AliSQL: Breakthrough for the Future

Wei Hu
Database Kernel Developer @ Alibaba Group
Percona live 2017
Apr 27, 2017
About The Talk

01 / A Bit of History about AliSQL
Introduction of AliSQL and what we did in the past years

02 / X-KV
AliSQL key-value interface

03 / X-Cluster
AliSQL cluster solution
HISTORY
A bit of history about AliSQL
Roadmap

2011
- Bug fix for DDL
  - Eliminate race condition

2014
- SQL firewall
  - Optimize Thread pool

2017
- X-KV
  - X-Cluster

AliSQL 5.1
- 2012
  - Parallel Replication
  - Optimize Hot SKU Version 1

AliSQL 5.5
- 2014
- Optimize Hot SKU Version 1

AliSQL 5.6
- 2016
- Group Update for Hot SKU
  - RocksDB support

AliSQL 5.7
- 2017
- X-KV
  - X-Cluster
First View of AliSQL

Percona live 2016 What's new in AliSQL- Alibaba's branch of MySQL

- commit_on_success hints
- Enhanced Thread Pool
- Hot SKU Optimization
- Dual redo log buffers
- PK access optimization
- Enhanced Semi-Sync Replication
- Select from Update
- Table-level Parallel Replication
- Table Lock optimization for select
- Single row update reached 100,000 per second
  By one AliSQL instance last year
Single’s Day is Always a challenge

We Need...

- Extreme performance
- Strong data consistency
- Low Cost
X-KV
Key-value interface of AliSQL
What is X-KV
AliSQL KV Interface

An extended Key-Value API for AliSQL which based on InnoDB memcache plugin

- There were so many simple queries according to unique keys
- Data Consistency is very important for us
- SQL parsing and Query Optimization is too expensive

Differences between InnoDB memcache plugin and X-KV?
What is X-KV
AliSQL KV Interface

Current InnoDB memcache plugin cannot satisfy our needs

- Support More data types and index types
- Design new protocol
- Optimize the performance
- Easy operation and maintenance
Data Types Support

X-KV

Char/Varchar/Blob are not enough!!!

Float/Double  Integer  ....

Decimal      DateTime

Char/Varchar  Custom Format  ....
Features extension

X-KV

- Multi get from multi tables
- Non-unique index
- Composite indexes & prefix search
New Protocol

X-KV

- How to deal with NULL value?
- How to choose delimiter?
- How to support composite index?

Field  = Meta info + Data

Comosite index support:

get [@@name.]key1 [key2] ... [keyN]

Meta info format:

Version | Length1 | Length2 | Length3 |
Convenience

X-KV

How to deal the cases when user table schema changed

Original solution is to stop request and then reinstall plugin

Unacceptable

X-KV Support Online Schema Change

- Not block user queries from KV interface
- Configurations are loaded dynamically
X-KV performance optimization

Profiling result shows bottlenecks in our production environment

- Efficiency of serialization
- Allocation of transactions
- Container info verifying

![X-KV Performance Optimization Chart]
X-Cluster
AliSQL Cluster base on X-Paxos
Background

Article: Existential Consistency: Measuring and Understanding Consistency at Facebook

Large web services faces a trade-off between stronger forms of consistency and higher performance properties.

For Alibaba:

- Data consistency is the most important thing
- High performance is also indispensable
Ideas

The cluster solution must...

- Using Consensus protocol to handle the data consistency problem
- Take advantages of AliSQL
- Using Batching/Pipelining/Asynchronization to boost performance
Our Problems

Huge Network Latency!!!
Performance?
What is X-Cluster

AliSQL Cluster With high availability and strong consistency

X-Paxos instead of M-S Replication
X-Paxos
Consensus Library With Extreme High Performance And Availability

- Multi Paxos Algorithm
- Async Framework
- Pluggable Log
X-Paxos

Features Description

- Online Member Change
- Election Weights
- Customized Role
X-Cluster System With X-Paxos

Witness SDK
X-Cluster

- X Cluster uses X-Paxos Pluggable log feature, and design the consensus log with binlog format.

