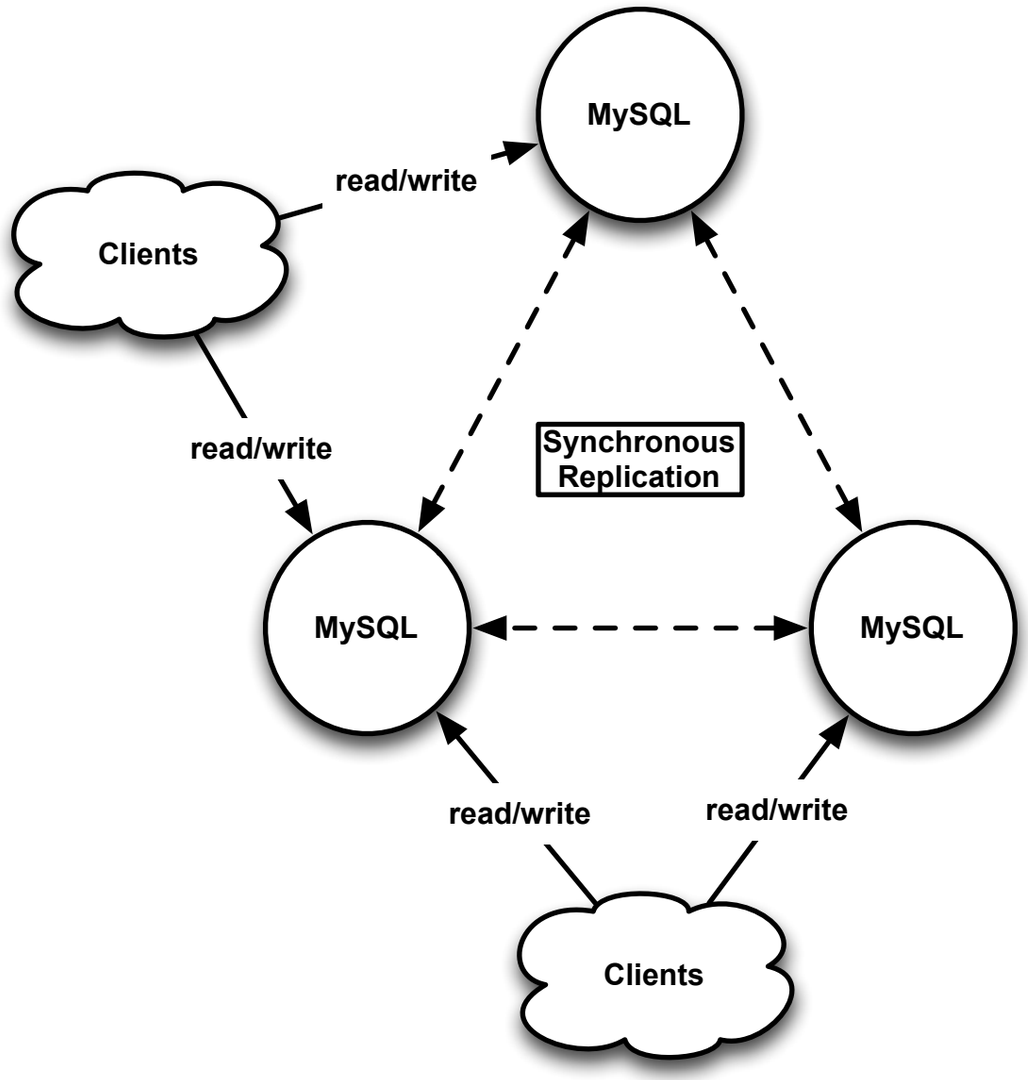
Migrating to XtraDB Cluster

Jay Janssen
MySQL Consulting Lead
March 22nd, 2013

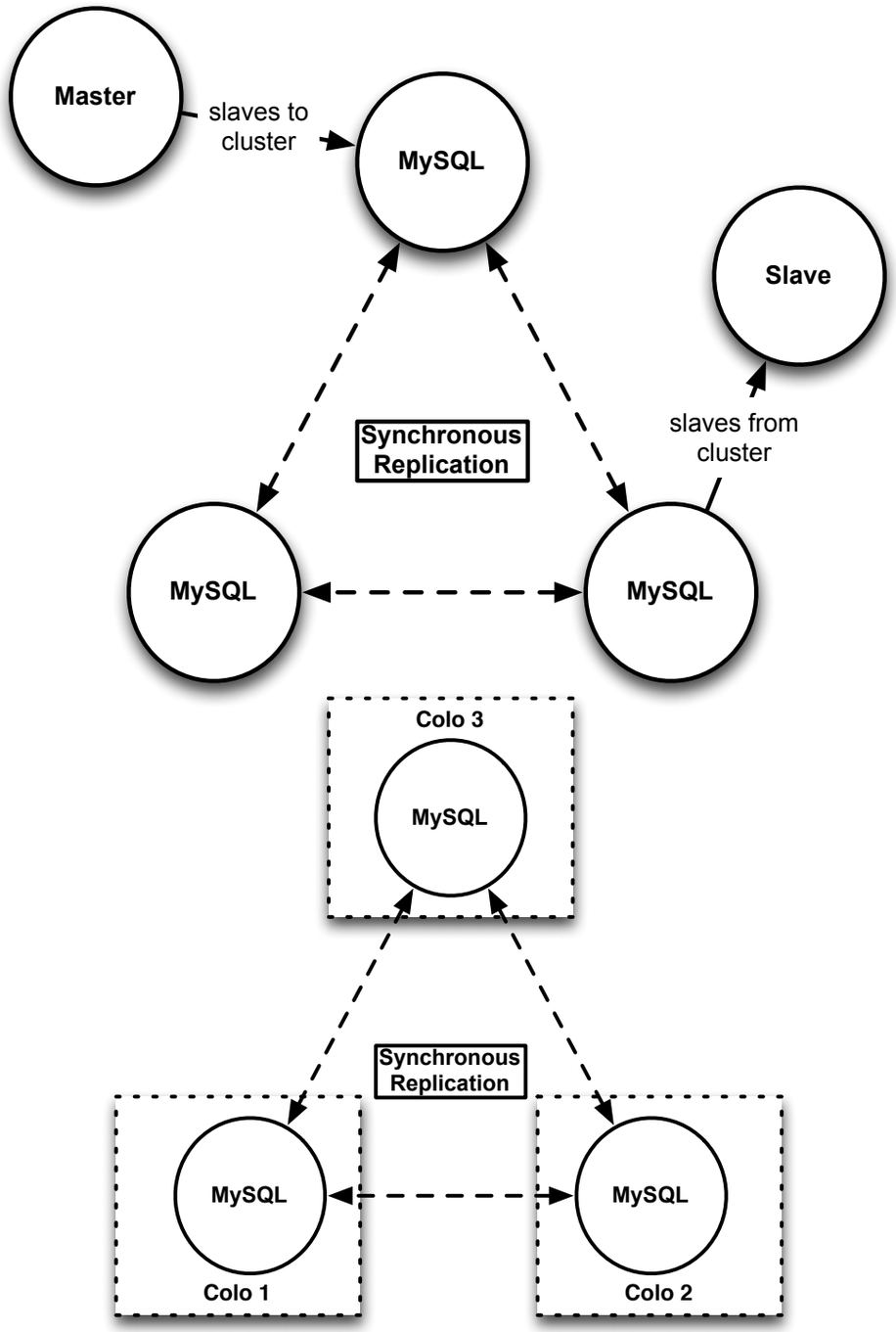
Overview of Xtradb Cluster

- Percona Server 5.5 + Galera
- Cluster of InnoDB nodes
  - Have all the data, all* the time
  - Readable and writeable
  - “Virtually” synchronous
- Established cluster:
  - Synchronizes new nodes
  - Handles node failures/resync
  - Split brain protection (quorum)



XtraDB Cluster FAQ

- Standard MySQL replication
  - into or out of the cluster
- Write scalable to a point
  - all writes still hit all nodes
- LAN/WAN architectures
  - write latency ~1 RTT for 2 DCs
- MyISAM experimental
  - big list of caveats
  - designed and built for Innodb



