2019
05
08-10
北京新云南皇冠假日酒店

数据风云十年变迁

DTCC 第十届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2019

IT168 ChinaUnix ITPUB

基于MGR的读写强一致性数据库

滴滴出行 数据库团队

SPEAKER

INTRODUCE

田佳伟 数据库开发专家

曾就职于百度、Thomson Reuters 等公司，现就职于滴滴出行，负责数据库产品的架构及研发工作；多年中间件和数据库开发经验，研究方向主要为分布式数据库系统的理论与实践。

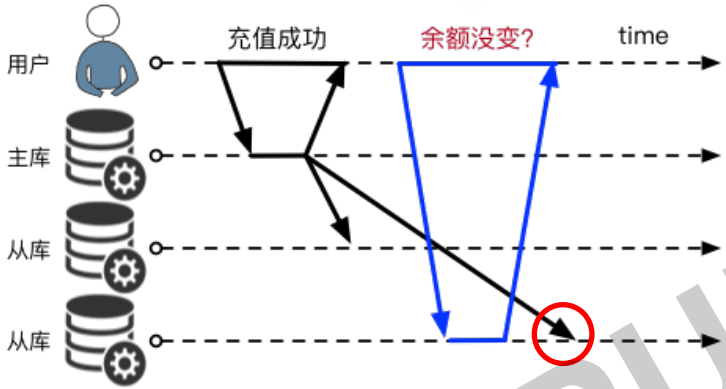
目录
Contents

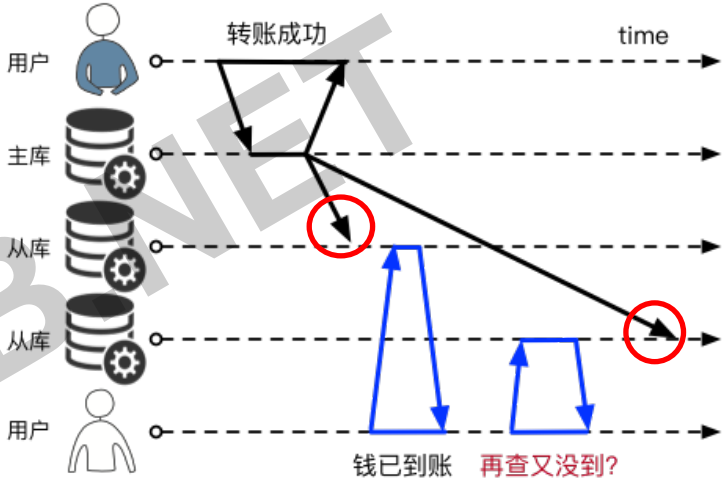
1

PART ONE

What's the reality ?

