Percona Xtradb Cluster

Reference Architecture 2016
Slides: http://bit.ly/1qBNFmW

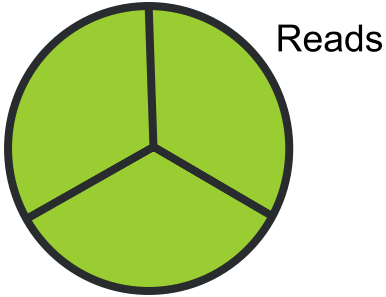
Jay Janssen
Managing Principal Architect

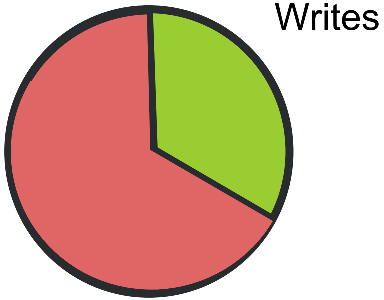
PERCONA

http://bit.ly/1qBNFmW

Cluster Sizing

- Quorum
  - Rule of 3's
  - Really: The loss of any one *thing* should not cause the cluster to lose quorum
- Read and Write Scalability
  - Reads scale like Master/Slave
  - Writes don't scale (much)
- Which nodes to read and write from
  - Reads can be most anywhere
  - Writes may best start on a single node
    - *Some App work can overcome this*





Cluster Setup and Configuration

PERCONA

Baseline Configuration

- 3 CentOS 7 Nodes
- IPs: 172.28.128.3, .4, .5
- EPEL Repo installed for 1 dependency (socat)
- SElinux --permissive
- Firewall open on 3306, 4567, 4568, 4444 / tcp

```
[root@node1 ~]# firewall-cmd --add-port=3306/tcp --add-port=4567/tcp --add-
port=4568/tcp --add-port=4444/tcp --permanent
success
[root@node1 ~]# firewall-cmd --reload
Success
```

http://bit.ly/1qBNFmW

Software Install (all 3 nodes)

- Package Install

```
# yum install http://www.percona.com/downloads/percona-release/redhat/0.1-
3/percona-release-0.1-3.noarch.rpm
```

```
# yum install Percona-XtraDB-Cluster-56.x86_64
```

- Monitoring Software (Myq Gadgets)
  - https://github.com/jayjanssen/myq-tools/releases

```
# wget `curl -s https://api.github.com/repos/jayjanssen/myq-tools/releases |
grep browser_download_url | head -n 1 | cut -d '"' -f 4` && tar xvzf myq_tools.
tgz -C /usr/local/bin --strip-components=1 && ln -sf /usr/local/bin/myq_status.
linux-amd64 /usr/local/bin/myq_status
```

http://bit.ly/1qBNFmW

Starter /etc/my.cnf

```
[mysqld]
binlog_format                  = ROW
datadir                        = /var/lib/mysql

innodb_flush_log_at_trx_commit = 0
innodb_autoinc_lock_mode       = 2

wsrep_cluster_address          = gcomm://172.28.128.3,172.28.128.4,172.28.128.5
wsrep_provider                 = /usr/lib64/galera3/libgalera_smm.so
wsrep_provider_options         = "gcache.size=128M; gcs.fc_limit=128;
                                  gcs.fc_master_slave=yes"

wsrep_cluster_name             = mycluster
wsrep_slave_threads            = 8
wsrep_sst_method               = xtrabackup-v2
wsrep_sst_auth                 = sst:secret

[mysqld_safe]
pid-file = /run/mysqld/mysql.pid
syslog
```

[root@node1 ~]# systemctl start mysql@bootstrap

Verify it worked

```
[root@node1 ~]# /usr/local/bin/myq_status wsrep
```

```
mycluster /  (idx: 0) / Galera 3.14(r53b88eb)
         Cluster  Node       Outbound      Inbound       FlowC     Conflct Gcache     Appl
    time P cnf  # stat laten msgs data que msgs data que pause snt lcf bfa   ist  idx  %ef
17:56:06 P   1  1 Sync 0.0ns    1 290b   0    2 133b   0   0ns   0   0   0     0    1  12%
17:56:07 P   1  1 Sync 0.0ns    0   0b   0    0   0b   0   0ns   0   0   0     0    1  12%
17:56:08 P   1  1 Sync 0.0ns    0   0b   0    0   0b   0   0ns   0   0   0     0    1  12%
```

Grant SST User

```
mysql> GRANT LOCK TABLES, RELOAD, REPLICATION CLIENT ON *.* TO
'sst'@'localhost' IDENTIFIED BY 'secret';
```