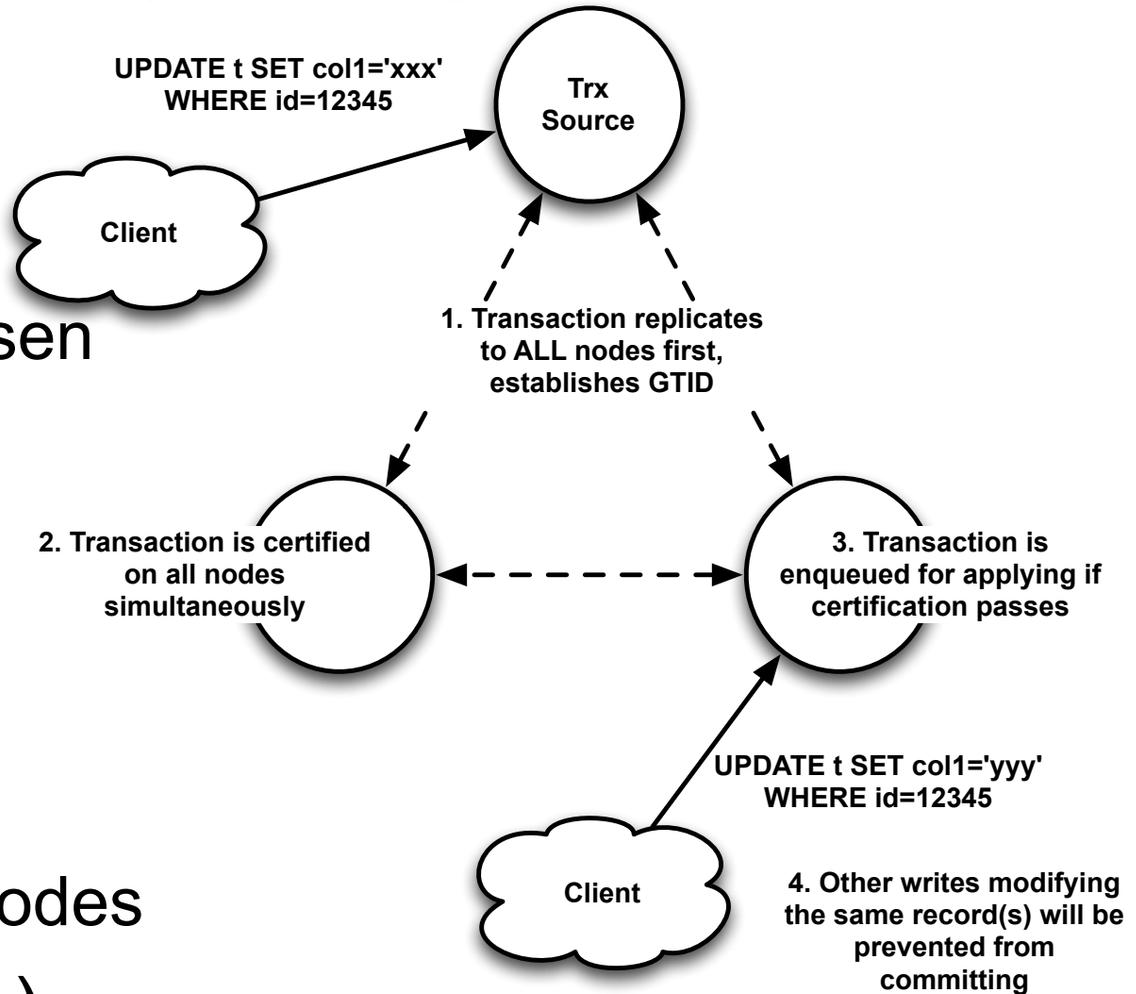
What you really want to know

- Can I drop PXC in place of my existing MySQL?
  - Yes and no (look for the tradeoffs)
  - Multi-writing will modify application behavior
  - Limitations: http://www.codership.com/wiki/doku.php?id=limitations
  - TEST, TEST, TEST
- Is it production worthy?
  - Lots of production users of Galera/PXC
  - Areas where increased prod usage has exposed bugs (e.g., FKs)
  - Test your workload to see if it's a good fit!

Application Workloads

What is Virtually Synchronous?

- Source node - pessimistic (Innodb) locking
- Cluster repl - optimistic locking
  - Before source returns commit:
    - replicates to all nodes, GTID chosen
    - source certifies
      - PASS: source applies
      - FAIL: source deadlock error (LCF)
- Other nodes
  - certify, apply or drop
  - Certification deterministic on all nodes
- Apply can abort open trxs (BFA)
  - First commit wins!



Why does the Application care?

- Latency Penalty on commit
- A given row can't be modified more than once per RTT (Callaghan's law)
  - Same node = lock wait
  - Other node = local cert failure or brute-force abort
- Increase of deadlock errors for multi-node writing!
  - Avoided by writing [some|all] data on one node
- Workload dependent!

Workloads that work best

- Multi-node writing
  - Low Data hotspots
  - Auto-increment-offset/increment is ok
- Small Transactions
  - Large trx expose serial points in repl and certification
- Tables
  - Innodb
  - With PKs
  - Avoid FKs -- supported, but problematic
    - remove this line in Q3 if no new FK bugs

Application to Cluster Connects

- For writes:
  - Best practice: (any) single node
- For Reads:
  - All/most nodes load-balanced to your choosing
    - Replication lag still possible, but minimal.
      - SET [SESSION|GLOBAL] wsrep_causal_reads = ON;
- Be sure that nodes are functioning members of the cluster!

Operational Considerations

Galera Replication

- Cluster replication (gcomm) port: tcp/4567
  - Supports multicast, SSL. Unicast by default.
  - Can be a separate network from mysqld (tcp/3306)
- Starting node needs to know one cluster node ip
  - you can list all the nodes you know and it will find one that is a member of the cluster
  - first node bootstraps cluster
- MySQL level tuning: wsrep* system variables
- Galera tuning: wsrep_provider_options

Other Cluster communication

- State Snapshot Transfer (SST)
  - Donor picked from running cluster, gives full backup to joiner node
  - Might be blocking (various methods allowed)
  - default tcp/4444
- Incremental State Transfer (IST)
  - Brief hiatus from the cluster
  - Donor as with SST, gives missing trxs to joiner
  - default tcp/4568

Example configuration

```
1.  [mysqld]
2.  datadir                          = /var/lib/mysql
3.  binlog_format                    = ROW
4.  # Other regular Server tuning

6.  innodb_buffer_pool_size          = 128M
7.  innodb_log_file_size             = 64M
8.  # Other regular Innodb tuning

10. innodb_locks_unsafe_for_binlog = 1
11. innodb_autoinc_lock_mode         = 2

13. wsrep_cluster_name               = our_cluster
14. wsrep_cluster_address            = gcomm://192.168.70.2,192.168.70.3,192.168.70.4

16. # Only use this before the cluster is formed
17. # wsrep_cluster_address          = gcomm://

19. wsrep_node_name                  = percona1
20. wsrep_node_address               = 192.168.70.2

22. wsrep_provider                   = /usr/lib64/libgalera_smm.so
23. wsrep_provider_options           = "gcache.size=2G; gcs.fc_limit=512"

25. wsrep_sst_method                 = xtrabackup
26. wsrep_sst_auth                   = backupuser:password

28. # Other wsrep options
```

Maintenance

- Rolling restarts
- Schema changes
  - potential for blocking the whole cluster
  - Galera supports a rolling schema upgrade feature
    - Isolates DDL to individual cluster nodes
    - Won't work if replication events become incompatible
  - pt-online-schema-change
- Prefer IST over SST
  - be sure you know when IST will and won't work!

