DTCC 2014中国数据库技术大会
DATABASE TECHNOLOGY CONFERENCE CHINA 2014
大数据技术探索和价值发现

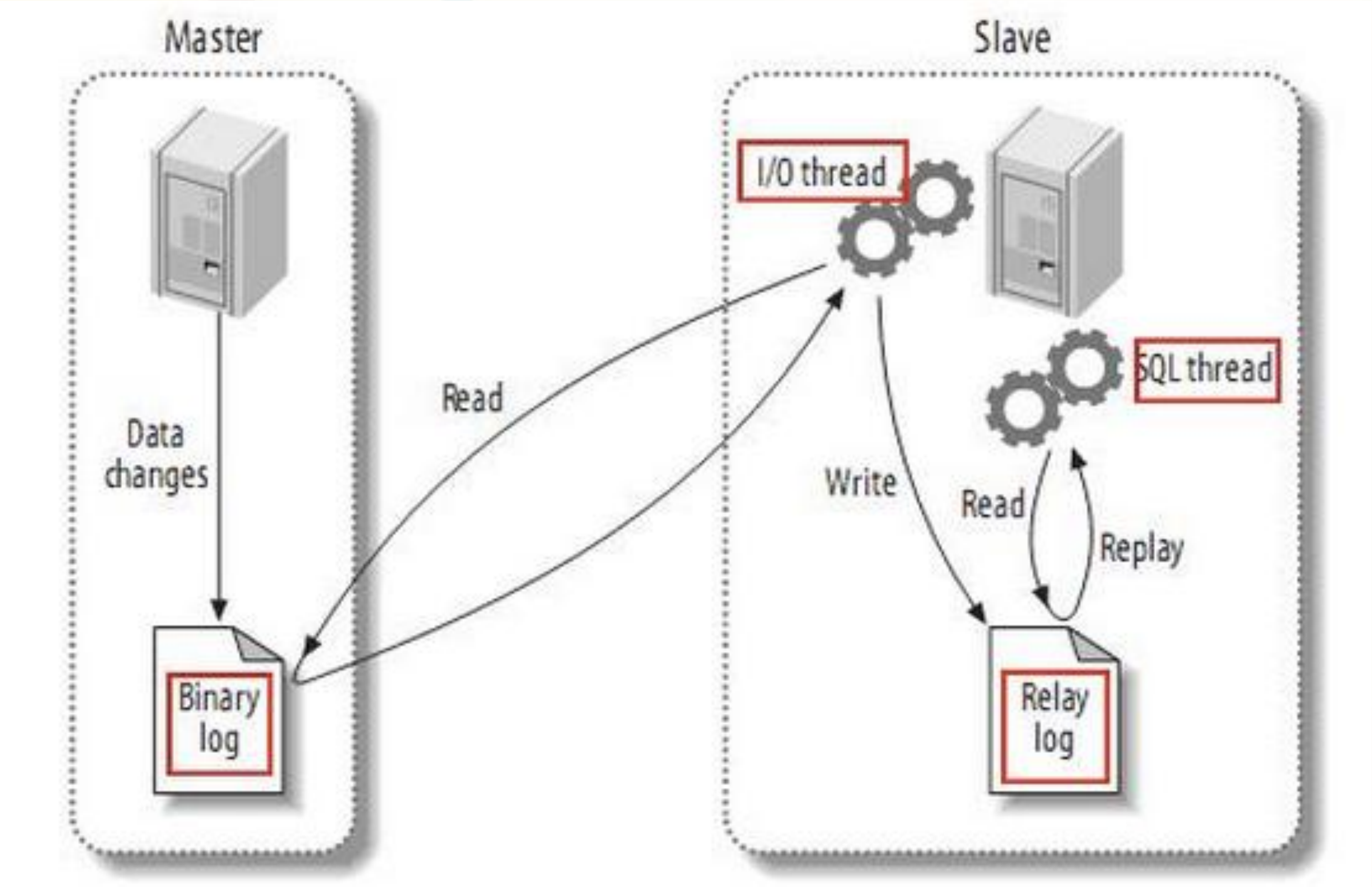
Percona XtraDB Cluster

@刘小成

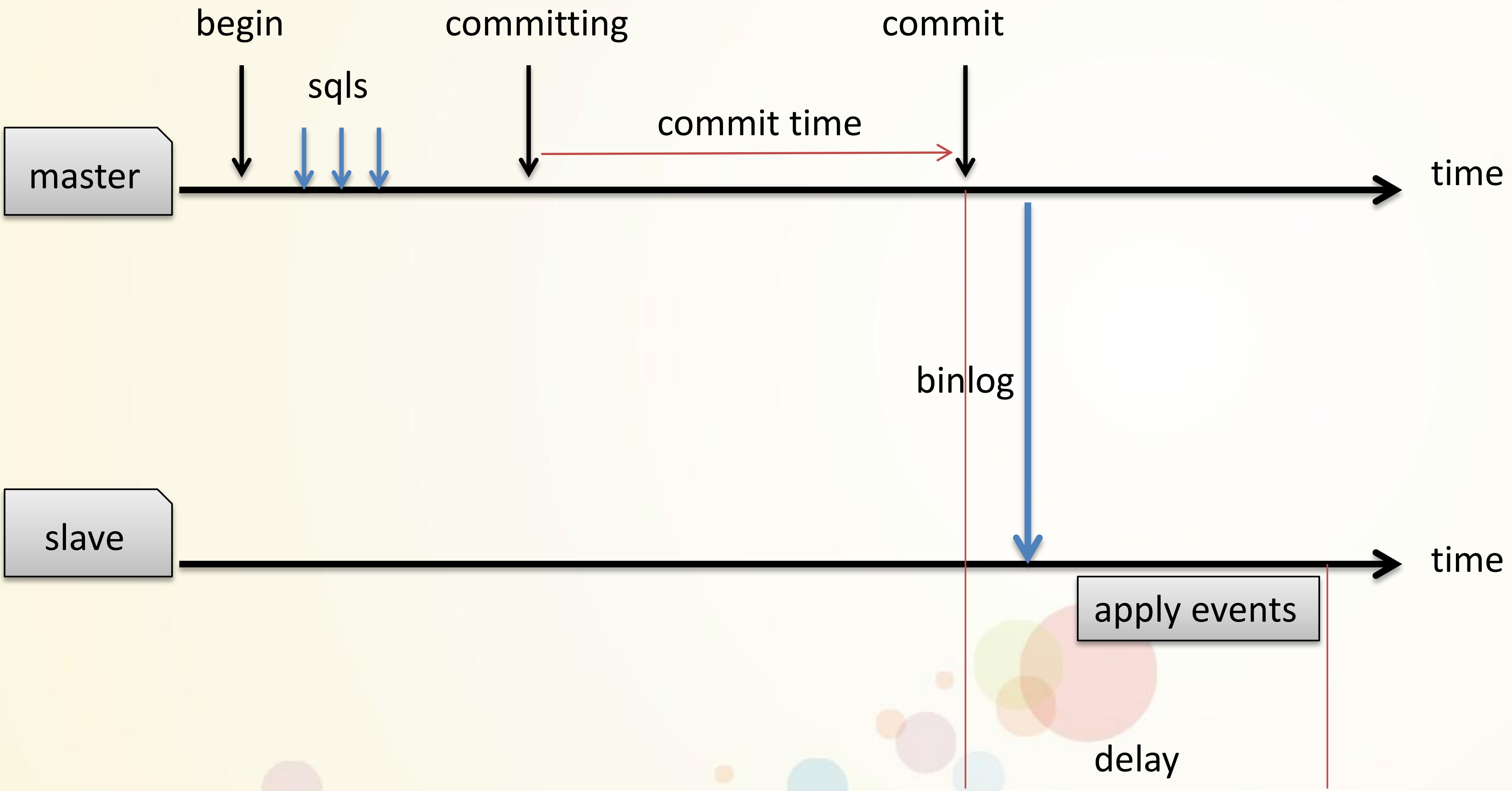
Agenda

- Replication
- Galera
- Percona XtraDB Cluster
- Refference

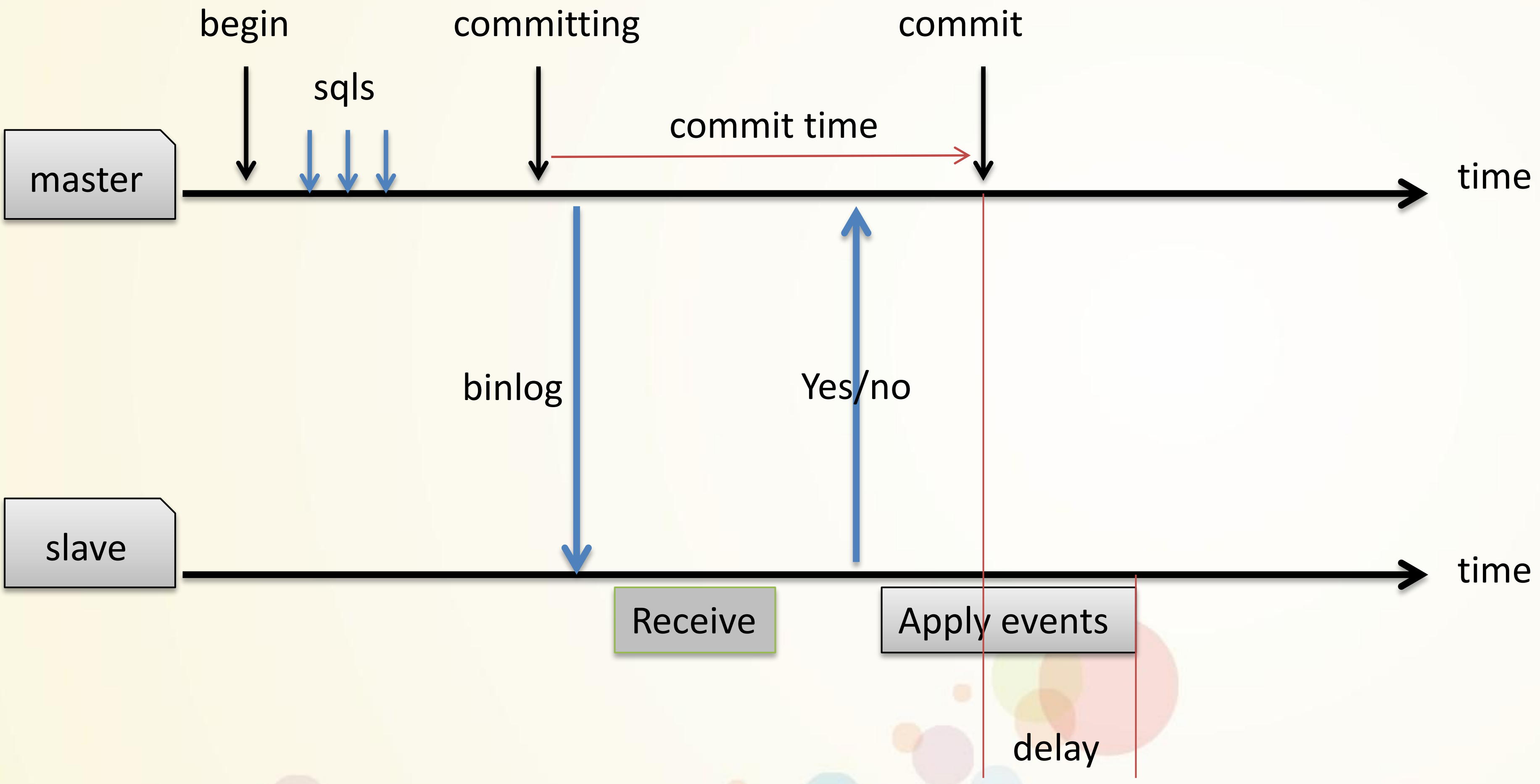
