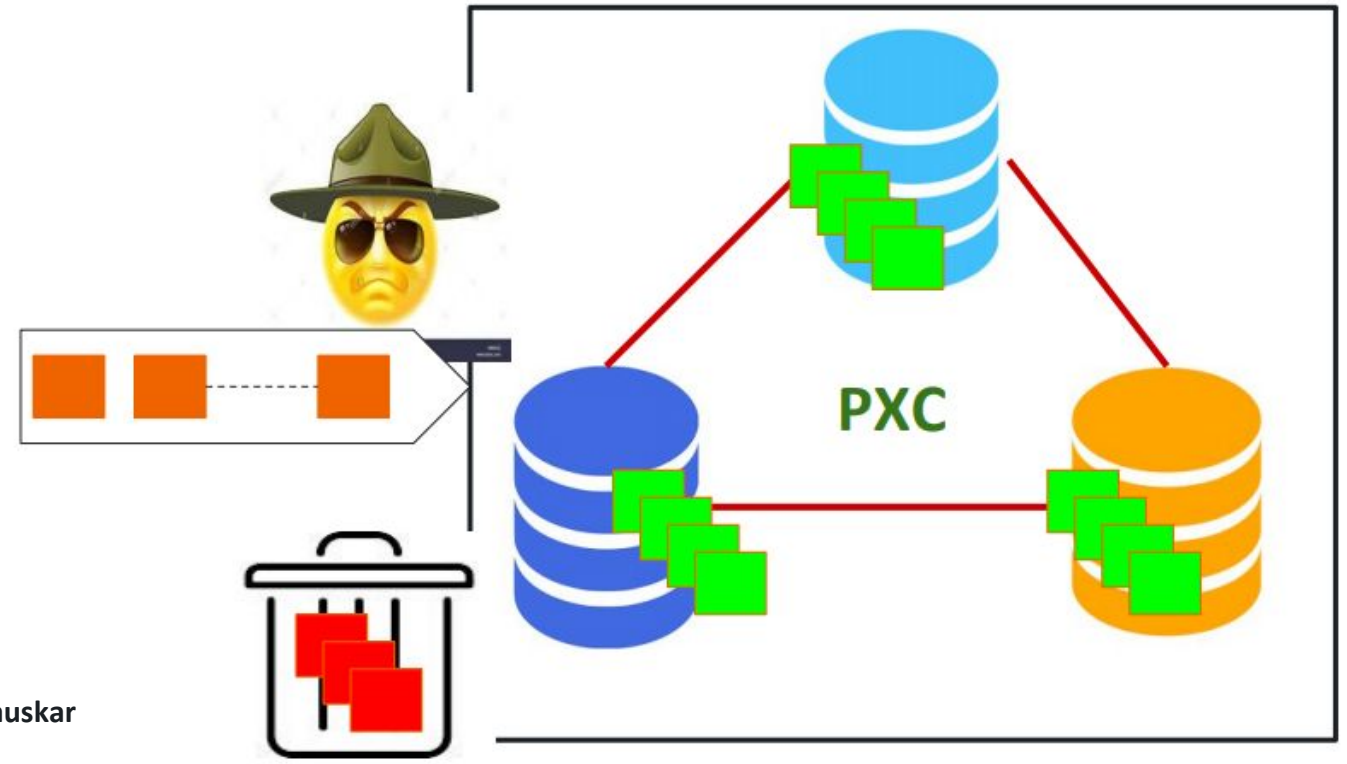


Image by **Krunal Bauskar**

© 2019 Percona

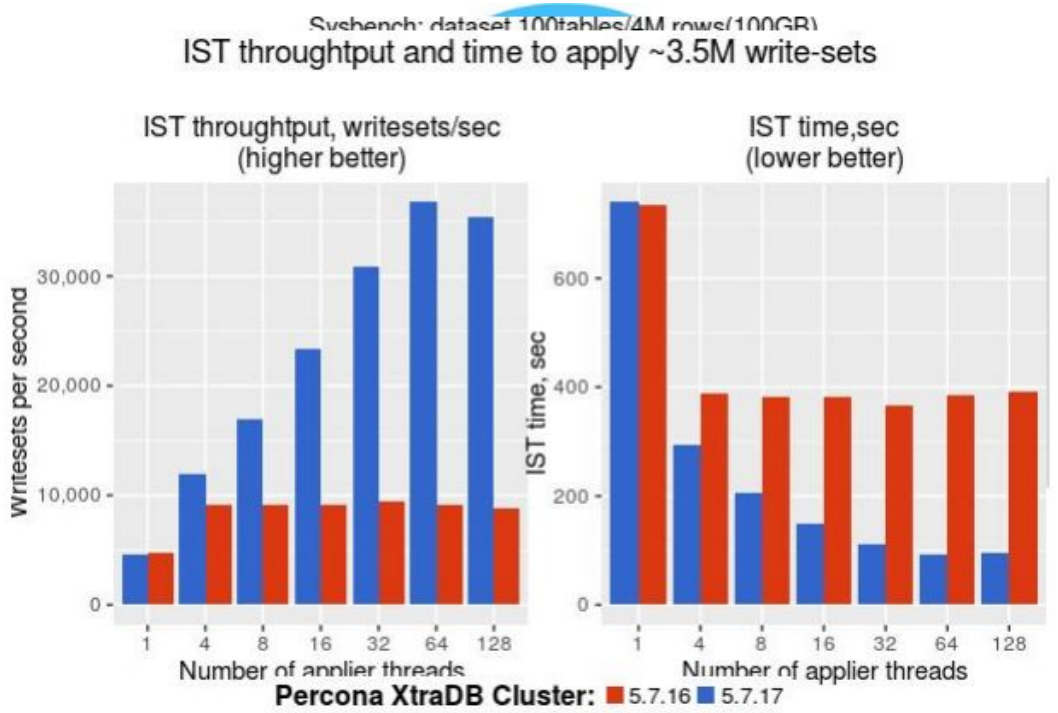
Security

- Compatible with At Rest tablespace encryption - 5.7
- SST + IST - can be done in encrypted channels
- Replication traffics - can be encrypted as well
- ALL security from PS applies!

© 2019 Percona

Performance

- Industry lead performance solution (5.7.17+)
- Fit several workloads - OLTP/POINT UPDATE/WH ...
- Quicker Node rejoins
- All Percona Server tuning
- Allow a Node to be Desync
- Allow a Node in Maintenance



© 2019 Percona