现实中的交易场景



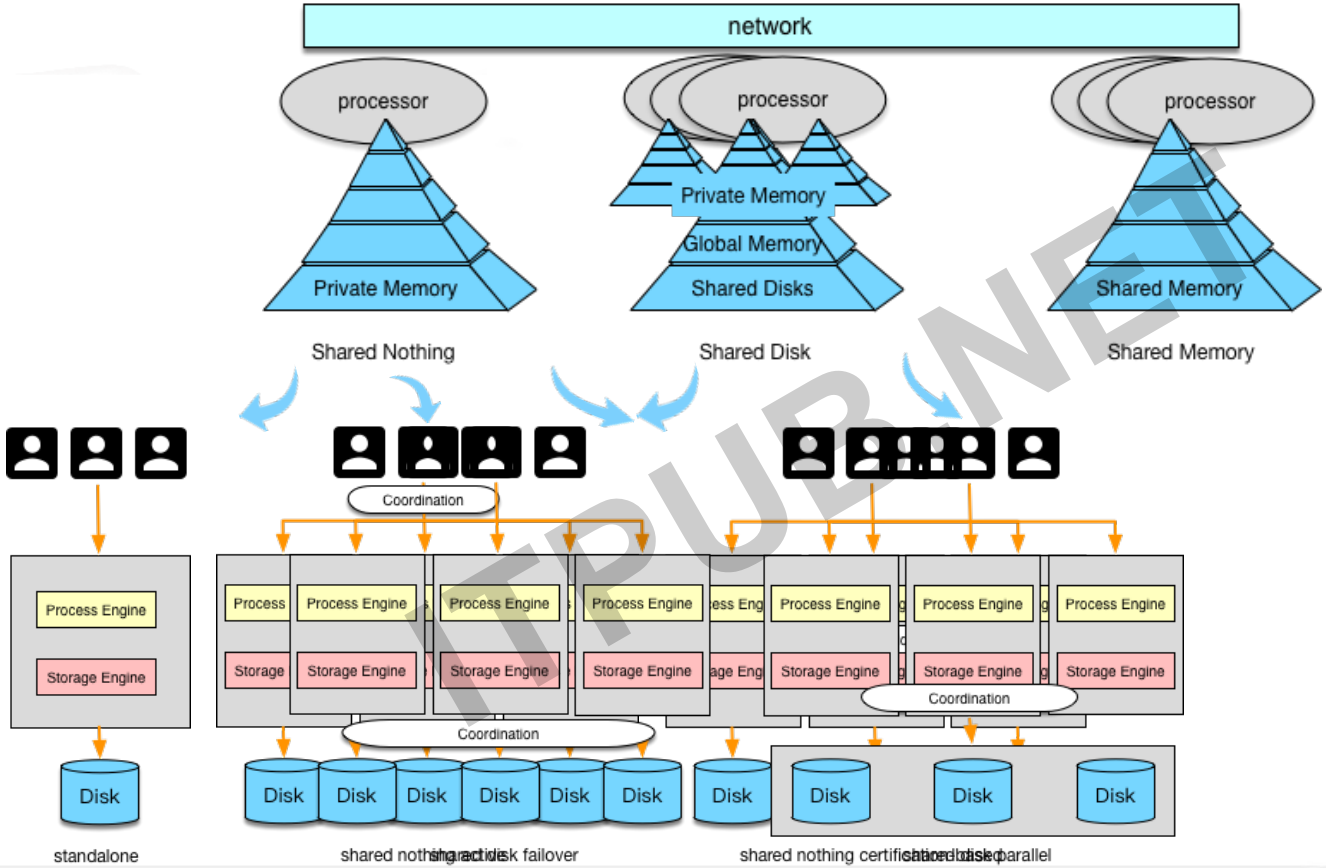


2

PART TWO

Write Consistency

数据库模型



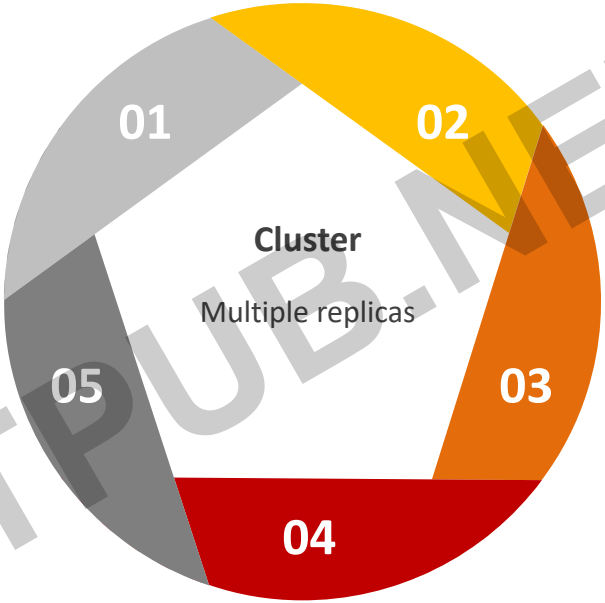
分布式数据库的核心能力

01. Scalability

弹性的扩展能力

02. Fault-Tolerant

可以容忍实例、机器、存储和网络的异常



03. Write Consistency

数据一旦写入成功，不会丢失

05. Availability

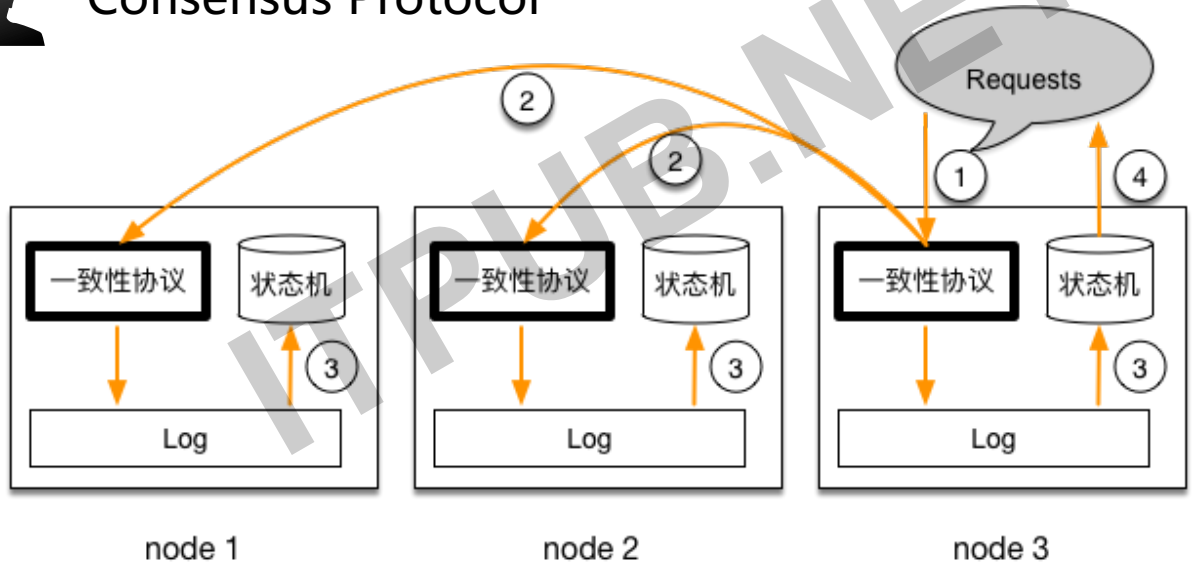
保证99.99%的可用性

04. Read Consistency

保证数据读取的线性一致性

Replicated State Machine

Consensus Protocol