Replication



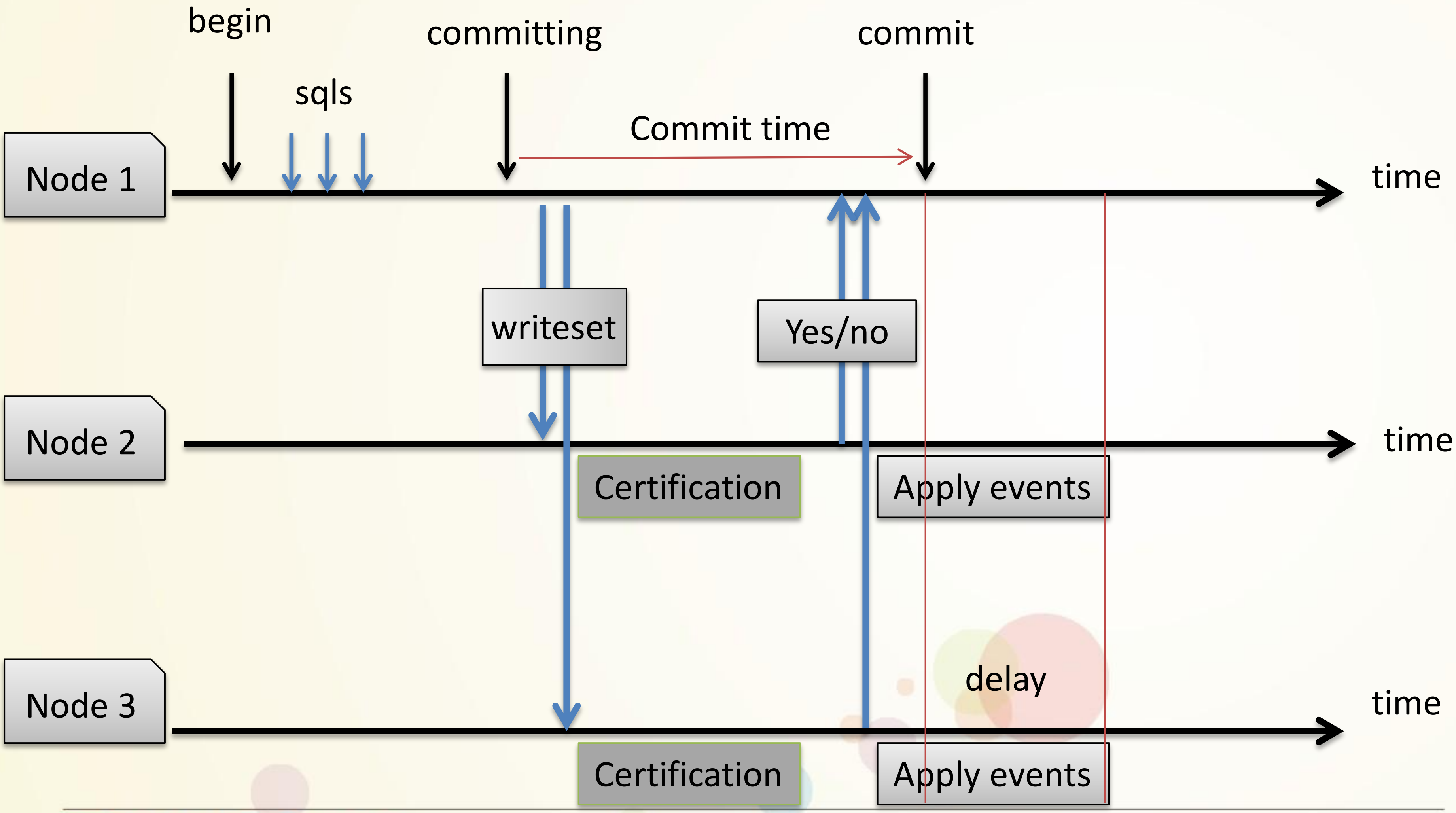
Replication-Async



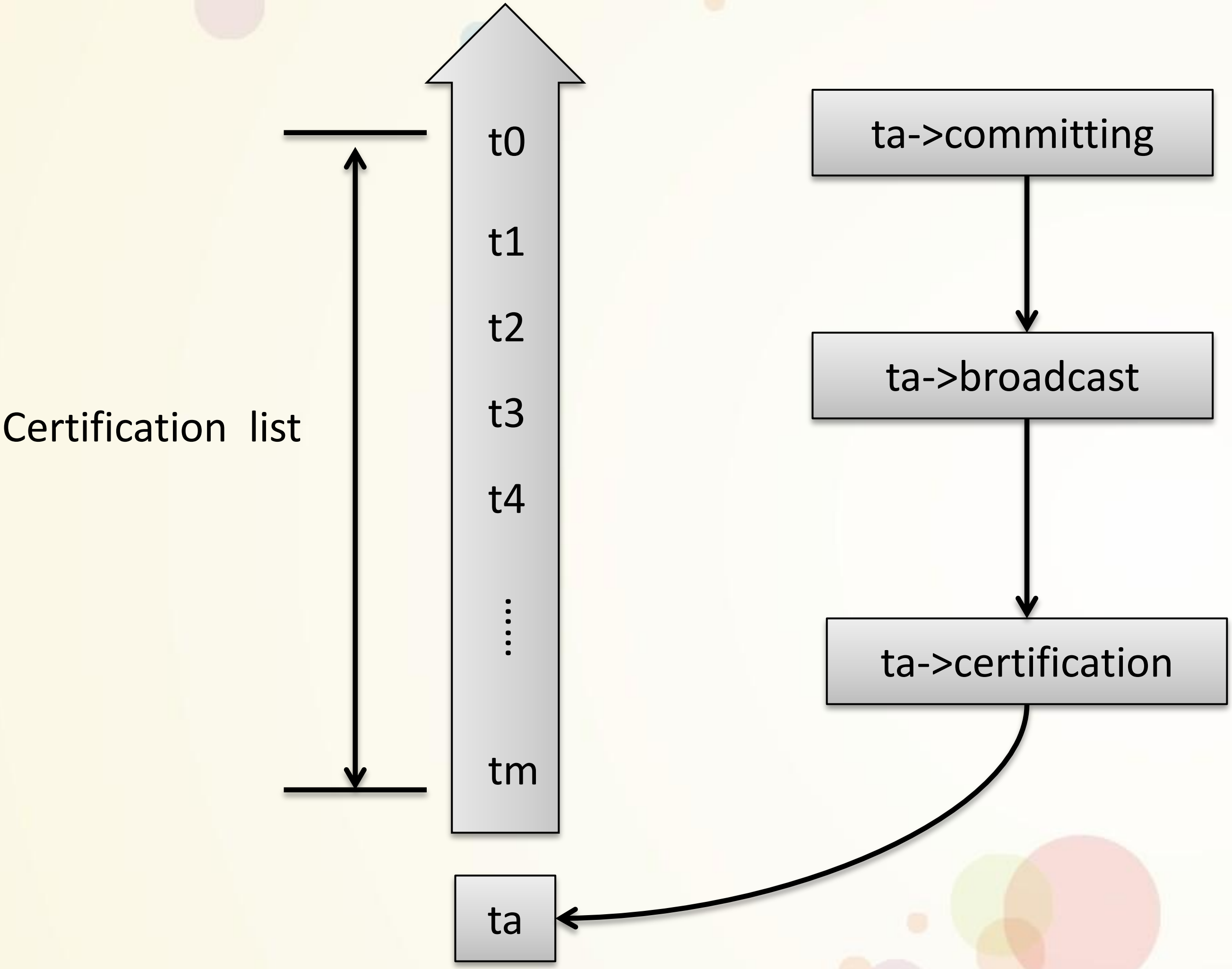
Replication-Semi



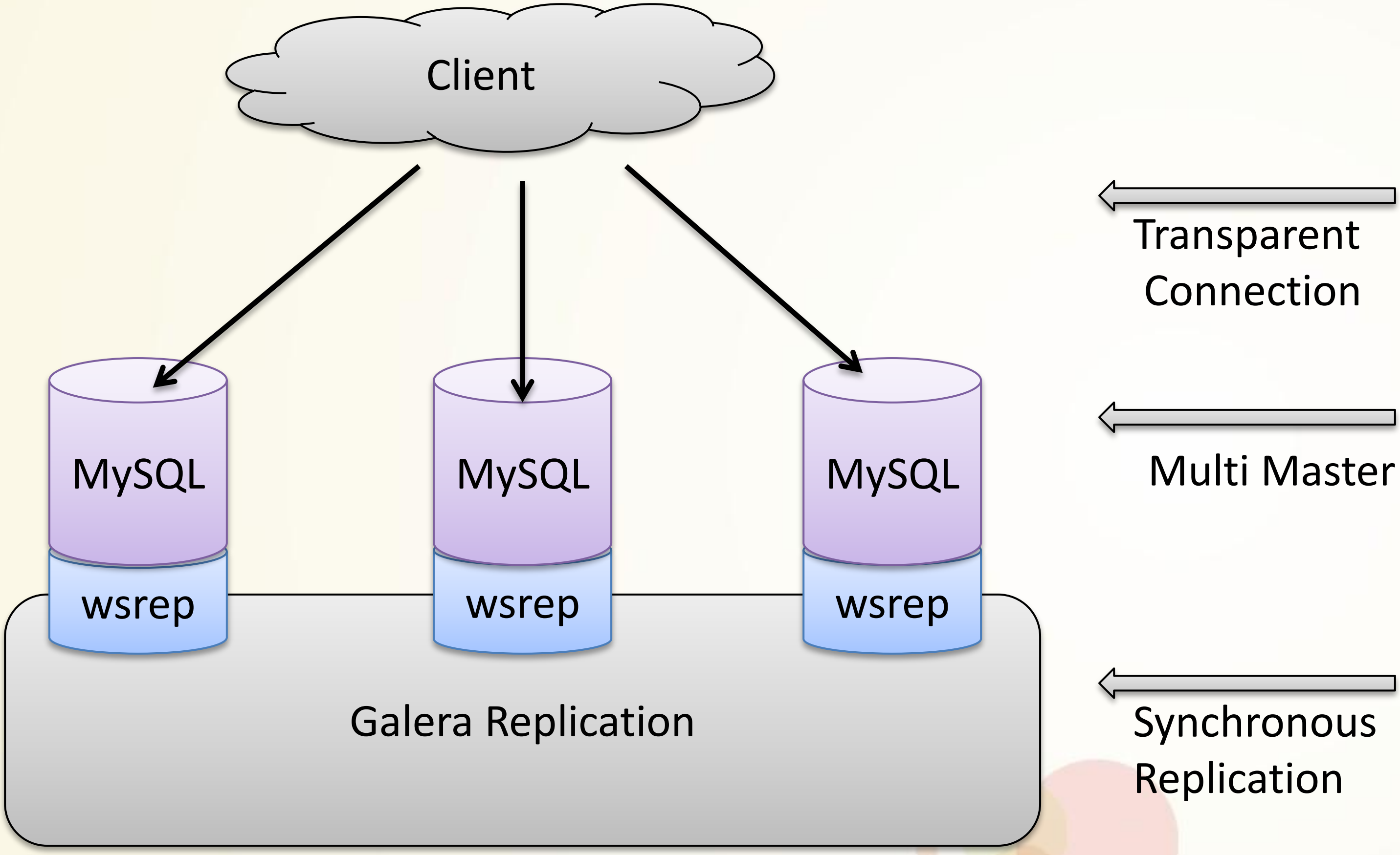
Replication-Certification



Certification



Galera Cluster



Client

MySQL

wsrep

MySQL

wsrep