More Verification

```
mysql> SHOW GLOBAL STATUS like 'wsrep%';

...
| wsrep_evs_state              | OPERATIONAL                                  |
| wsrep_gcomm_uuid             | 019add5c-f769-11e5-89e8-72592f726546         |
| wsrep_cluster_conf_id        | 1                                            |
| wsrep_cluster_size           | 1                                            |
| wsrep_cluster_state_uuid     | 019d9782-f769-11e5-9292-72a6b97aa153         |
| wsrep_cluster_status         | Primary                                      |
| wsrep_connected              | ON                                           |
| wsrep_local_bf_aborts        | 0                                            |
| wsrep_local_index            | 0                                            |
| wsrep_provider_name          | Galera                                       |
| wsrep_provider_vendor        | Codership Oy <info@codership.com>            |
| wsrep_provider_version       | 3.14(r53b88eb)                               |
| wsrep_ready                  | ON                                           |
+------------------------------+----------------------------------------------+
58 rows in set (0.00 sec)
```

http://bit.ly/1qBNFmW

[root@node2 ~]# systemctl start mysql
[root@node3 ~]# systemctl start mysql

Node1’s myq_status

```
mycluster /   (idx: 0) / Galera 3.14(r53b88eb)
         Cluster  Node       Outbound      Inbound       FlowC     Conflct Gcache     Appl
    time P cnf  # stat laten msgs data que msgs data que pause snt lcf bfa   ist  idx  %ef
18:09:08 P   2  2 Sync 1.3ms    0   0b   0    1 192b   0   0ns   0   0   0     1    0   0%
18:09:09 P   2  2 Dono 1.3ms    0   0b   0    1  64b   0   0ns   0   0   0     1    0   0%
...
18:09:19 P   2  2 Dono 1.3ms    0   0b   0    0   0b   0   0ns   0   0   0     1    0   0%
18:09:20 P   2  2 Sync 1.3ms    0   0b   0    2  16b   0   0ns   0   0   0     1    0   0%
18:09:21 P   2  2 Sync 1.3ms    0   0b   0    0   0b   0   0ns   0   0   0     1    0   0%
```

Node2’s myq_status

```
mycluster /   (idx: 1) / Galera 3.14(r53b88eb)
         Cluster  Node       Outbound      Inbound       FlowC     Conflct Gcache     Appl
    time P cnf  # stat laten msgs data que msgs data que pause snt lcf bfa   ist  idx  %ef
18:10:30 P   2  2 Sync 0.0ns    0   0b   0    3 208b   0   0ns   0   0   0     1    0   0%
18:10:31 P   2  2 Sync 0.0ns    0   0b   0    0   0b   0   0ns   0   0   0     1    0   0%
```

Async Slaves

PERCONA

http://bit.ly/1qBNFmW

Add Binary log with GTID

- All nodes in the same cluster get the same server-id
- Binary logs NOT required except for:
  - Async replication
  - Point in time backups

```
# Async replication setup
server_id                       = 1
log_bin                         = async_log
log_slave_updates
enforce_gtid_consistency        = 1
gtid_mode                       = ON
```

PERCONA

Slave Setup and Failover

- Enable binary logging on at least 2 cluster nodes
- Build Slave as normal from one of them
- If async master in cluster fails, use master_auto_position to repoint the slave:
  - CHANGE MASTER TO MASTER_HOST="<host/ip of other binlogging cluster node>", MASTER_AUTO_POSITION=1;
  - This could use a VIP, etc provided all possible nodes are binlogging.

Proxies and Load Balancing

MaxScale Install

- Old Reference Arch used HAproxy
  - Today we have some choices for auto-RW splitting
  - I selected MaxScale because:
    - *Probably the most production worthy at the moment*
    - *Open-source candidate*
    - *I've seen it used in production*
  - ProxySQL is a close contender
- Download or build binaries
  - https://www.percona.com/blog/2016/04/11/downloading-mariadb-maxscale-binaries/

```
# yum http://downloads.mariadb.com/enterprise/<secret link>/generate/10.1/mariadb-enterprise-repository.rpm
# yum install maxscale -y
```

