Alibaba Group 阿里巴巴集团

阿里数据库内核技术演进之路

阿里巴巴集团 数据库事业部 胡炜

Weibo: @我的书包不见了

Email: droopy.hw#alibaba-inc.com

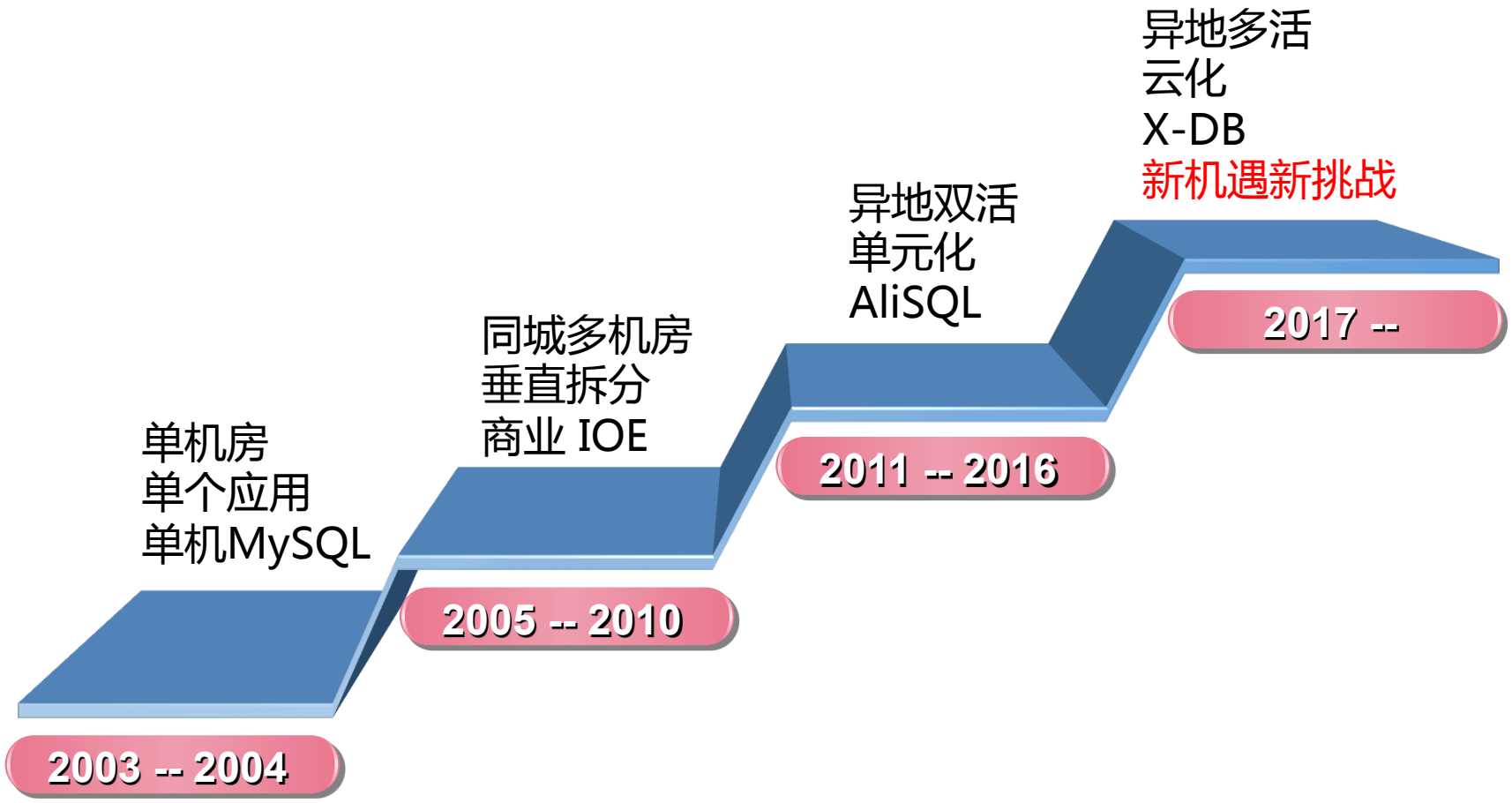
阿里巴巴的场景

- 阿里巴巴的数据库能
- 业务高速发展，体量巨大
- 业务多样，场景丰富、从电商开始逐渐覆盖各种复杂的业务场景
- 性能要求高，数据一致性要求高

Alibaba Group
阿里巴巴集团

阿里数据库历史