MySQL

wsrep

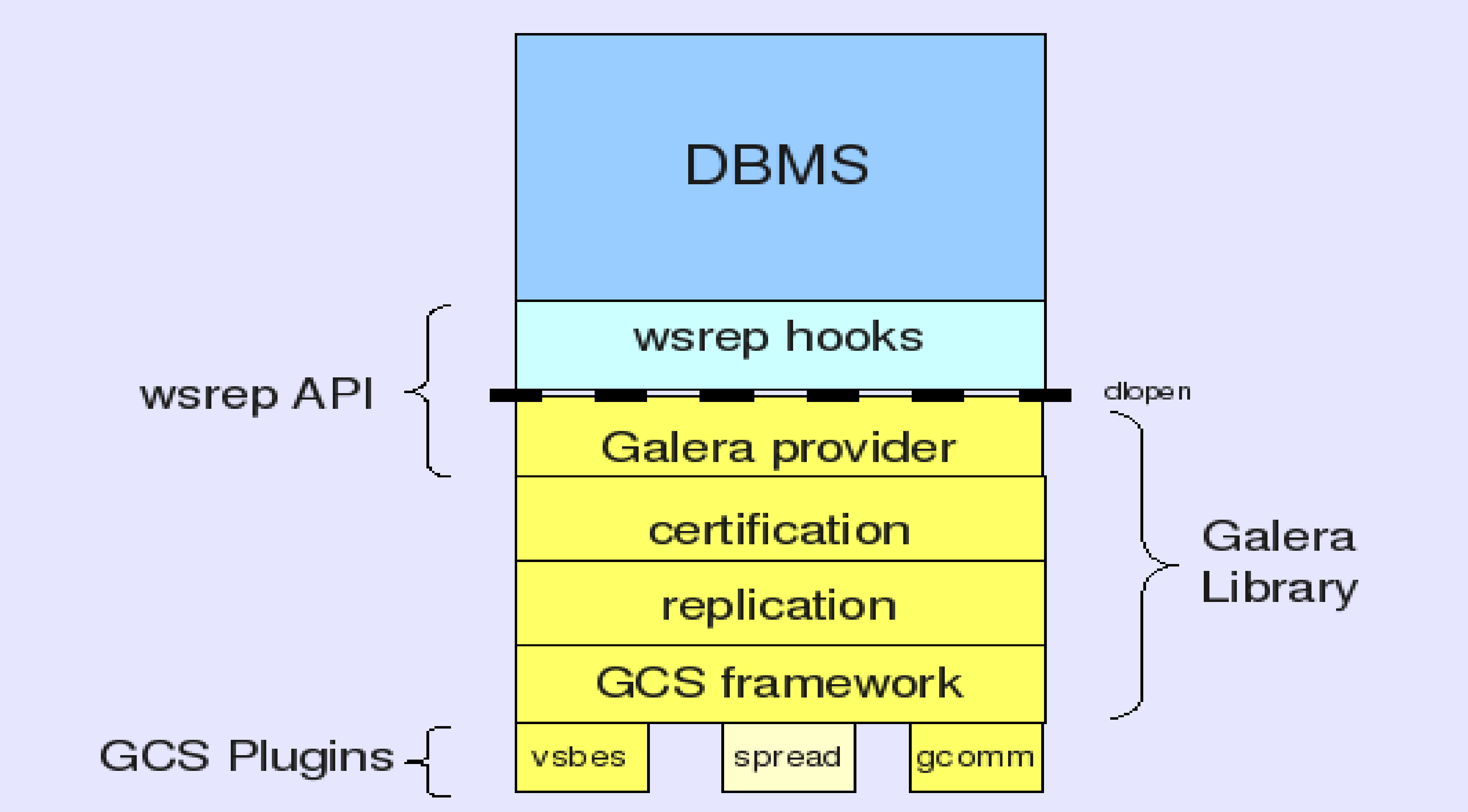
Galera Replication

Transparent Connection

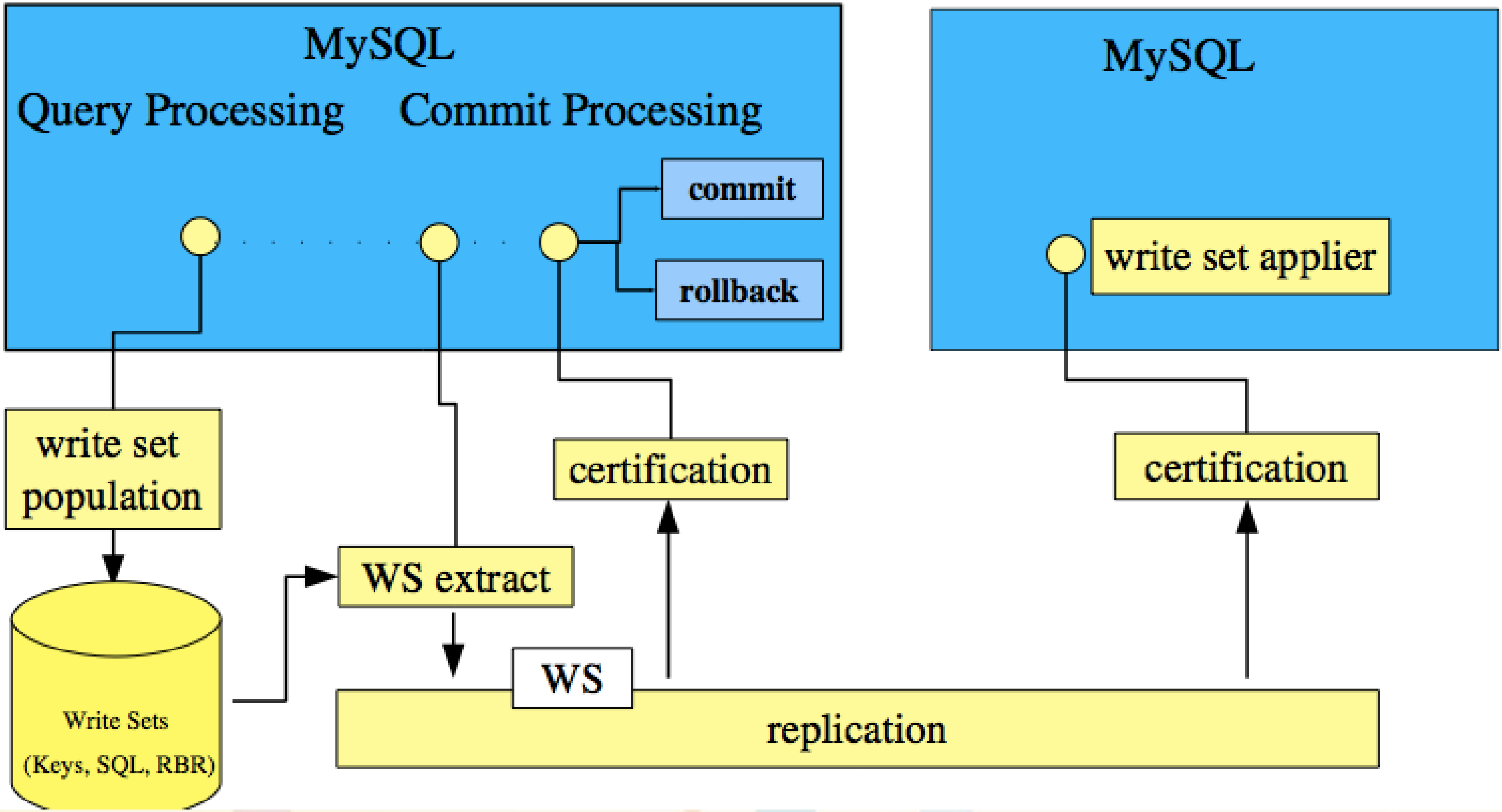
Multi Master

Synchronous Replication

Galera


DBMS
wsrep hooks
wsrep API
dlopen
Galera provider
certification
replication
GCS framework
Galera Library
