Galera Replicator IRL

Art van Scheppingen
Head of Database Engineering
Overview

1. Who are we?
2. What is Galera?
3. What is Spil Games using Galera for?
4. What have we learned?
5. Future technologies
6. Conclusion
Who are we?

Who is Spil Games?
Facts

- Company founded in 2001
- 350+ employees world wide
- 180M+ unique visitors per month
- Over 60M registered users
- 45 portals in 19 languages
  - Casual games
  - Social games
  - Real time multiplayer games
  - Mobile games
- 35+ MySQL clusters
- 60k queries per second (3.5 billion qpd)
Geographic Reach

180 Million Monthly Active Users(*)

Source: (*) Google Analytics, August 2012
What is Galera?
How to get Highly Available and beyond
What is Galera?

1. Replication plugin for MySQL by Codership
   • Synchronous (parallel) replication
   • Supports InnoDB
   • MyISAM “works”
   • Committing transactions actually replicates data

2. Allows clustering of nodes
   • Minimum of 3 nodes for HA
     • Galera Arbitrator allows 2 nodes
   • One node elected as Primary Component
How does Galera work?

Server-1
Server-2
Server-n

MySQL
MySQL
MySQL

Connect/read/write to any node

Galera
Synchronous replication
Galera replication

Client receives OK

MySQL

MySQL

MySQL

commit

Galera replication

Transaction applied to slaves
High Availability (1)
High Availability (2)

Server-1  
Server-2  
Server-n

Load balancer / Query router

MySQL  
MySQL  
MySQL

Galera

spilgames
High Availability (3)

Server-1
Load balancer
MySQL

Server-2
Load balancer
MySQL

Server-n
Load balancer
MySQL

Galera
High Availability (4)

Server-1  Server-2  Server-n

Load balancer (port 3306 read, 3307 write)

MySQL  MySQL  MySQL
read+write  read only  read only

Galera
Node joining SST (State Snapshot Transfer)

Server-1  Server-2  Server-n

Load balancer

MySQL  MySQL  MySQL  MySQL

Cluster drains node

SST

Requests to join cluster

Galera

Synchronous replication

read/write to two nodes

spilgames
Node joining IST (Incremental State Transfer)
Galera replication over WAN

Commit is delayed by RTT
WAN replication Galera 2.x
WAN replication Galera 3.x
What are we using Galera for?

Synchronous replication for the masses
Our systems

1. Legacy services databases
   • MySQL Master-Master
2. SSP (Spil Storage Platform)
   • MySQL Master-Master (to be phased out)
   • Galera
3. ROAR (Read Often, Alter Rarely)
   • Galera
Master-Master setup used at Spil Games

Server-1

MySQL active master

db-something (192.168.1.1)
db-something-r1 (192.168.1.2)

Server-2

MySQL active master

Server-n

MySQL inactive master

db-something-r2 (192.168.1.3)

Asynchronous replication

MMM
Master-Master setup used at Spil Games

MySQL active master

MySQL slave

MySQL inactive master

MySQL slave

db-something (192.168.1.1)
db-something-r1 (192.168.1.2)
db-something-r3 (192.168.1.4)
db-something-r2 (192.168.1.3)
db-something-r4 (192.168.1.5)
Migrating legacy dbs to Galera (lab)

- **Legacy 1 (inactive master)**
  - Clone database (innobackupex)
  - Start slaving

- **Legacy 2 (inactive master)**
  - Feed database dump (mysqldump)

- **Legacy 3 (inactive master)**

Galera
Scaling Galera (1)

Server-1

Server-2

Server-n

Load balancer (port 3306 write+read)

MySQL

MySQL

MySQL

MySQL

MySQL

MySQL

Galera
Scaling Galera (2)

Server-1 | Server-2 | Server-n
---|---|---
Load balancer (port 3306 write, 3307 read)

MySQL | MySQL | MySQL | MySQL | MySQL
---|---|---|---|---
read only | read only

MySQL - MySQL
asynchronous replication

Galera
Why consolidate legacy systems?