阿里数据库体系的四个时代



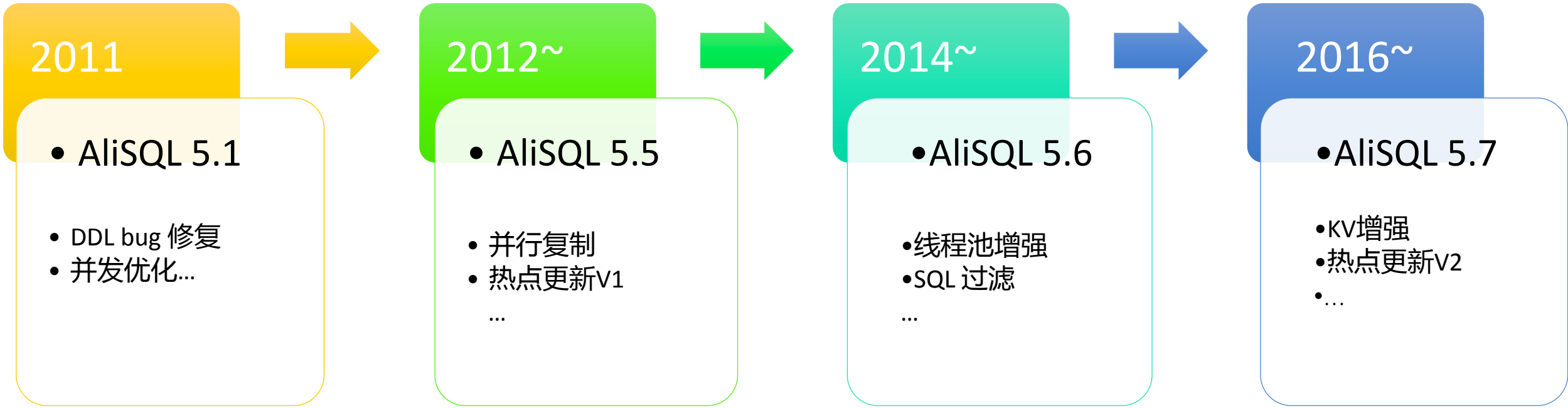
Alibaba Group
阿里巴巴集团

AliSQL

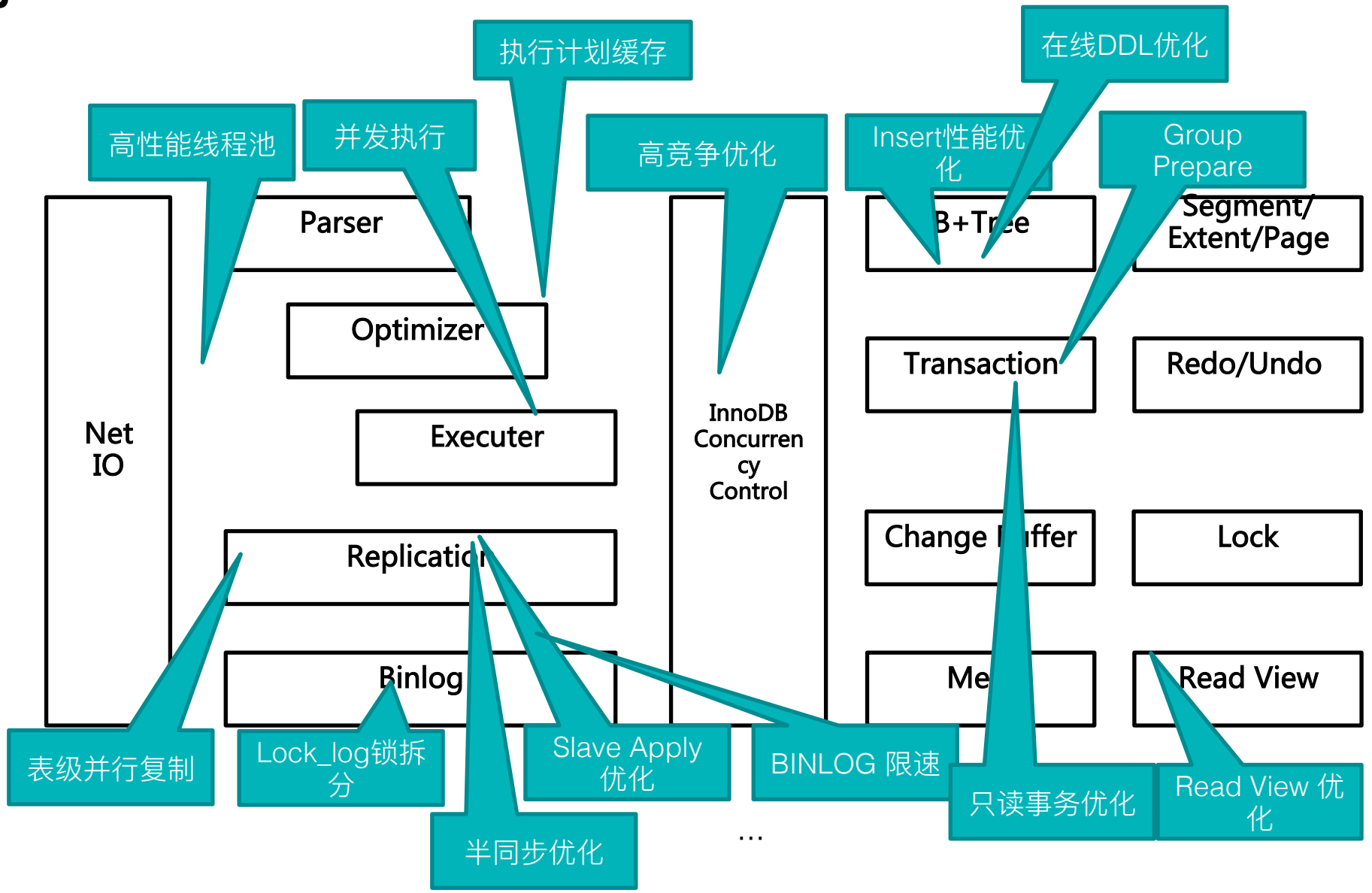
AliSQL的分布式架构

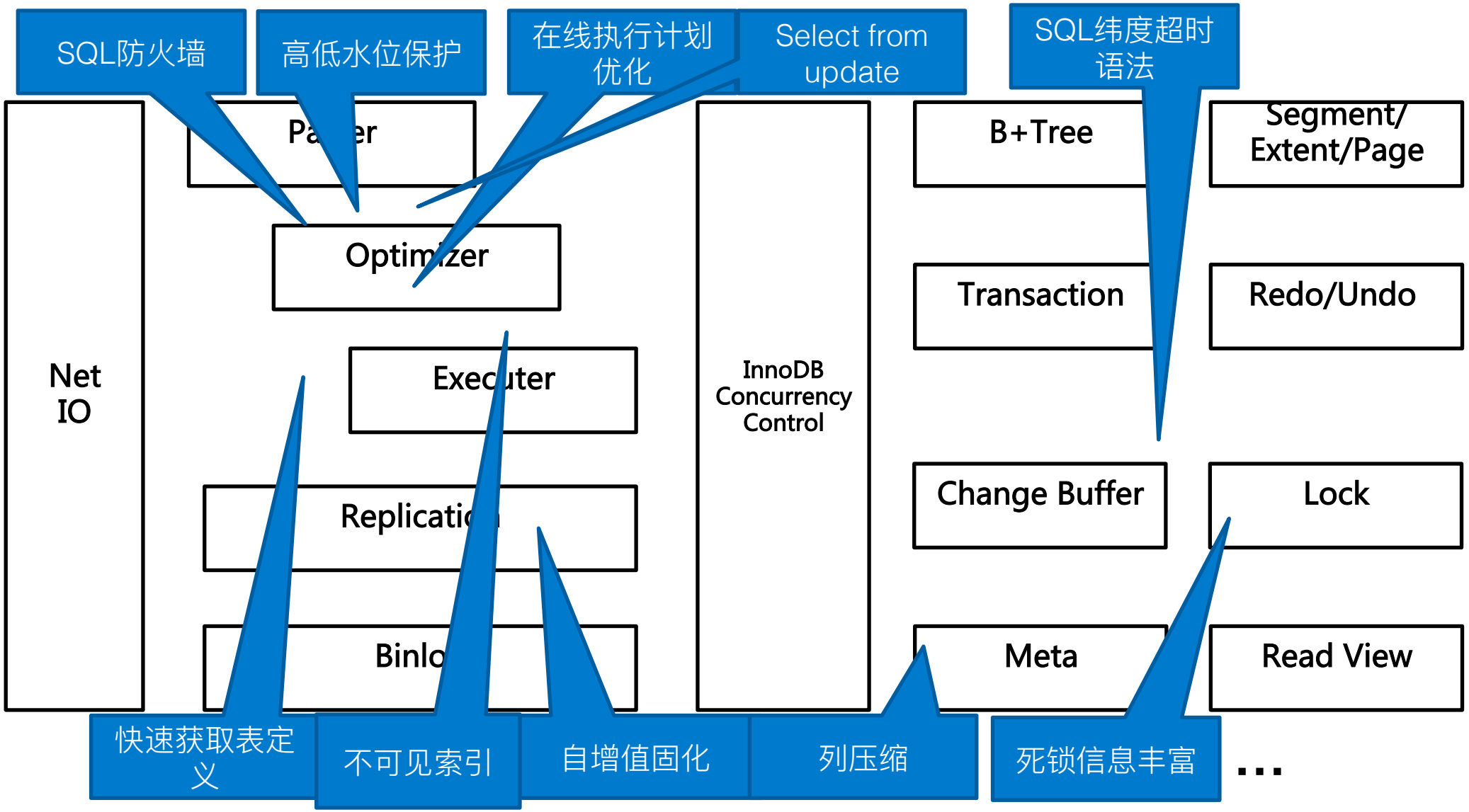
✓ 结合分布式数据库中间件TDDL的分库分表架构做到对业务透明

✓ 吸收开源生态的精华， 充分结合业务，获得数据库层面的自主掌控力



性能优化


执行计划缓存
在线DDL优化
高性能线程池
并发执行
高竞争优化
Insert性能优化
Group Prepare
Parser
B+Tree
Segment/ Extent/Page
Optimizer
Transaction
Redo/Undo
Net IO
Executer
InnoDB Concurrency Control
Change Buffer
Lock
Replication
Binlog
Me
Read View
表级并行复制
Lock_log锁拆分
Slave Apply 优化
BINLOG 限速
只读事务优化
Read View 优化
半同步优化
…


SQL防火墙
高低水位保护
在线执行计划优化
Select from update
SQL纬度超时语法
Net IO
Optimizer
Executer
Replicati
Binlo
InnoDB Concurrency Control
B+Tree
Segment/ Extent/Page
Transaction
Redo/Undo
Change Buffer
Lock
Meta
Read View
快速获取表定义
不可见索引
自增值固化
列压缩
死锁信息丰富
...