GCS Plugins
vsbes
spread
gcomm

Galera Cluster



codership

PERCONA XTRADB CLUSTER

MariaDB Galera Cluster

Galera Cluster for MySQL

- High-Availability
- Multi-Master replication
- Parallel Replication
- Automatic node provisioning

Percona XtraDB Cluster

- UUID
  - 01aea6cd-1496-11e3-92d9-ae81f48f9fe4
- GTID
  - 01aea6cd-1496-11e3-92d9-ae81f48f9fe4:seqno(*64-bit signed integer)*
- PC:Primary Component

Percona XtraDB Cluster

- Parallel Applying
- Flow Control
- Split-Brain
- Wsrep causal reads

WriteSet

- WriteSet
  - key， sql， RBR
- Gcache:

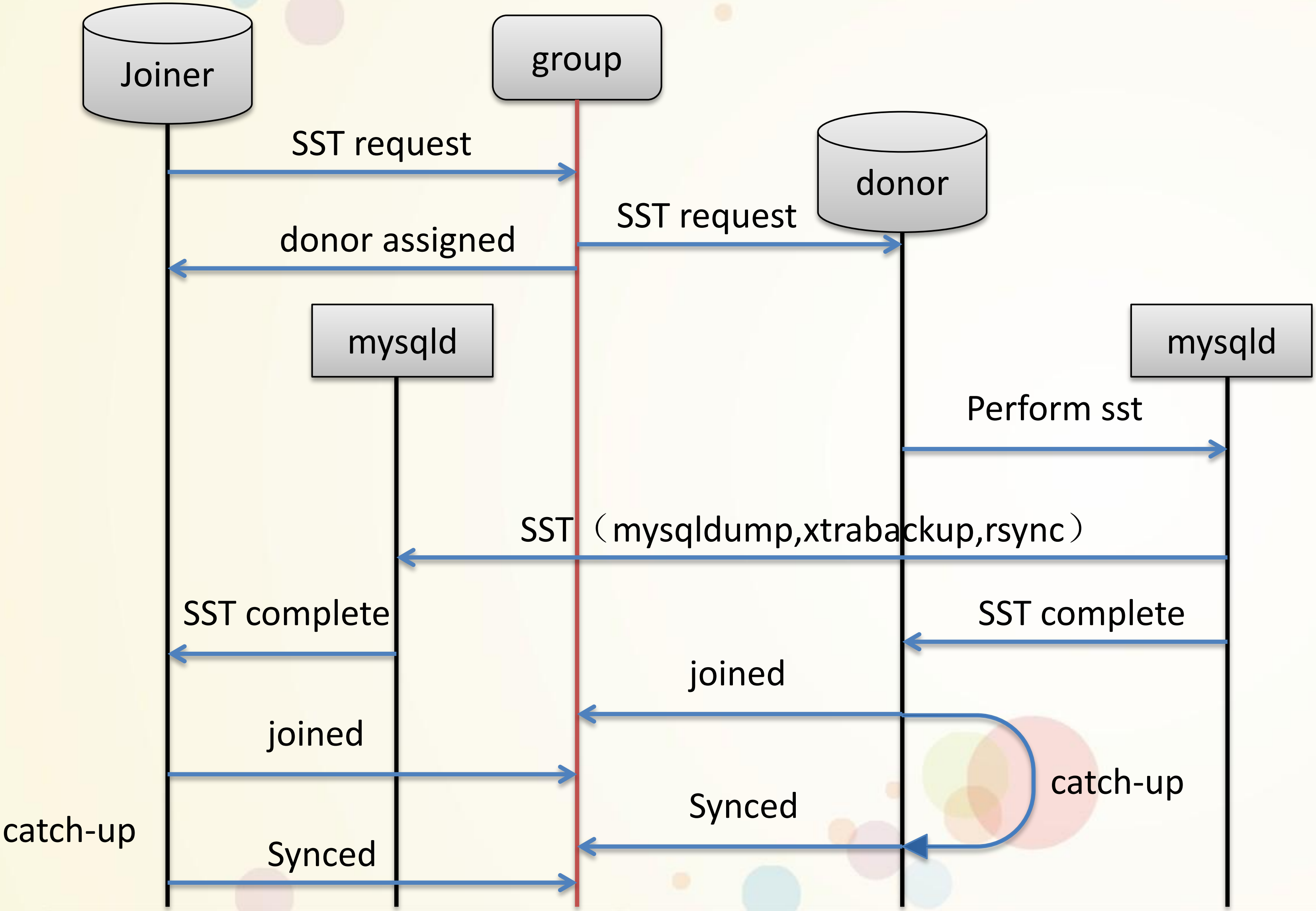
  Memory

  Disk: (128M default)

  ring buffer

  on-demand

IST & SST


Joiner
group
donor
SST request
donor assigned
SST request
mysqld
mysqld
Perform sst
SST（mysqldump,xtrabackup,rsync）
SST complete
SST complete
joined
joined
catch-up
Synced
catch-up
Synced