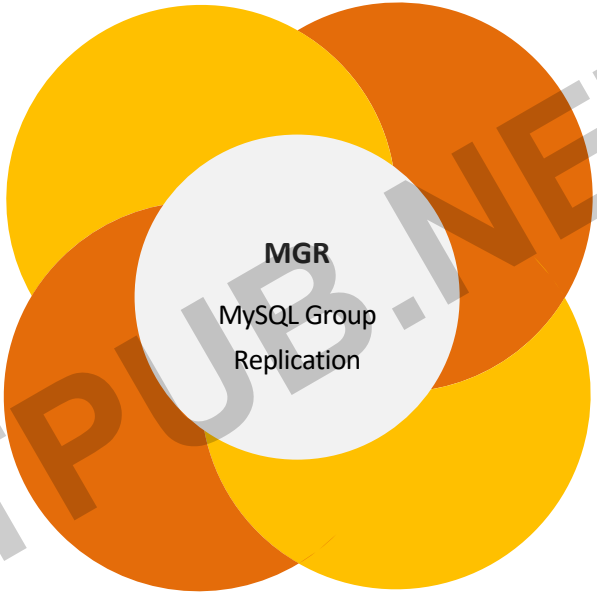
为什么选择MySQL Group Replication

01. Consensus Protocol

Mencious (Paxos variant)
保证线性一致性
全局同序处理

02. Stability

官方支持
社区生态完善



03. Multiple Primary

多点写入
提升可用性

04. Focus

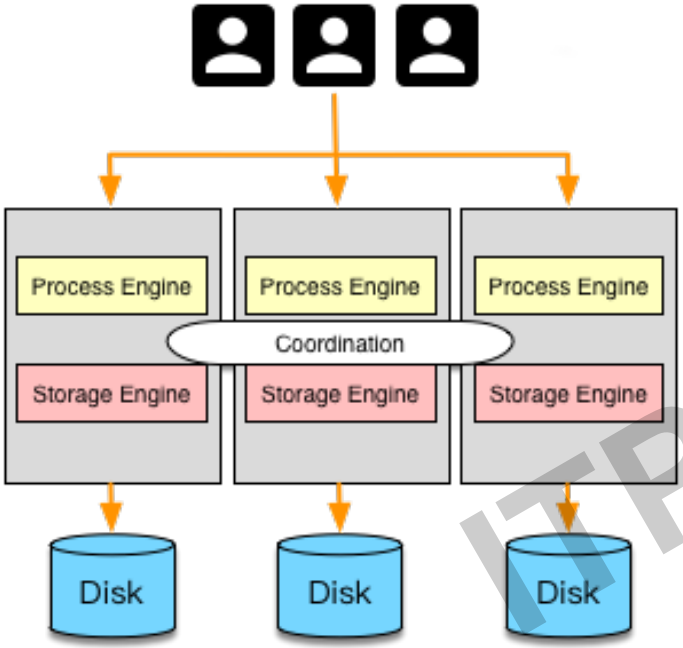
读取的线性一致性
完善一致性语义

3

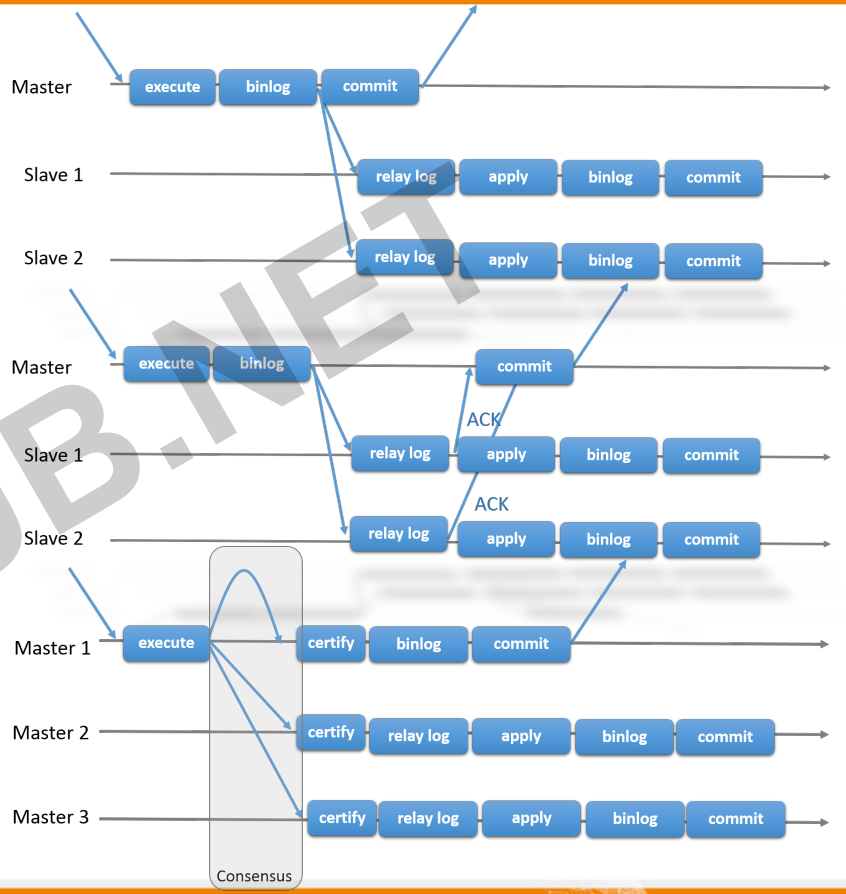
PART THREE

Read Consistency

MGR的架构和复制方式



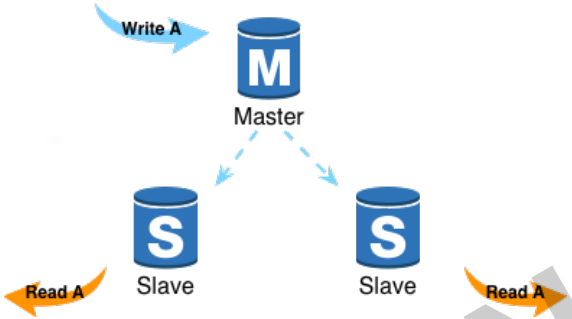