Monitoring

- SHOW GLOBAL STATUS like ‘wsrep%’;
- Cluster integrity - same across all nodes
  - wsrep_cluster_conf_id - configuration version
  - wsrep_cluster_size - number of active nodes
  - wsrep_cluster_status - should be Primary
- Node Status
  - wsrep_ready - indicator that the node is healthy
  - wsrep_local_state_comment - status of this node
  - wsrep_flow_control_paused/sent - replication lag feedback
  - wsrep_local_send_q_avg - possible network bottleneck
- http://www.codership.com/wiki/doku.php?id=monitoring

Realtime Wsrep status

```
1.   $ ./myq_status -t 1 wsrep
2.   Wsrep    Cluster         Node              Queue   Ops      Bytes      Flow          Conflct
3.       time  name P cnf  #  name  cmt sta  Up  Dn  Up  Dn   Up    Dn pau snt dst lcf bfa
4.   14:38:13 myclu P  73  2 node2 Sync T/T   0   0   0  10    0  15K 0.0   0  25   0   0
5.   14:38:14 myclu P  73  2 node2 Sync T/T   0   0   0  17    0  26K 0.0   0  32   0   0
6.   14:38:15 myclu P  73  2 node2 Sync T/T   0   0   0  14    0  21K 0.0   0  38   0   0
7.   14:38:16 myclu P  73  2 node2 Sync T/T   0   0   0  12    0  18K 0.0   0  43   0   0
8.   14:38:17 myclu P  73  2 node2 Sync T/T   0   0   0  13    0  19K 0.0   0  48   0   0
9.   14:38:18 myclu P  74  3 node2 Sync T/T   0   0   0   9    0  11K 0.0   0   4   0   0
10.  14:38:19 myclu P  74  3 node2 Sync T/T   0   0   0   7    0  10K 0.0   0   8   0   0
11.  14:38:20 myclu P  74  3 node2 Dono T/T   0   1   0   4    0 4.2K 0.0   0   8   0   0
12.  14:38:21 myclu P  74  3 node2 Dono T/T   0   7   0   0    0    0 0.0   0   8   0   0
13.  14:38:22 myclu P  74  3 node2 Dono T/T   0  20   0   0    0    0 0.0   0   8   0   0
14.  14:38:23 myclu P  74  3 node2 Dono T/T   0  29   0   0    0    0 0.0   0   8   0   0
15.  14:38:25 myclu P  74  3 node2 Dono T/T   0  38   0   0    0    0 0.0   0   8   0   0
16.  14:38:26 myclu P  74  3 node2 Dono T/T   0  47   0   0    0    0 0.0   0   8   0   0
17.  14:38:27 myclu P  74  3 node2 Dono T/T   0  55   0   0    0    0 0.0   0   8   0   0
18.  14:38:28 myclu P  74  3 node2 Dono T/T   0  69   0   0    0    0 0.0   0   8   0   0
19.  14:38:29 myclu P  74  3 node2 Dono T/T   0  83   0   0    0    0 0.0   0   8   0   0
20.  14:38:30 myclu P  74  3 node2 Dono T/T   0  91   0   0    0    0 0.0   0   8   0   0
21.  14:38:31 myclu P  74  3 node2 Dono T/T   0 100   0   0    0    0 0.0   0   8   0   0
22.  14:38:32 myclu P  74  3 node2 Dono T/T   0 115   0   0    0    0 0.0   0   8   0   0
23.  14:38:33 myclu P  74  3 node2 Dono T/T   0 126   0   0    0    0 0.0   0   8   0   0
24.  14:38:35 myclu P  74  3 node2 Dono T/T   0  68   0  67    0  98K 0.0   0  42   0   0
25.  14:38:36 myclu P  74  3 node2 Sync T/T   0   0   0  77    0 112K 0.0   0  70   0   0
26.  14:38:37 myclu P  74  3 node2 Sync T/T   0   0   0  13    0  20K 0.0   0  76   0   0
27.  14:38:38 myclu P  74  3 node2 Sync T/T   0   0   0  13    0  19K 0.0   0  80   0   0
28.  14:38:39 myclu P  74  3 node2 Sync T/T   0   0   0  12    0  17K 0.0   0  85   0   0
29.  14:38:40 myclu P  74  3 node2 Sync T/T   0   0   0   6    0 9.3K 0.0   0  86   0   0
30.  14:38:41 myclu P  74  3 node2 Sync T/T   0   0   0  11    0  16K 0.0   0  91   0   0

32.  https://github.com/jayjanssen/myq_gadgets
```

ClusterControl
severalnines

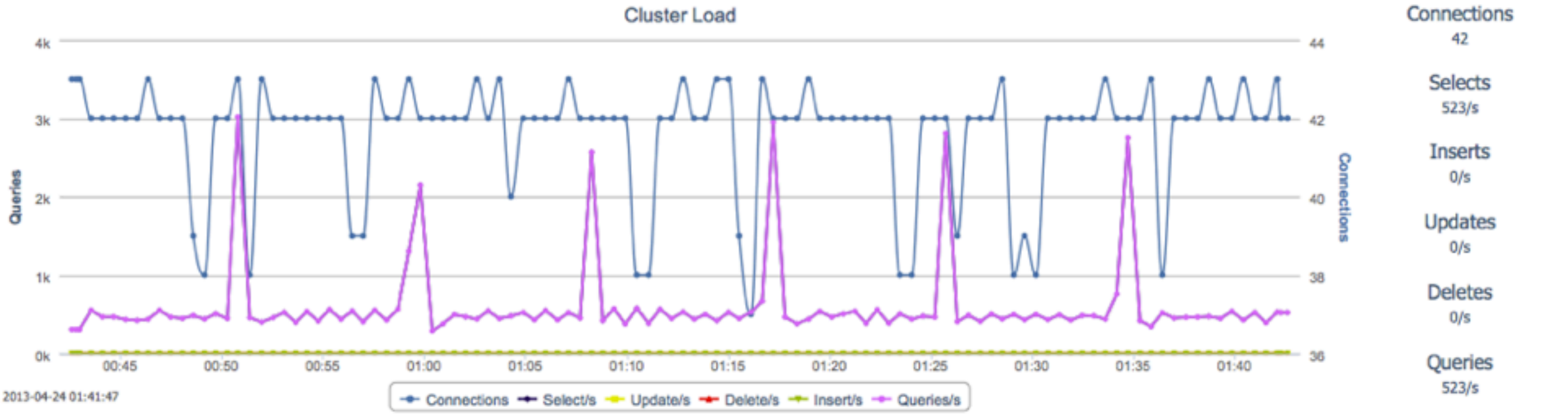
Help | Cluster Registrations | Admin | Service Providers | Log Out