DDL

• TOI	Total Order Isolation
• RSU	Rolling Schema Upgrade

Configuration

- query_cache_size=0
- binlog_format=ROW
- default_storage_engine=innodb
- innodb_autoinc_lock_mode=2
- innodb_doublewrite=1

Configuration

```
wsrep_cluster_address=gcomm://10.10.58.168:4030,10.10.58.232:4030
wsrep_sst_receive_address=10.10.58.209:4020
wsrep_cluster_name=PXCS_10-10-57-2
wsrep_provider_options="ist.recv_addr = tcp://10.10.58.209:4031;…"
wsrep_node_name=PXCN_10-10-58-209
wsrep_sst_method=xtrabackup
wsrep_sst_auth=user:pwd
…
```

Monitor

wsrep_cluster_status	Primary
wsrep_cluster_size	3
wsrep_cluster_state_uuid	e2c9a15e-5485-11e0-0800-6bbb637e7211
wsrep_local_state_comment	Synced
wsrep_local_state	4
wsrep_ready	ON
wsrep_local_state_uuid	e2c9a15e-5485-11e0-0800-6bbb637e7211
wsrep_incoming_addresse	10.10.58.168:3306,10.10.58.209:3306, 10.10.58.232:3306

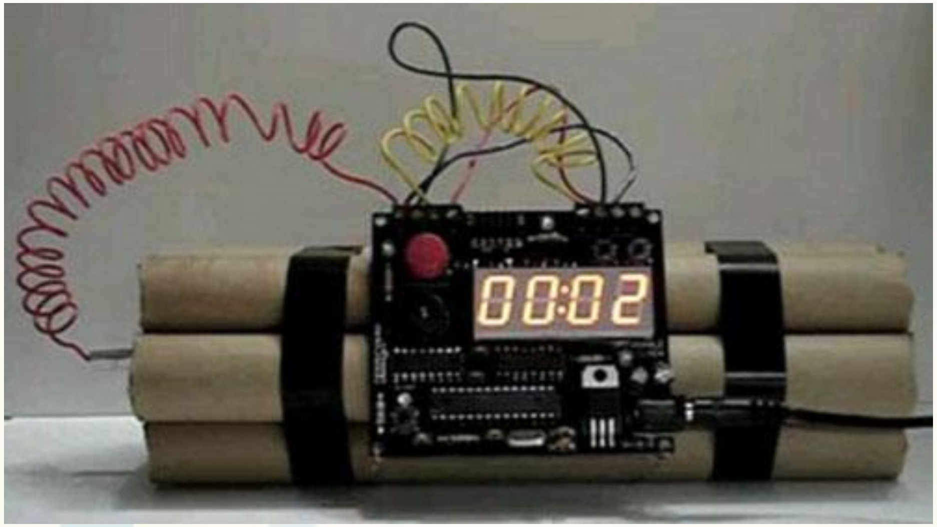
Monitor

wsrep_replicated	16109
wsrep_received	17831
wsrep_local_cert_failures	333
wsrep_local_bf_aborts	960
wsrep_local_send_queue_avg	0.145
wsrep_local_recv_queue_avg	3.348452
wsrep_flow_control_paused	0.184353
wsrep_cert_deps_distance	23.88889
wsrep_commit_window	0