http://bit.ly/1qBNFmW

MaxScale Config - /etc/maxscale.cnf

```
[maxscale]
threads=4

[node1]
type=server
address=172.28.128.3
port=3306
protocol=MySQLBackend

[node2]
type=server
address=172.28.128.4
port=3306
protocol=MySQLBackend

[node3]
type=server
address=172.28.128.5
port=3306
protocol=MySQLBackend
```

```
[Galera Monitor]
type=monitor
module=galeramon
servers=node1,node2,node3
user=maxscale
passwd=test
monitor_interval=5000

[Read Connection Router]
type=service
router=readconnroute
servers=node1,node2,node3
user=maxscale
passwd=test
router_options=slave

[RW Split Router]
type=service
router=readwritesplit
servers=node1,node2,node3
user=maxscale
passwd=test
max_slave_connections=100%
```

```
[Read Connection Listener]
type=listener
service=Read Connection Router
protocol=MySQLClient
address=0.0.0.0
port=4008
socket=/var/lib/maxscale/readconn.sock

[RW Split Listener]
type=listener
service=RW Split Router
protocol=MySQLClient
port=4006
#socket=/var/lib/maxscale/rwsplit.sock

[MaxAdmin Service]
type=service
router=cli

[MaxAdmin Listener]
type=listener
service=MaxAdmin Service
protocol=maxscaled
port=6603
```

http://bit.ly/1qBNFmW

Grant user and start MaxScale and Test

```
mysql> grant all on *.* to maxscale@'%' identified by 'test';

# systemctl start maxscale

[root@node2 ~]# mysql -u maxscale -ptest -h 127.0.0.1 -P 4008 -e "show global
variables like 'wsrep_node_name';"
+-----------------+-------+
| Variable_name   | Value |
+-----------------+-------+
| wsrep_node_name | node2 |
+-----------------+-------+
[root@node2 ~]# mysql -u maxscale -ptest -h 127.0.0.1 -P 4008 -e "show global
variables like 'wsrep_node_name';"
+-----------------+-------+
| Variable_name   | Value |
+-----------------+-------+
| wsrep_node_name | node3 |
+-----------------+-------+
```

PERCONA

http://bit.ly/1qBNFmW

Test Read/Write split

```
[root@node2 ~]# mysql -u maxscale -ptest -h 127.0.0.1 -P 4006 -e "show global
variables like 'wsrep_node_name';"
+-----------------+-------+
| Variable_name   | Value |
+-----------------+-------+
| wsrep_node_name | node3 |
+-----------------+-------+

[root@node2 ~]# mysql -u maxscale -ptest -h 127.0.0.1 -P 4006 -e "start transaction;
show global variables like 'wsrep_node_name';"
+-----------------+-------+
| Variable_name   | Value |
+-----------------+-------+
| wsrep_node_name | node1 |
+-----------------+-------+
```

Backups and Monitoring

Backups

- Remove backup node from Prod rotation
  - Possible to leave in rotation, but watch:
    - *resource consumption*
    - *flow control*
  - Dedicated Backup nodes not uncommon
- mysql> set global wsrep_desync=ON;
- Use Xtrabackup
- Wait for node's apply queue to drain
- mysql> set global wsrep_desync=OFF;
- Restore to rotation

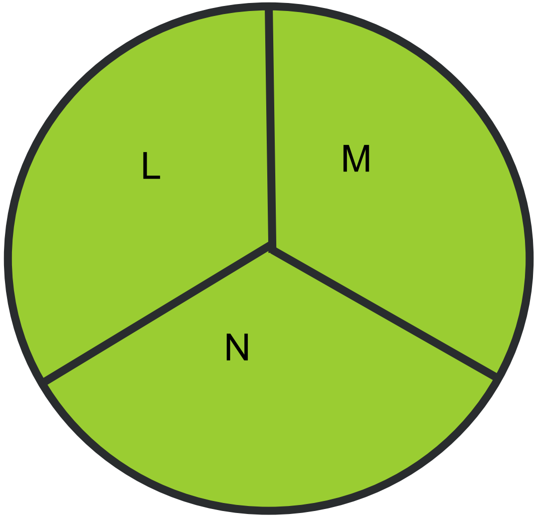
Monitoring

- Myq-status for real-time
  - You will learn a lot more about the cluster watching it second-by-second, esp in triage situations
- Alerting
  - Old info still applies, even if you don't use the same tools:
    - https://www.percona.com/blog/2013/10/31/percona-xtradb-cluster-galera-with-percona-monitoring-plugins/
- Trending
  - Use/Reproduce graphs from here:
    - https://www.percona.com/doc/percona-monitoring-plugins/1.1/cacti/galera-templates.html
  - Vividcortex does nicely, commercial