Database Clusters

● GALERA_STAGING_APAC (ACTIVE)
GALERA CONNECTIONS: 42 MASTER: ✔ ✔ ✔ HAProxy: ✔

Overview | Nodes | Query Monitor | Performance 3 | Backup | Manage | Jobs/Alarms 4 | Logs | Settings

Showing Range: 1 Hour Ago



Galera Nodes

Node	State	Cluster Status	WSREP Cluster Size	WSREP Ready	Local Queue (Send/Receive)	Flow Control Paused	Flow Control Sent	Cert Ceps Distance	Last Committed	Uptime	Last Updated
10.132.172.158	Synced	Primary	3	ON	0.000000 / 0.000000	0.000000	0	0.000000	70962	4 Days 8 Hours 40 Minutes	2013-04-23 23:42:35
10.132.179.159	Synced	Primary	3	ON	0.000000 / 0.000000	0.000000	0	0.000000	70962	1 Day 14 Hours 51 Minutes	2013-04-23 23:42:35
10.132.181.147	Synced	Primary	3	ON	0.000000 / 0.000000	0.000000	0	0.000000	70962	18 Days 18 Hours 23 Minutes	2013-04-23 23:42:35

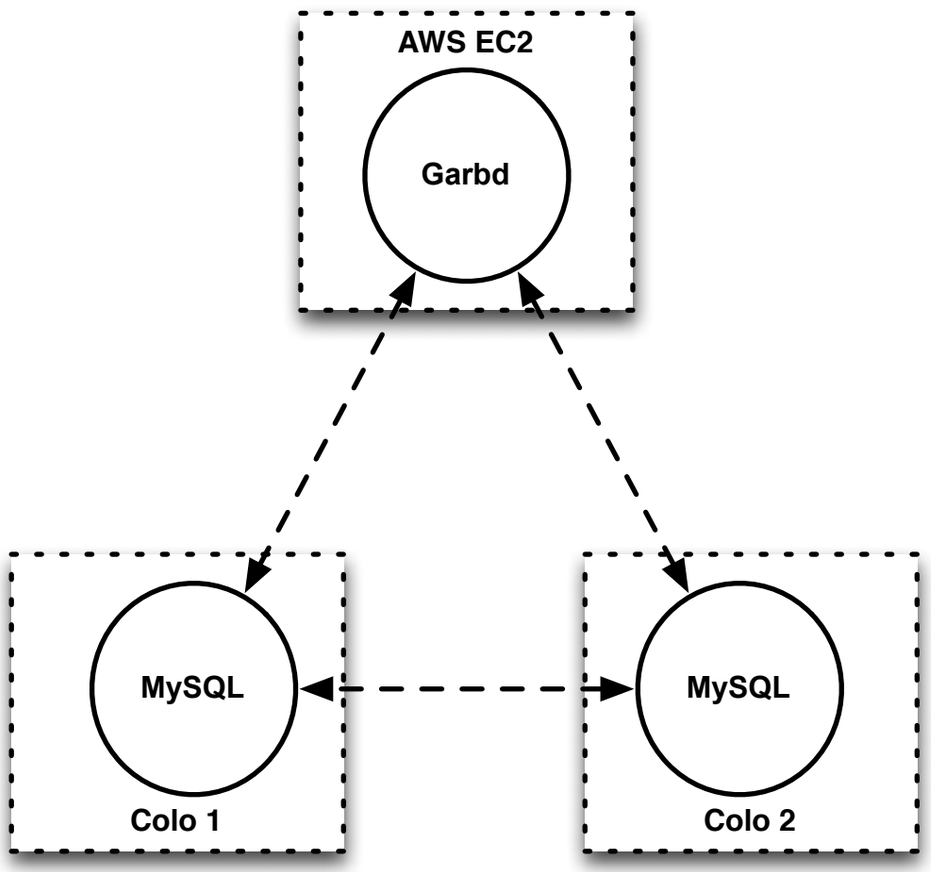
Last updated. Apr 24, 2013 01:41:48

Hosts

Host	Ping(us)	CPU Util(%)	Loadavg(1)	Loadavg(5)	Loadavg(15)	Net tx/s	Net rx/s	Disk read	Disk write	Uptime	Last Updated
10.132.172.158	344	99.79	4.15	4.35	4.44	4.68 KB	2.61 KB	0.00 B | 0/s	63.87 KB | 5/s	1 Month 13 Days 17 Hours	2013-04-23 23:42:33
10.132.179.159	64747	80.70	3.21	3.88	4.01	4.79 KB	2.77 KB	0.00 B | 0/s	29.47 KB | 3/s	1 Month 13 Days 17 Hours	2013-04-23 23:42:25
10.132.181.147	30311	96.87	4.36	4.48	4.43	5.63 KB	3.17 KB	136.00 B | 0/s	56.00 KB | 5/s	1 Month 13 Days 17 Hours	2013-04-23 23:42:17