shared nothing certification-based



01

选择方案

1. 读一致性的定义
2. 现有的通用一致性方式是什么
3. 我们可以用哪些方式来做到读一致性
4. 我们的最佳选择



- read-your-writes consistency
- monotonic-reads
- consistent prefix reads

HOW

01. Proxy Route

通过Proxy流量
调度的方式

02. Read convert Write

读转换为写
走一遍Paxos协议

03. Quonum Read

W + R > N
Drawer Principle

04 Read with Write-Point

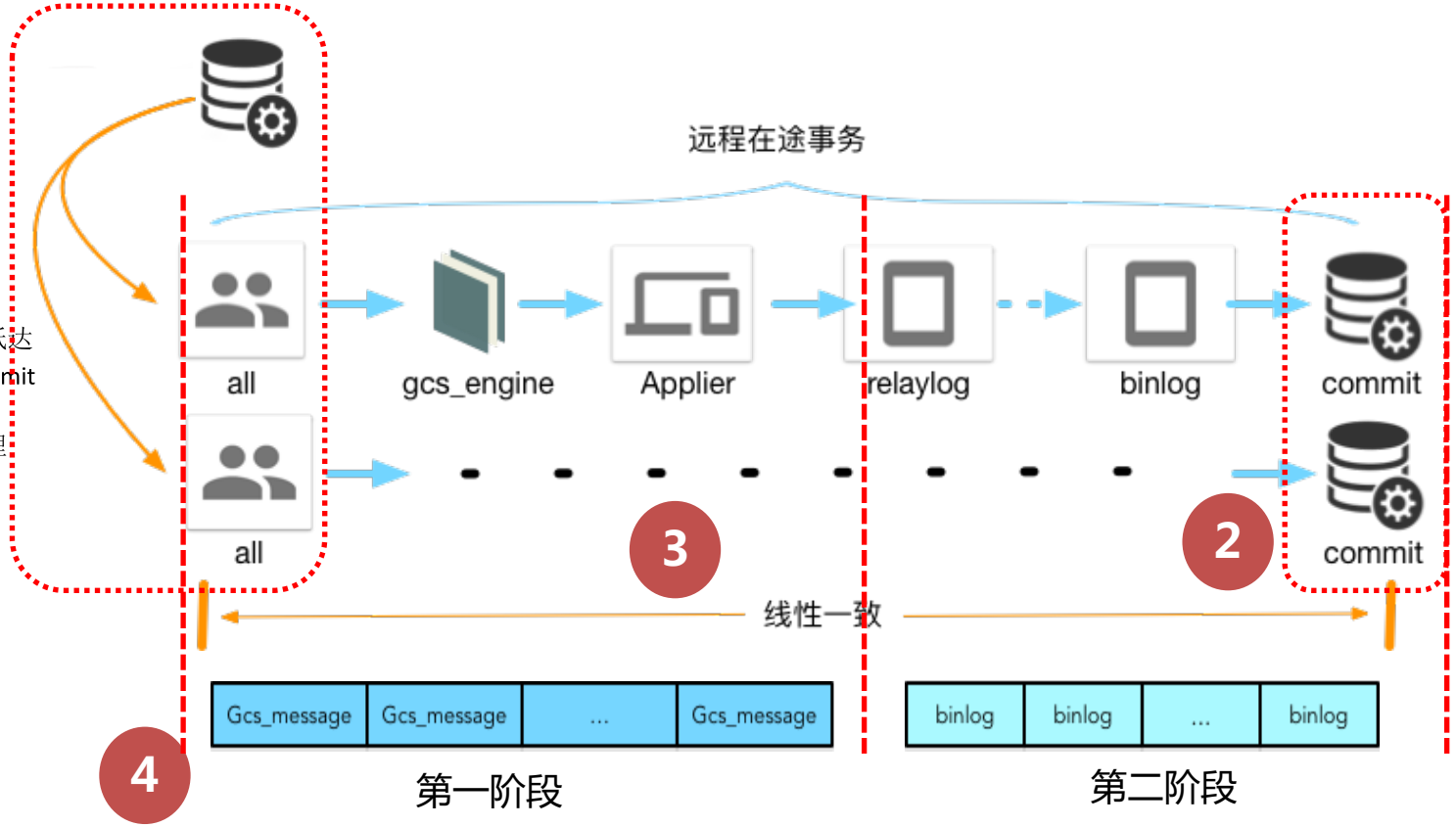
记录写入点
读取时携带写入点
推进到写入点后下放读