Special Setups

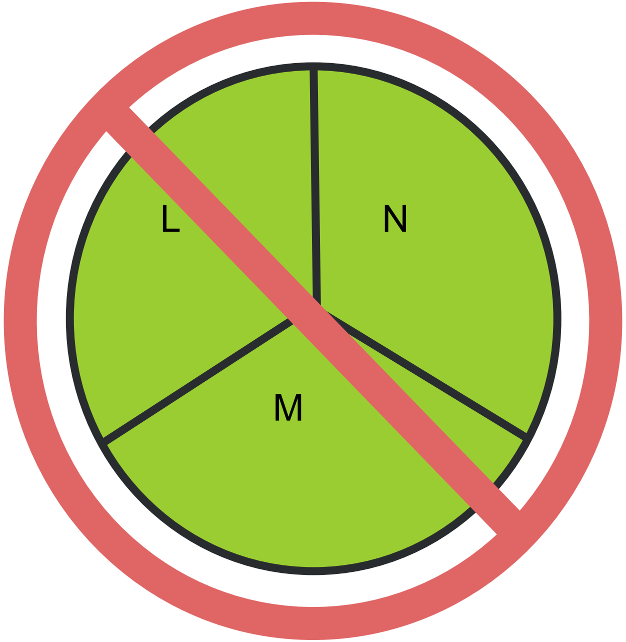
http://bit.ly/1qBNFmW

WAN Architecture Best Practices



L = M = N
Or

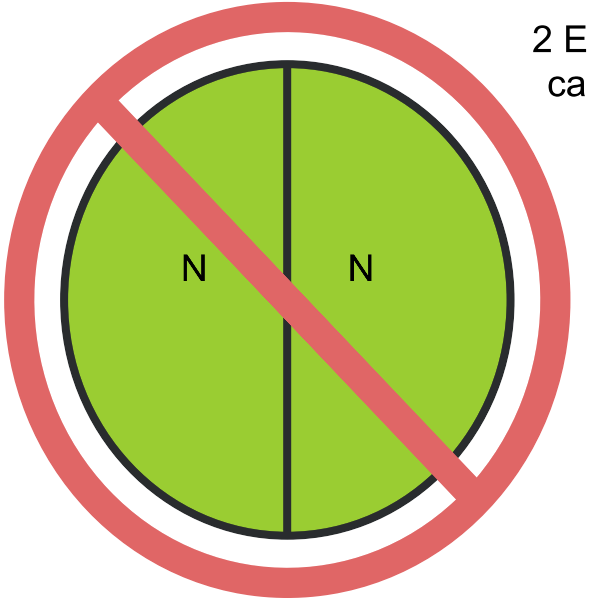
L+M>N
AND
L+N>M
AND
M+N>L



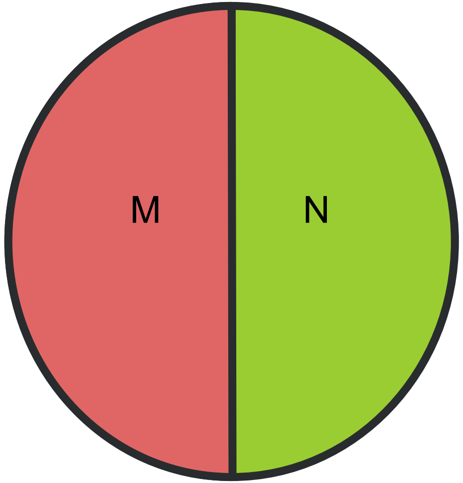
Any of:
L+N <=M
M+N <=L
L+M <=N

http://bit.ly/1qBNFmW

WAN Architecture Best Practices



2 Equal Active DCs can't auto failover

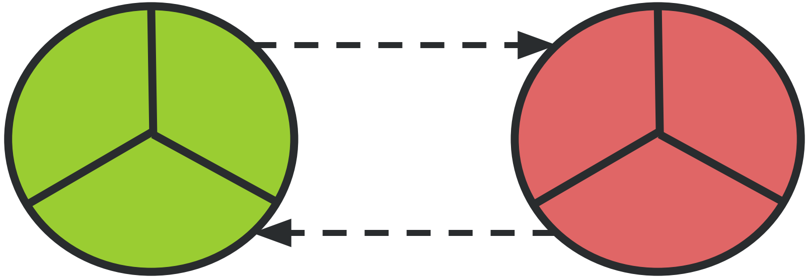


2 DCs work for a Primary/DR when N > M. Failover to DR is manual

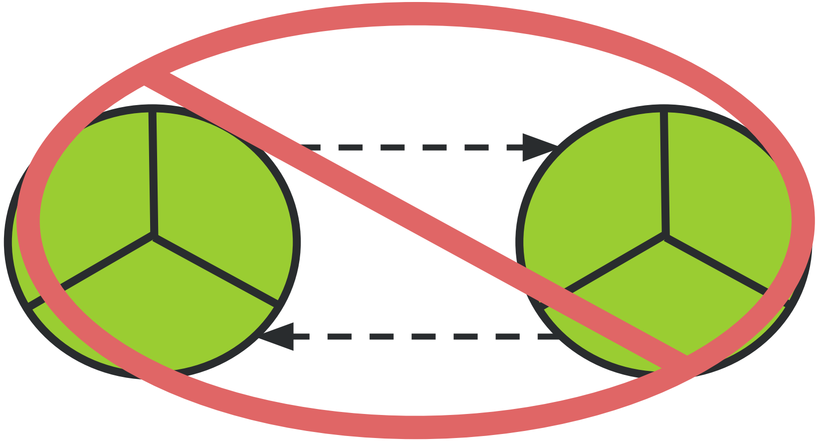
http://bit.ly/1qBNFmW

WAN Architecture Best Practices

Two clusters with Async.
DR Failover is manual



Two cluster Active/Active with Async



PERCONA

http://bit.ly/1qBNFmW

WAN Deployments

- All nodes must ack replication, commits will be slower!
- Tunings
  - Distinct gmcast.segment for each datacenter (e.g., 1, 2, etc.)
    - *Every node in that datacenter gets the same number*
  - Increase Replication windows
    - *Higher latency allows more "in-flight" replication at once*
  - Increase Timeouts above Cluster's Max RTT
    - *Larger timeouts will increase recovery time on a failure*
  - Only set wsrep_provider_options once in your my.cnf!

```
wsrep_provider_options="gmcast.segment=1; evs.send_window=512; evs.
user_send_window=512; evs.keepalive_period = PT3S; evs.suspect_timeout = PT30S;
evs.inactive_timeout = PT1M; evs.install_timeout = PT1M; evs.
join_retrans_period = PT3S; evs.delayed_margin = PT3S"
```

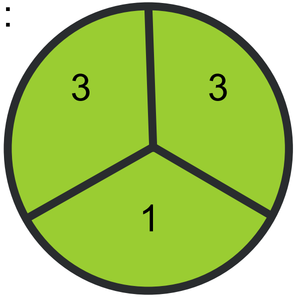
Dedicated Galera Networks

- Each node will use the default NIC (eth0) by default
  - Includes
    - *Galera Replication (port 4567)*
    - *State Snapshot Transfer (full backup) (port 4444)*