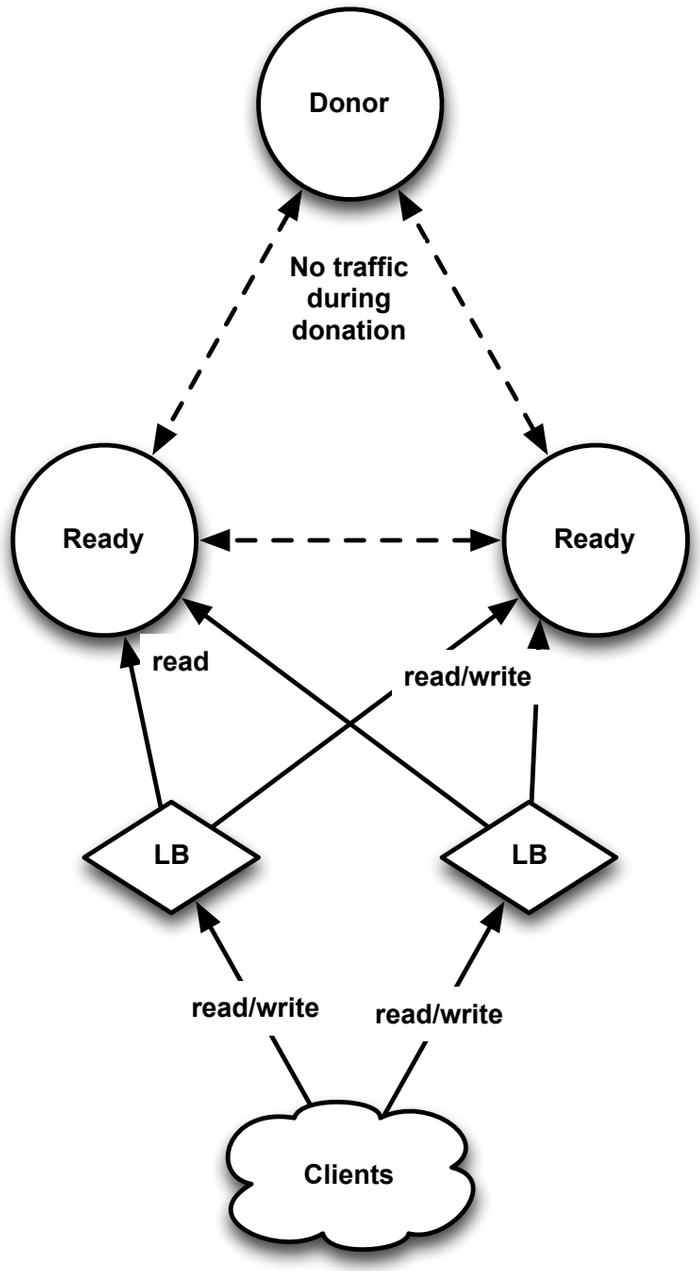
Architecture

- How many nodes should I have?
  - <= 50% is not a quorum
  - garbd - Galera Arbitrator Daemon
    - Contributes as a voting node for quorum
    - Does not store data, but does replicate
- What gear should I get?
  - Writes as fast as your slowest node
  - Standard MySQL + Innodb choices
  - garbd could be on a cloud server



Load balancing and Node status

- Health check:
  - TCP/3306
  - SHOW GLOBAL STATUS
    - wsrep_ready = ON
    - wsrep_local_state_comment !~ m/Donor/?
    - /usr/bin/clustercheck
- Maintain a separate rotations:
  - Reads
    - RR or Least Connected all available
  - Writes
    - Single node with backups on failure