05 all sync + read wait until commit

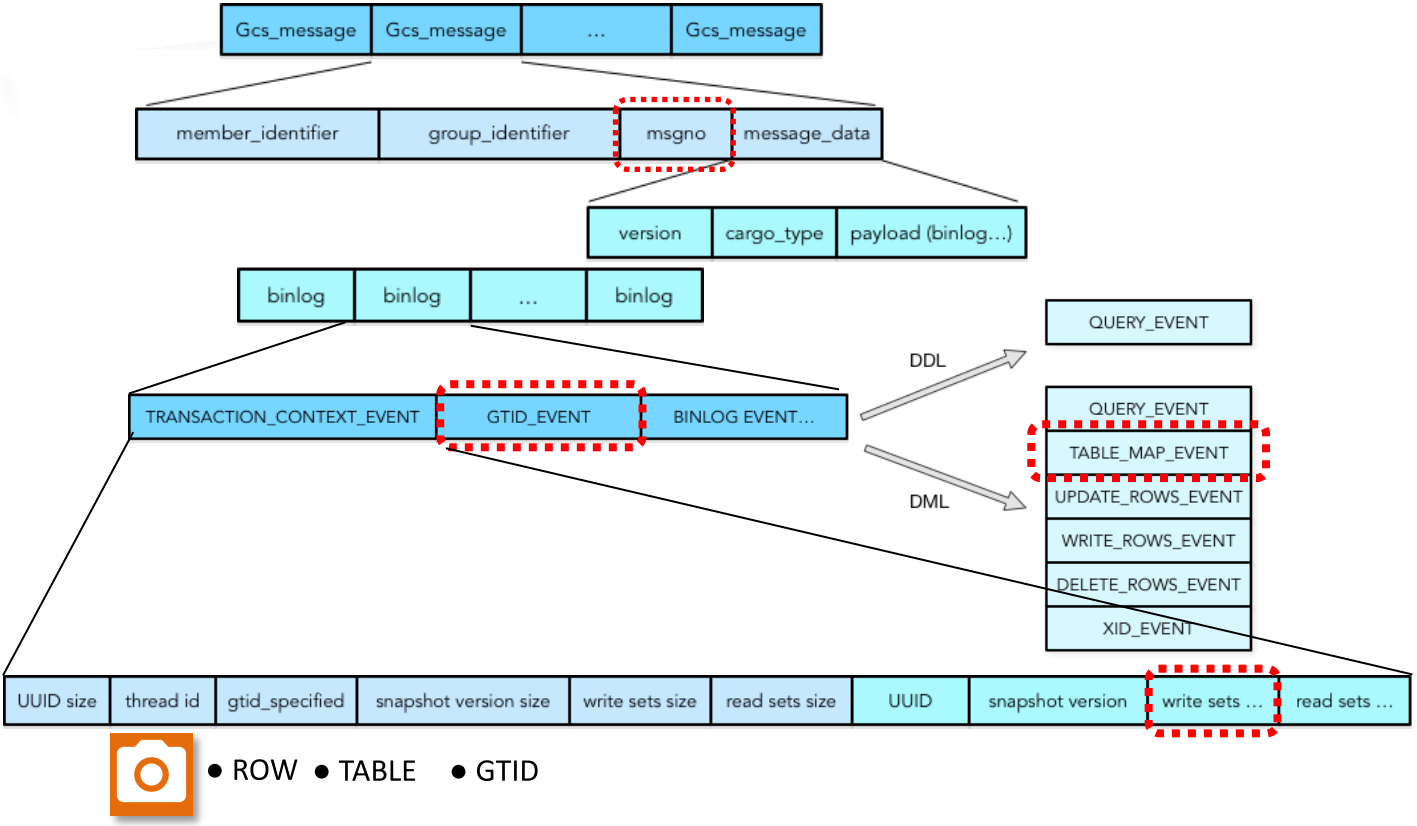
02

设计演进

1. 全同步，保证写入的数据抵达
2. 读达到时，等待engine commit 后查询
3. 各个节点保证全局同序处理
4. 远程在途事务拆解



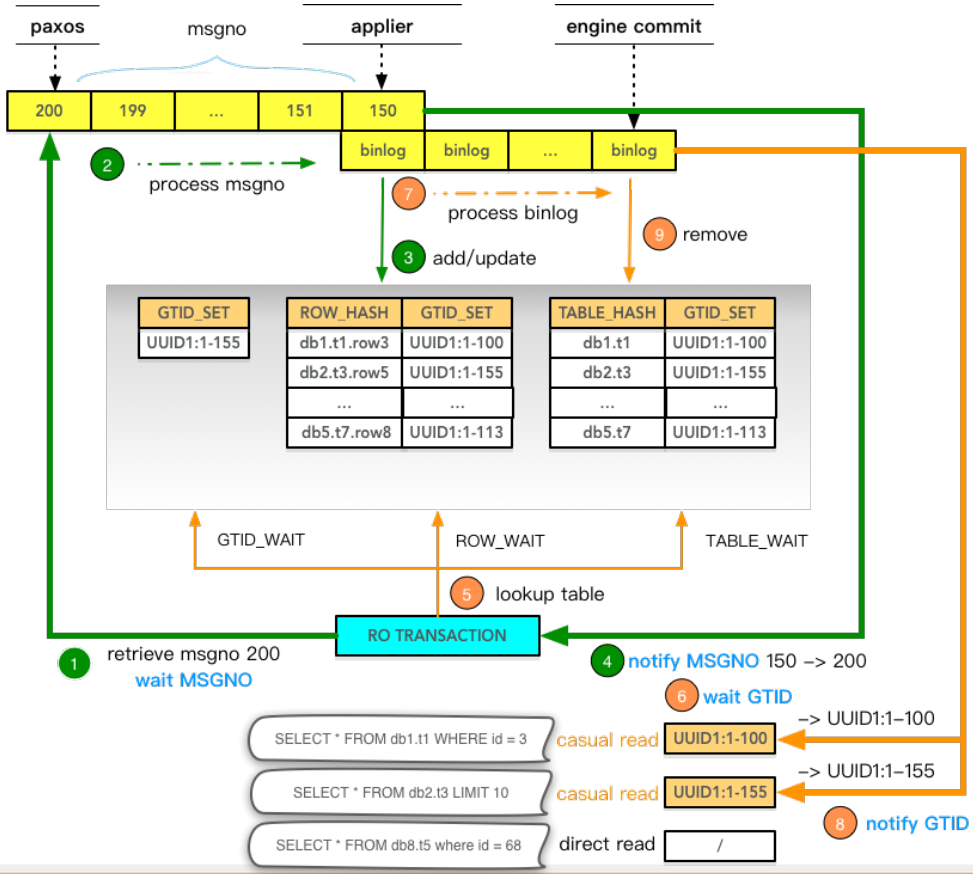
读一致性



03

两阶段&Casual Read

1. 分析gcs_message的组成
2. 分析binlog的组成
3. 因果读减少等待成本



04

详细流程

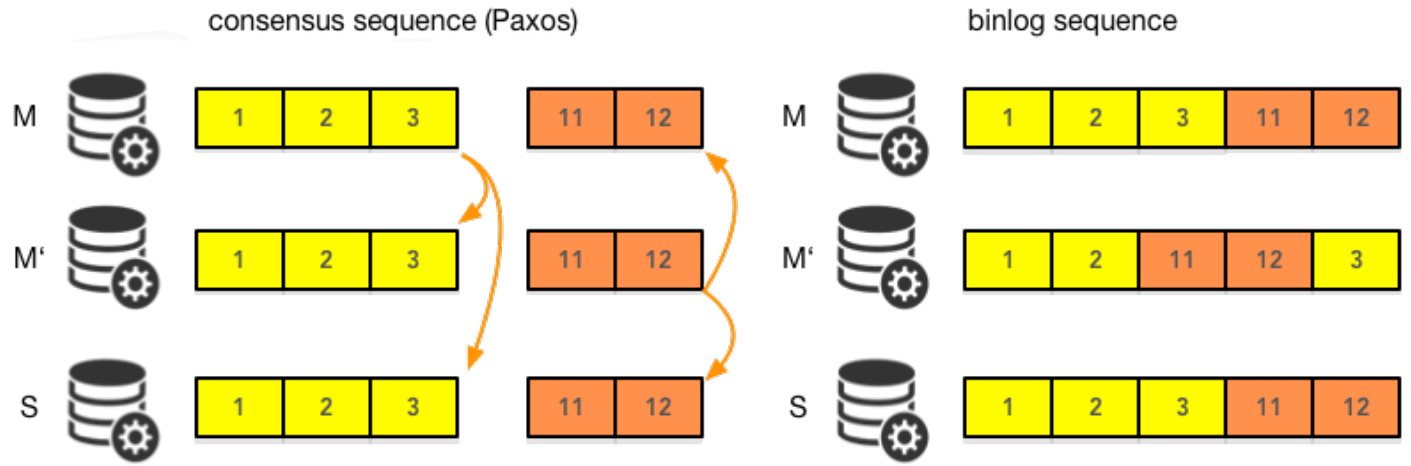
1. 第一阶段 停-等 msgno
2. 第二阶段 停-等 gtid

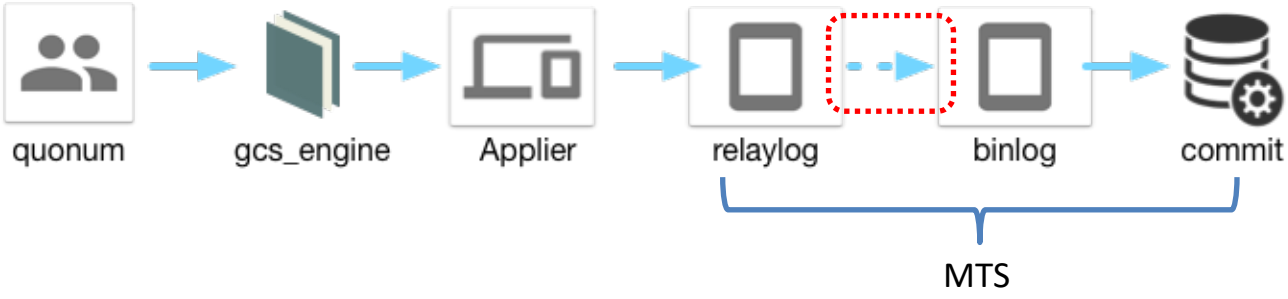