Performance - Trackability

- TRACK Node behavior in the cluster (delaying ot being delayed)
- Overload of replication replicas or network
- show status (Flow-control, IST progress, queue size, replication latency, applied/committed upto, cache size)
- performance_schema.pxc_cluster_view
- Other internal objects exposed through Perf. Schema (PFS)
- Clear error/information logs
- Improved debugging messages for SST
- Improved visibility of thread state through show processlist

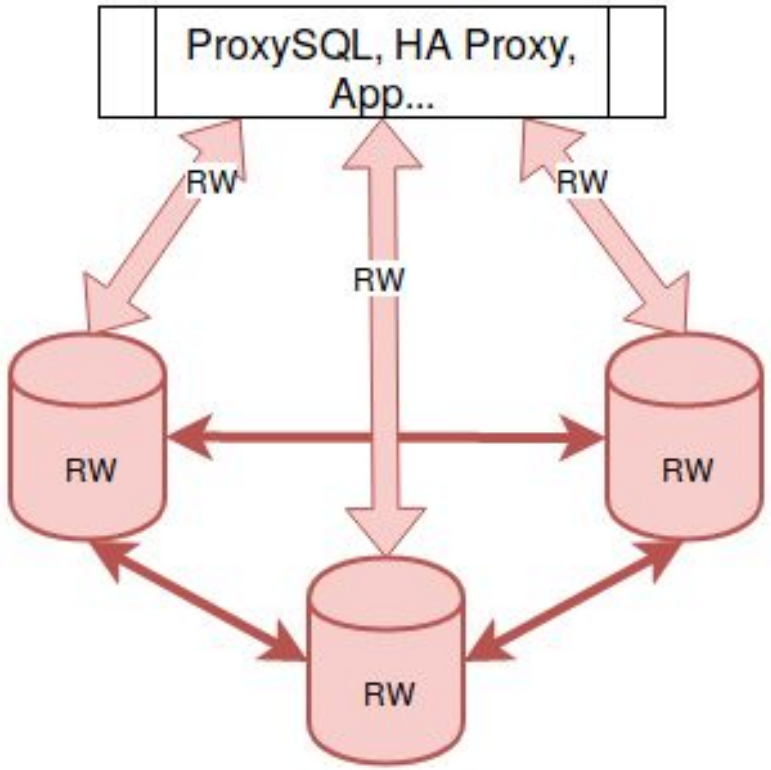
© 2019 Percona

HA Solution using PXC

Complete HA solution should have:

- **Stable cluster**
- **Load Balance**
- **Simplified and Unified view of the system**

- **Easy/Quick Failover / Switchover**
- **Easy Business Continuity Solution**



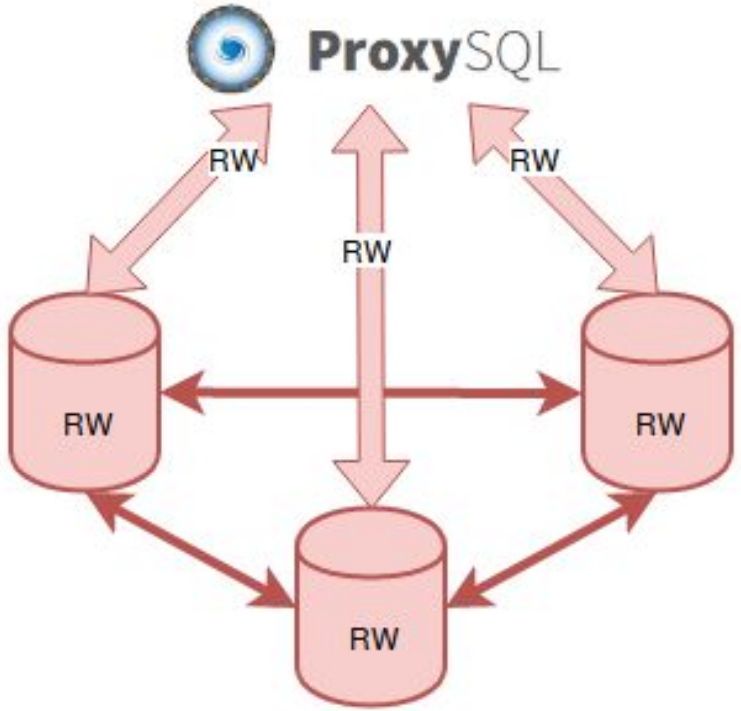
© 2019 Percona

HA - Load Balancer

- PXC can operate with multiple load/setups
- Balancers like HAProxy, ProxySQL, etc...

[PXC suggests use of ProxySQL]

  - Integrated and closed development.
  - Feature rich load balancer (lot more features getting added)
  - Custom PXC script aids simplified PXC configuration (auto-discovery of PXC nodes).



© 2019 Percona

HA - Load Balancer

PXC maintenance mode

- Abrupt graceful shutdown can cause disruption in workload till Load-Balancer adjust the load/connections.
- Maintenance mode is like advance warning helping load balancer (ProxySQL only) to make a note of this and pre-adjust workload.
- Also applicable if node needs to stop active

traffic for maintenance purpose.

- In 5.7.16 - pxc_maint_transition_period

tempo de transição dos estados (10s padrão)

pxc_maint_mode
DISABLED
SHUTDOWN
MAINTENANCE

© 2019 Percona

PMM - complete dashboard

- PMM (Percona Monitoring and Management)
- Complete integrated with PXC
- Simple way to take control of complete HA system

https://pmmdemo.percona.com/

SIMPLIFIED AND UNIT VIEW OF THE CLUSTER SYSTEM

© 2019 Percona

PERCONA

PXC as complete HA solution

Complete HA solution should have:

- **Stable cluster** ★
- **Load Balance** ★
- **Simplified and Unified view of the system** ★

- **Easy/Quick Failover / Switchover** ★
- **Easy Business Continuity Solution** ★

© 2019 Percona

Keep research / Keep Contact

PXC forum https://www.percona.com/forums/questions-discussions/percona-xtradb-cluster

PXC @ JIRA https://jira.percona.com/projects/PXC/issues

PXC Webinar 101 - Questions and Answers

PXC QUIZ

Reference: Percona XtraDB Cluster (PXC) 101 **by Krunal Bauskar**

QUESTIONS / PERGUNTAS

marcelo.goncalves@percona.com
Twitter: @billmask
LinkedIn: https://www.linkedin.com/in/billmask/

© 2019 Percona

Champions of Unbiased
Open Source Database Solutions