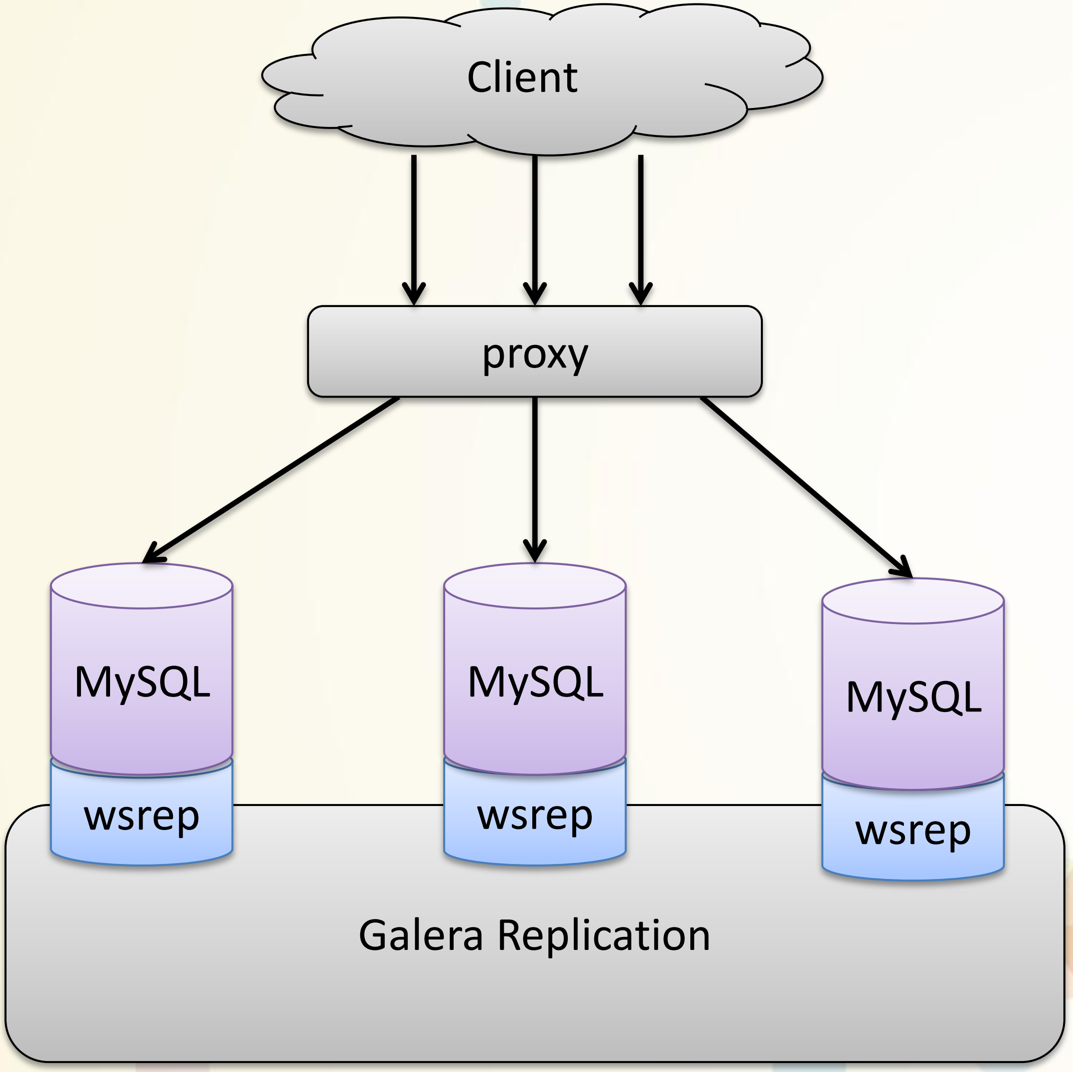
Tips

- Primary key
- InnoDB tables
- No MyISAM tables
- No forien key
- Commit exception
- Small transaction
- innodb_flush_log_at_trx_commit=2
- wsrep_slave_threads=32（4*core）

- MyISAM
- Update mysql.user set …/insert into host…
- LOCK/UNLOCK TABLES
- SET wsrep_on=0;
- SET sql_log_bin=0;

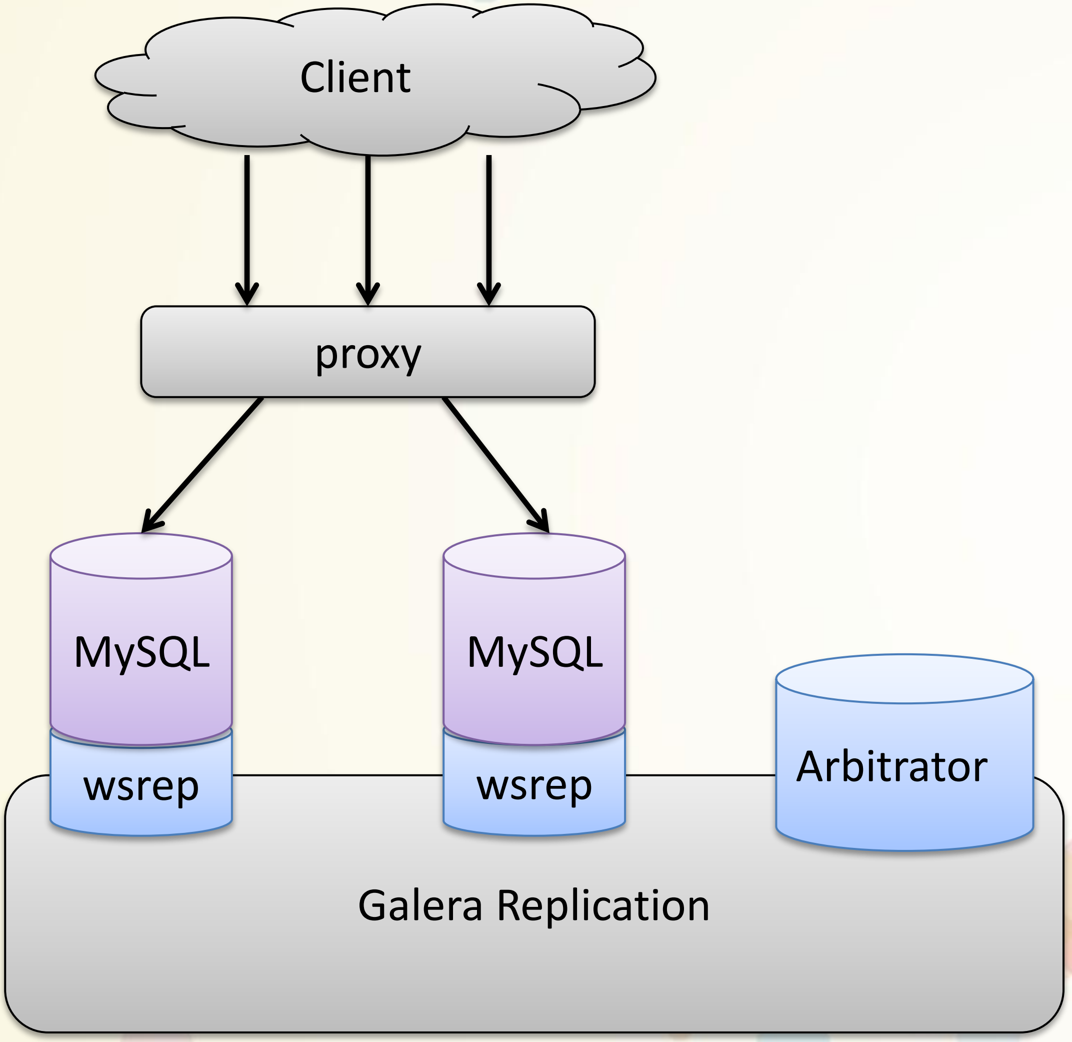


Galera Cluster-Architecture



Three or more

Galera Cluster-Arbitrator



Refference

- CAP：http://en.wikipedia.org/wiki/CAP_theorem
- The database state machine and group communication issues

http://infoscience.epfl.ch/record/32566/files/EPFL_TH2090.pdf

- Integrity Dangers in Certification-Based Replication Protocols：

http://web.iti.upv.es/~fmunyoz/research/pdf/TR-ITI-ITE-0813.pdf

- Comparison of Database Replication Techniques Based on Total Order Broadcast
- Codership : www.codership.com
- Percona: www.percona.com
- Percona Xtradb Cluster的设计与实现

http://www.cnblogs.com/bamboos/p/3543309.html

Q&A

THANKS