Conclusion

Make simple

appear as only one copy of data

日志的一致性



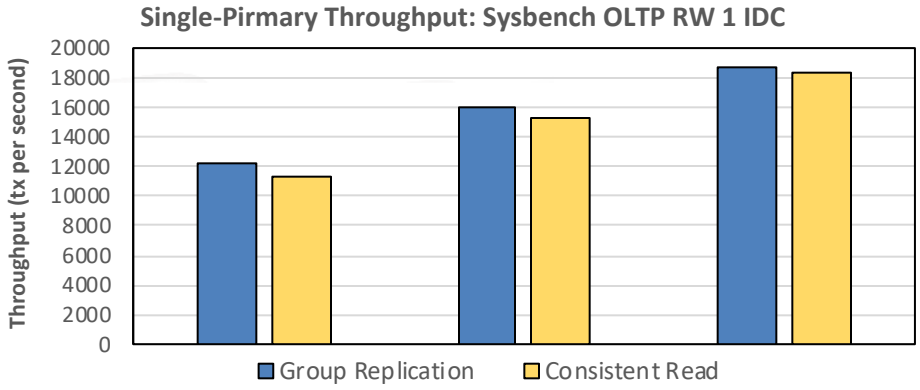


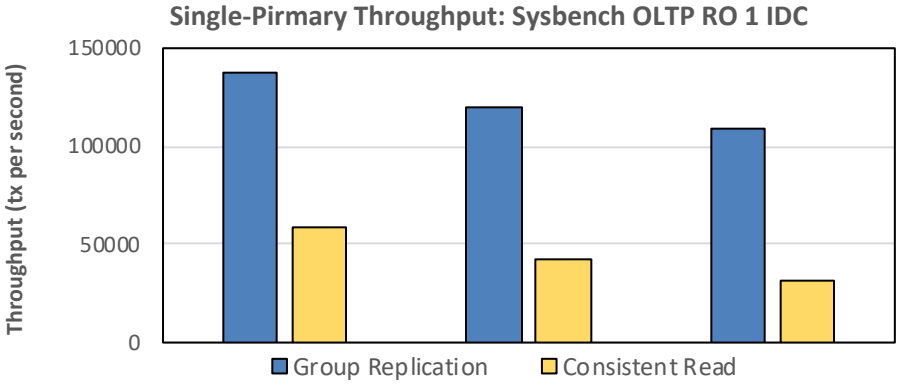
05

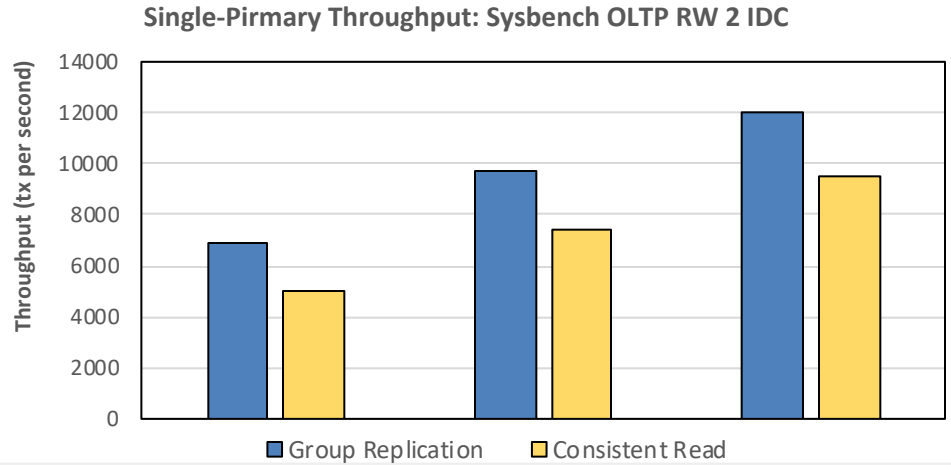
主从切换场景

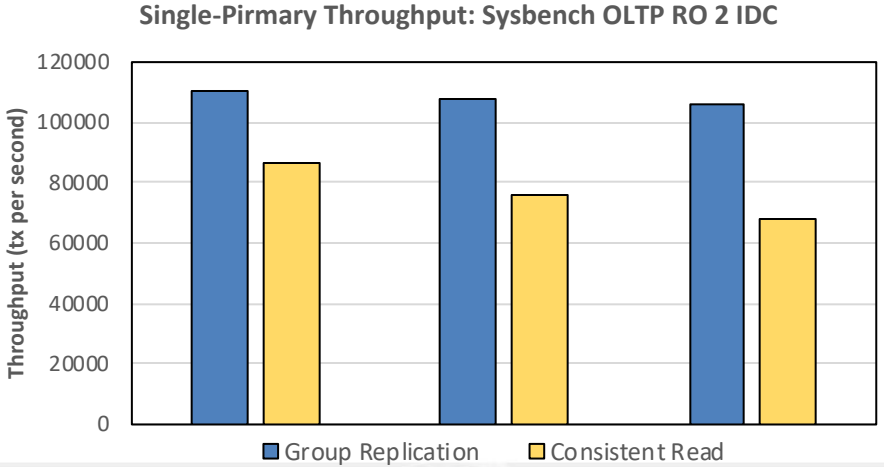
1.一致性协议保证传播线性一致
2. 应用层的日志binlog有可能不一致

性能


Single-Pirmary Throughput: Sysbench OLTP RW 1 IDC
Throughput (tx per second)
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Group Replication
Consistent Read


Single-Pirmary Throughput: Sysbench OLTP RO 1 IDC
Throughput (tx per second)
150000
100000
50000
0
Group Replication
Consistent Read


Single-Pirmary Throughput: Sysbench OLTP RW 2 IDC
Throughput (tx per second)
14000
12000
10000
8000
6000
4000
2000
0
Group Replication
Consistent Read


Single-Pirmary Throughput: Sysbench OLTP RO 2 IDC
Throughput (tx per second)
120000
100000
80000
60000
40000
20000
0
Group Replication
Consistent Read

最佳实践

内存使用

冲突检测库，在途事务，恢复时对增量数据的缓存
Paxos消息缓存，writesets开销

节点管理

失败节点重新加入
3+1容灾

读一致性

错误节点摘除读流量
开启流量控制、大事务限制

4

PART FOUR

Blueprint

规划和展望

1

Optimization

冲突检测库的gc优化
内存容量管理
writesets版本

2

Multiple Primary

多写试点
整合一致性读
跨机房优化

3

Enhance Consistency Read

优化性能
提前回放
版本管理

4

Port with MySQL 8.0.14

一致性增强
自动rejoin
指定主库


THANKS