- Consensus Log = Binlog + Relay Log, One instance One Consensus log.

- Transaction is permitted to commit only if majority of instances in the cluster synchronized the log.
Optimization on Performance
Log Replication & Server layer
Optimization on X-Paxos
Log Replication batching & Pipelining

Leader

Batch 1
N log Entries

Batch 2
M log Entries

Batch 3
K log Entries

WAN

ACK Batch 1

ACK Batch 2

ACK Batch 3

Follower

Batch 1
N log Entries

Batch 2
M log Entries

Batch 3
K log Entries

WAN

ACK Batch 2

ACK Batch 3

ACK Batch 1

VS

Leader

Batch 1
N log Entries

Batch 2
M log Entries

Batch 3
K log Entries

WAN

ACK Batch 2

ACK Batch 3

ACK Batch 1

Follower

Batch 2
M log Entries

Batch 3
K log Entries

Batch 1
N log Entries

Reorder & Comit in Order
Performance of X-Paxos

Low latency & High latency

Replication Benchmark

Throughput of X-Paxos is almost same in low and high latency network
Optimization of X-Cluster

Another problem in High Latency network

Thread pool limit the number of concurrent worker threads to eliminate the data race, but most of the threads are waiting for Committing !!!
Benchmark

Sysbench Insert & Oltp TPS Performance

INSERT

<table>
<thead>
<tr>
<th></th>
<th>AliSQL-5.7</th>
<th>X-Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insert</td>
<td>121913.16</td>
<td>110432.3</td>
</tr>
<tr>
<td>OLTP</td>
<td>311701.2</td>
<td>275127.17</td>
</tr>
</tbody>
</table>

Sysbench Insert & Oltp RT Performance

INSERT

<table>
<thead>
<tr>
<th></th>
<th>AliSQL-5.7</th>
<th>X-Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insert</td>
<td>2.8</td>
<td>3.4</td>
</tr>
<tr>
<td>OLTP</td>
<td>16.3</td>
<td>17.7</td>
</tr>
</tbody>
</table>
X-Cluster vs GR

Sysbench QPS in low latency (0.1ms) network

<table>
<thead>
<tr>
<th></th>
<th>INSERT</th>
<th>OLTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-Cluster</td>
<td>110432.3</td>
<td>275127.17</td>
</tr>
<tr>
<td>5.7.17 GR</td>
<td>41290.5</td>
<td>241317.4</td>
</tr>
</tbody>
</table>

Sysbench RT in low latency (0.1ms) network

<table>
<thead>
<tr>
<th></th>
<th>INSERT</th>
<th>OLTP</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-Cluster</td>
<td>3.43</td>
<td>22</td>
</tr>
<tr>
<td>5.7.17 GR</td>
<td>7</td>
<td>25</td>
</tr>
</tbody>
</table>
X-Cluster vs GR

Sysbench QPS in high latency (30ms) network

- INSERT: X-Cluster 50366.62 vs 5.7.17 GR 11275
- OLTP: X-Cluster 265616.02 vs 5.7.17 GR 104974

Sysbench RT in high latency (30ms) network

- INSERT: X-Cluster 58 vs 5.7.17 GR 150
- OLTP: X-Cluster 98 vs 5.7.17 GR 304
Summary

X-KV and X-Cluster is Our new solution base AliSQL

- **X-KV**
  - More data types
  - New protocol
  - Better performance
  - Easy maintenance

- **X-Cluster**
  - Strong consistency
  - Geo distributed
  - High performance
Wei Hu

Database Kernel Developer @ Alibaba Group

Contact info:
Email: droopy.hw#alibaba-inc.com
Facebook: Droopy Hu
Q&A

Thank you

Alibaba Database Team