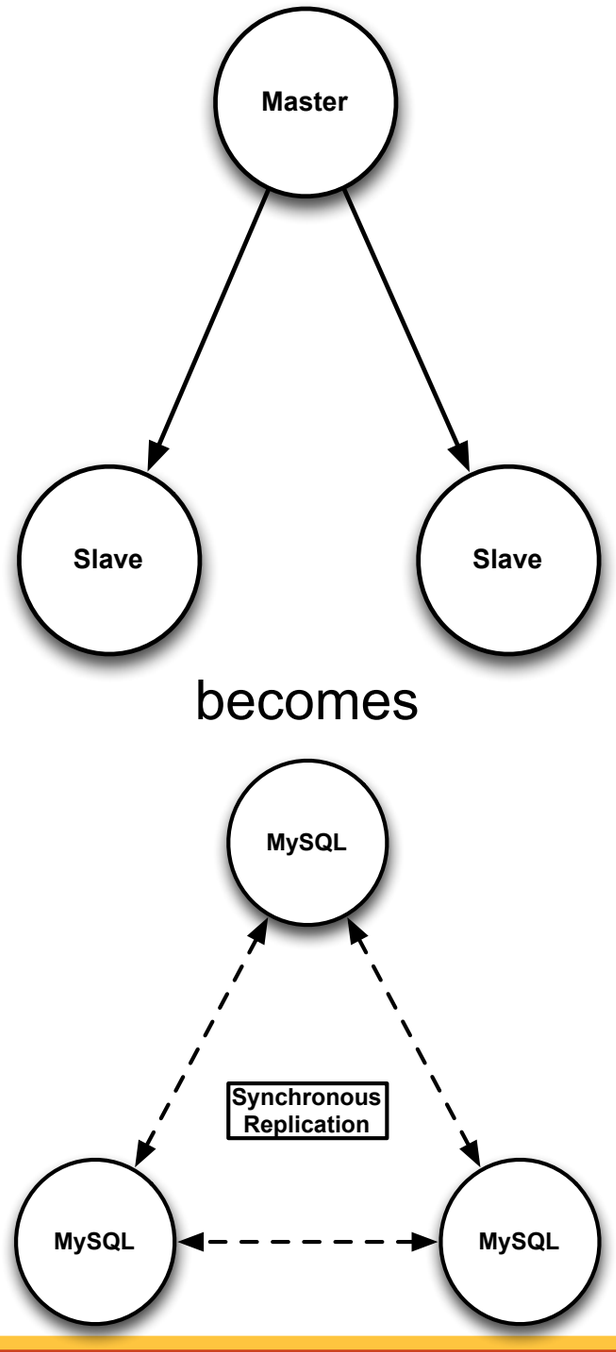
Converting Standalone MySQL to Xtradb Cluster

Method 1 - Single Node

- Migrating a single server:
  - stop MySQL
  - replace the packages
  - add Galera (wsrep_*) settings
  - start MySQL
- A stateless, peerless node will form its own cluster
  - IFF an empty cluster address is given (gcomm://)
- That node has the baseline data for the cluster
- Easiest from equivalent Percona Server 5.5

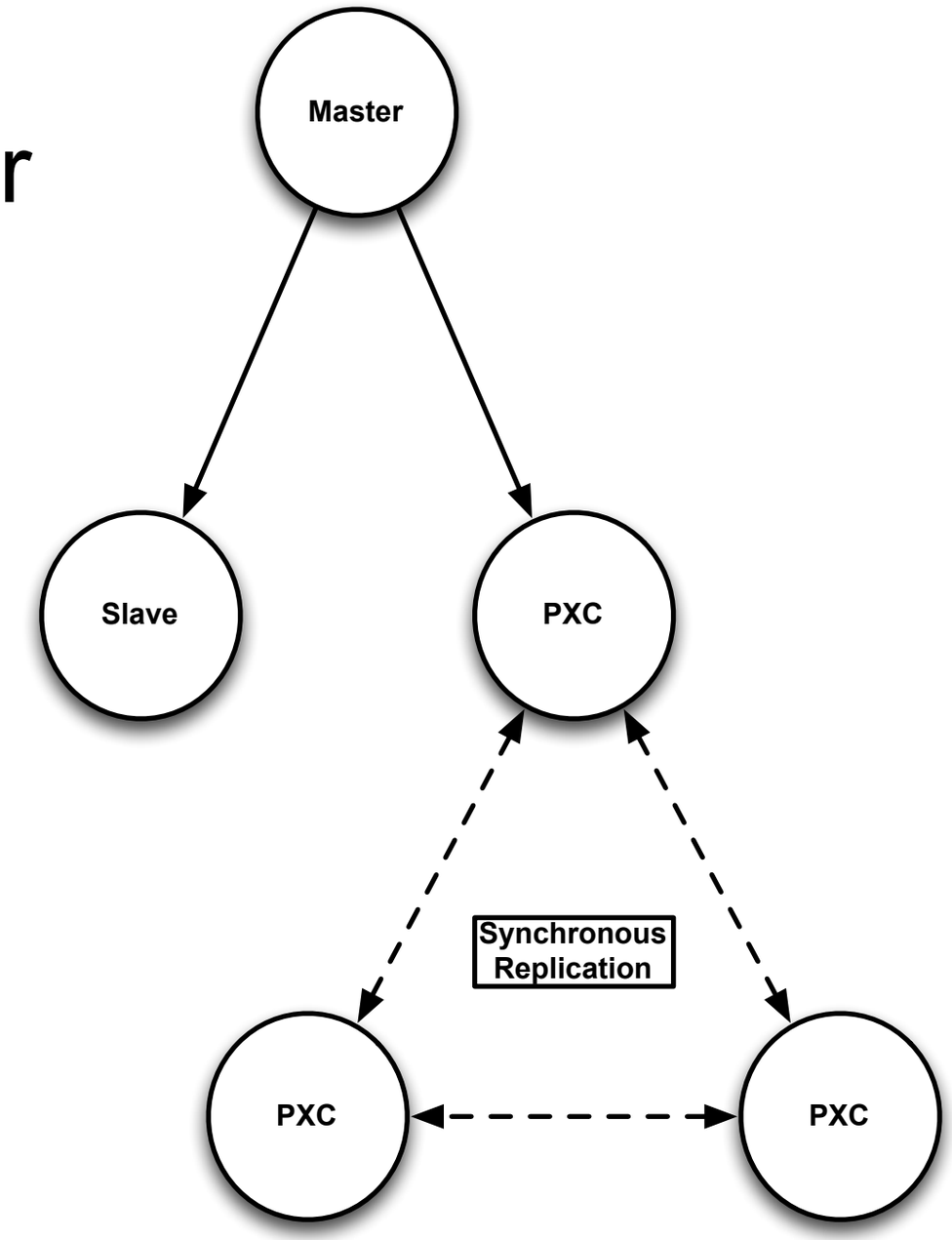
Method 2 - Blanket changeover

- All at once (with downtime):
  - Stop all writes, wait for replication
  - RESET SLAVE, stop mysql
  - Upgrade software
  - Start first node - initial cluster
  - Start the others with wsrep_sst_mode=skip
- The slaves will join the cluster, skipping SST
- Change wsrep_sst_method!