    - *Incremental State Transfer (port 4568)*
- Override all to another NIC with:
  - wsrep_node_address = <ip>
  - Possible to separate each item above if needed
- State transfers can max out a network in the right circumstances

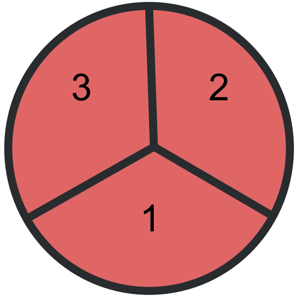
AWS Hosting

- Multi-AZ setups are best practice
  - Region with at least 3 AZs (some only have 2!)
  - At least one node in each AZ
  - One AZ loss, the rest should not lose quorum
- Elastic IPs
  - Could be useful for failover
  - Assumes the public IP of an instance
  - Assumes you are using private IPs for Galera

Good:



Bad:



Super Consistent Cluster

- Apply in Galera is still async
  - Write on node1, Read immediately from node2, might not be consistent yet (race)
- Flow control keeps apply lag low
  - Delays commits until node gets apply queue under threshold
- Can guarantee some/all reads are time-consistent
  - These can timeout however

```
[mysqld]
wsrep_sync_wait = 7
wsrep_provider_options = "gcs.fc_limit=16; gcs.fc_master_slave=yes; repl.
causal_read_timeout=PT5S"
```

Percona Live Amsterdam -- October 3-5

- **Super Saver tickets are already on sale but won’t last long!**
  - **Prices go up July 3rd.**
  - https://www.percona.com/live/plam16/registration
- **Call for Papers is Open!**
  - Do you have a MySQL, MongoDB and ODBMS use case to share, a skill to teach, or a big idea to share? We invite you to submit your speaking proposal for either breakout or tutorial sessions.
  - **The deadline to submit is July 18th, 2016.**
  - https://www.percona.com/live/plam16/program
- **Sponsorship opportunities Available**
  - https://www.percona.com/live/plam16/be-a-sponsor

Questions?

PERCONA