1. Around 20 legacy database clusters
   • 50 servers in total
2. Maintenance
   • Master-Master requires a lot of (manual) maintenance
3. Replacement is needed
   • 35 of them will be older than 3 years in 2014
4. Current state: tested in lab
SSP (Spil Storage Platform)

• Storage API between application and databases
• All data is sharded
  • User
  • Function
  • Location
• Every cluster (two masters) will contain two shards
• Data written interleaved
• HA for both shards
• Both masters active and “warmed up”
SSP Master-Master setup

Server-1

MySQL active master
db-ssp001 (192.168.2.1)

Server-2

MySQL active master
db-ssp002 (192.168.2.2)

Server-n

Asynchronous replication

read+write

MMM

MySQL ackve master

MySQL ackve master
SSP Master-Master setup

Server-1
Server-2
Server-n

MySQL broken master

MySQL active master

read+write

MMM
db-ssp001 (192.168.2.1)
db-ssp002 (192.168.2.2)

spilgames
SSP Galera setup

Server-1  Server-2  Server-n

Load balancer

MySQL  MySQL  MySQL

Galera  Synchronous replication

read/write to any node
Current state of the SSP

1. Total of 4 old style SSP shard nodes (2 clusters)
2. Total of 6 Galera SSP shard nodes (2 clusters)
3. Add Galera nodes/clusters when necessary
What have we learned so far?

Pitfalls, hurdles, etc
Creating backups

1. Two ways to make backups:
   • Issue SST
     • Either mysqldump or Innobackupex
   • Regular Innobackupex
     • --galera_info
     • set global wsrep_desync=on to remove node
Backup SST

- Server-1
- Server-2
- Server-n

Load balancer

MySQL
MySQL
MySQL

SST
Backup done
Request SST

Cluster drains node

Synchronous replication

Galera

Read/write to two nodes
Backup Innobackupex

Server-1 -> Server-2 -> Server-n

Load balancer

wsrep_desync=ONF

Stream backup

BackupPC -> MySQL -> MySQL -> MySQL

Galera
Synchronous replication

MySQL
read/write to two nodes
Restoring backups

1. Restored backup can be used to prevent SST of new joiners
2. Automated backup verification
   • Restores (randomly) chosen backup
   • Installs necessary MySQL version (5.1/5.5)
   • Perform basic checks
   • Enable replication
     • Will not work fully as it needs a working cluster to join
Monitoring

1. Cluster
   • Nodes in the cluster
     • Warning at 2, critical at 1
   • Availability of the address
2. Load balancer
   • Node checks
3. Performance monitoring
   • Adding metrics to mysql_statsd is easy
   • wsrep_flow_control
Flow control

1. Usage of replication threads
   • Scale from 0.0 to 1.0
2. Recommended to stay below 0.1 (10% blocked)
3. Adding more nodes will not solve your problem
4. Increase replication threads
   • Recommended 2*CPU cores
   • What if 64 is not enough?
   • How do you close flood gates?
Other things we bumped into…

1. MySQL version updates
   • Update one by one
   • PXC SST changes

2. Availability after restart
   • Joins cluster after IST/SST
   • LRU still loading

3. In descriptive errors during SST
   • Local user authentication (after starting mysqld with sudo!)

4. Schema changes

spilgames
Future for Galera at Spil Games

What will we do in the near future?
Openstack

1. Offer DAAS to our (internal) customers
2. Spawning (automated) database nodes and clusters when necessary
3. Mix and match Galera and regular MySQL replication
WAN Replication

1. No immediate use case (yet)
   • No need for WAN in sharded environment
   • Game catalogue might need it in the future
2. Wait for Galera 3.0
   • Datacenter awareness
MaxScale

1. Beta testing MaxScale for SkySQL
   • Works flawless in the lab (so far)
   • Not yet tested with mixed Galera/MySQL replication
2. MaxScale itself is not HA (yet)
   • Keepalived?
Conclusion

What is our verdict?

CONCLUSION: Bubble-flippin-tastic!
Conclusion(s)

1. Galera definitely live up to expectations
2. Decreased cluster wide performance
3. Increased replication performance
4. High investment in time for initial setup/tools
5. Maintenance is easier
6. Well worth the investment for us
Thank you!

• Presentation can be found at: http://spil.com/fosdem2014
• Mysql_statsd can be found at: http://spil.com/mysqlstatsd
  http://github.com/spilgames/mysql-statsd
• If you wish to contact me:
  Email: art@spilgames.com
  Twitter: @banpei
• Engineering @ Spil Games
  Blog: http://engineering.spilgames.com
  Twitter: @spilengineering
Photo sources

Our current HA environment:
http://thinkaurelius.com/2013/03/30/titan-server-from-a-single-server-to-a-highly-available-cluster/

What we have learned so far:
http://renaiissanceronin.wordpress.com/2009/10/05/playing-with-plasma-cutters/

Near future:
http://www.example-infographics.com/envisioning-the-near-future-of-technology/

Conclusion:
http://www.flickr.com/photos/louisephotography/5796499806/in/photostream/