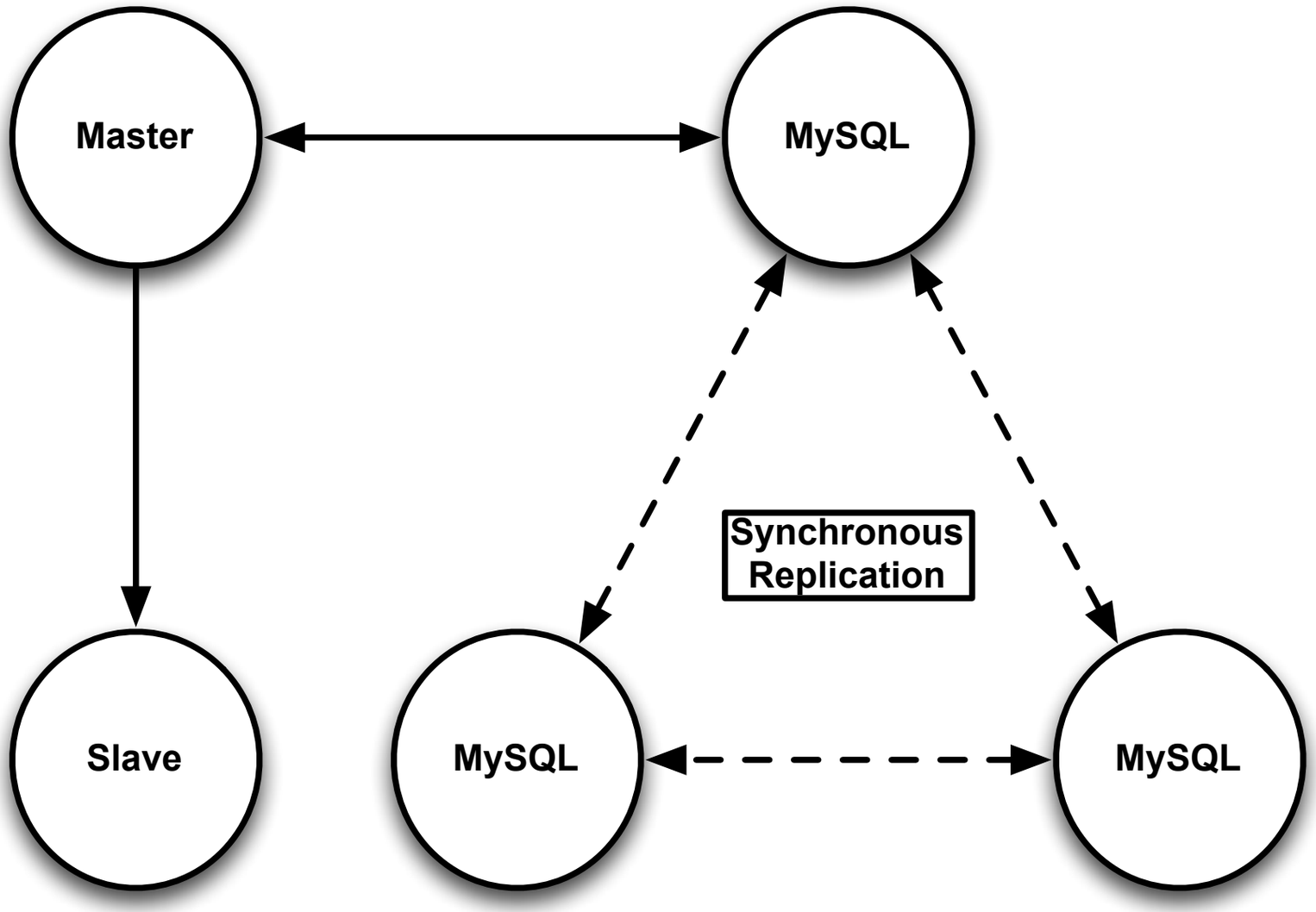


Method 3 - Slave cluster

- New cluster from one slave
- Node replicates from old master
  - log-slave-updates on this node
- Test like any other slave
- Move more slaves to cluster
  - Real SST
  - More Testing
- Cut writes over to the cluster
- Move old master into cluster



Method 3b - Dual-Master



PXC Migration Recommendations

- Test your workload carefully
  - Regression
  - Load
- Stay with Master/slave, at least at the beginning
- Avoid major MySQL version upgrades
- Dual-master your cluster with standalone master for rollback
- Be prepared for the operational learning curve
- Perform plenty of HA testing before migration

Resources

- XtraDB Cluster homepage and documentation:
  - http://www.percona.com/software/percona-xtradb-cluster/
- Galera Documentation:
  - http://www.codership.com/wiki/doku.php
- PXC tutorial (self-guided or at a conference):
  - https://github.com/percona/xtradb-cluster-tutorial
- http://www.mysqlperformanceblog.com/category/xtradb-cluster/
- Google Groups:
  - codership-team
  - percona-discussion