热点更新的问题

✓ 第一代的优化（热点排队）　　5000 热点tps

- Innodb Strict Concurrency
- Commit On Success
- Select From Update

✓ 第二代的优化（热点合并 ）　　100000 热点tps

- Group Update
- New Innodb Row Lock Type
- Row Cache

经典的秒杀事务模型:

① begin;
② insert normal row;
③ update hot row;
④ select hot row;
⑤ commit;

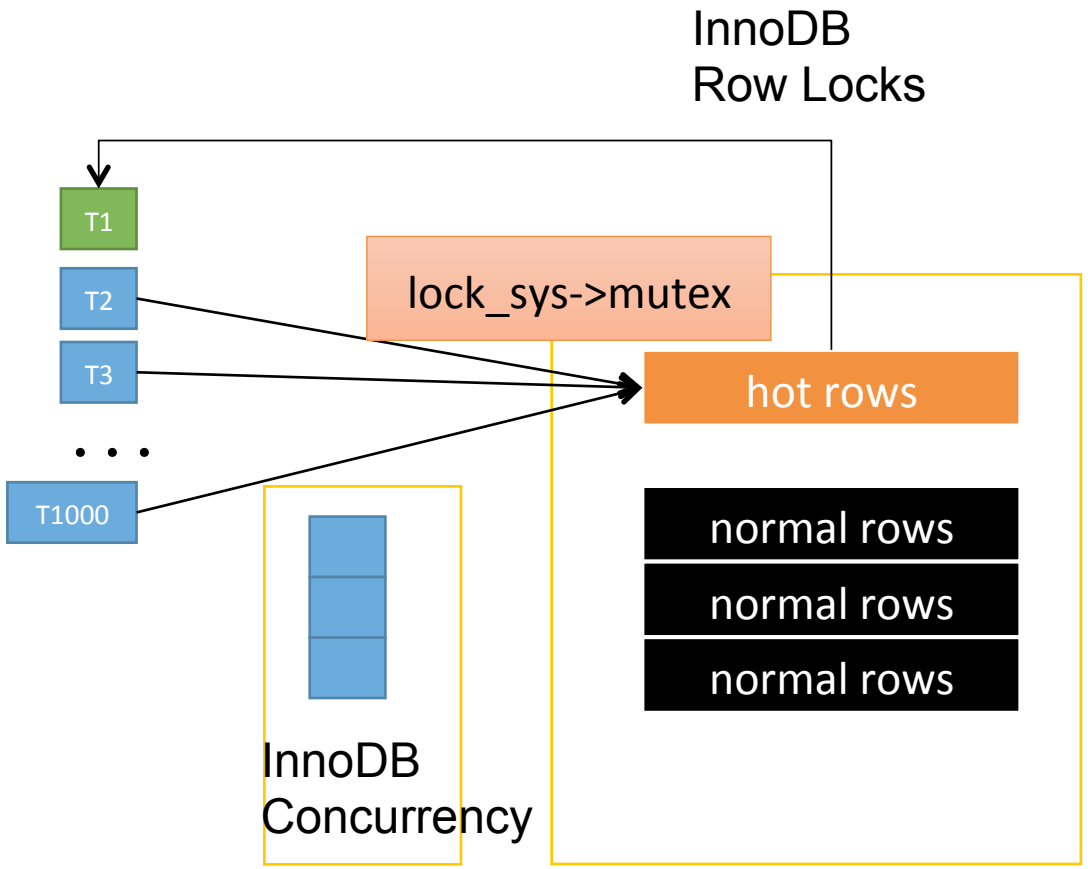
热点优化排队

✓ 原生MySQL热点更新低效的原因

- Innodb 中大量的锁等待
- 死锁检测占用大量CPU

✓ 核心思路

- 通过排队减少Innodb中并发更新热点事务的数量
- 改造Innodb Concurrency的实现
- 通过select from update以及commit on success 减少事务持锁时间



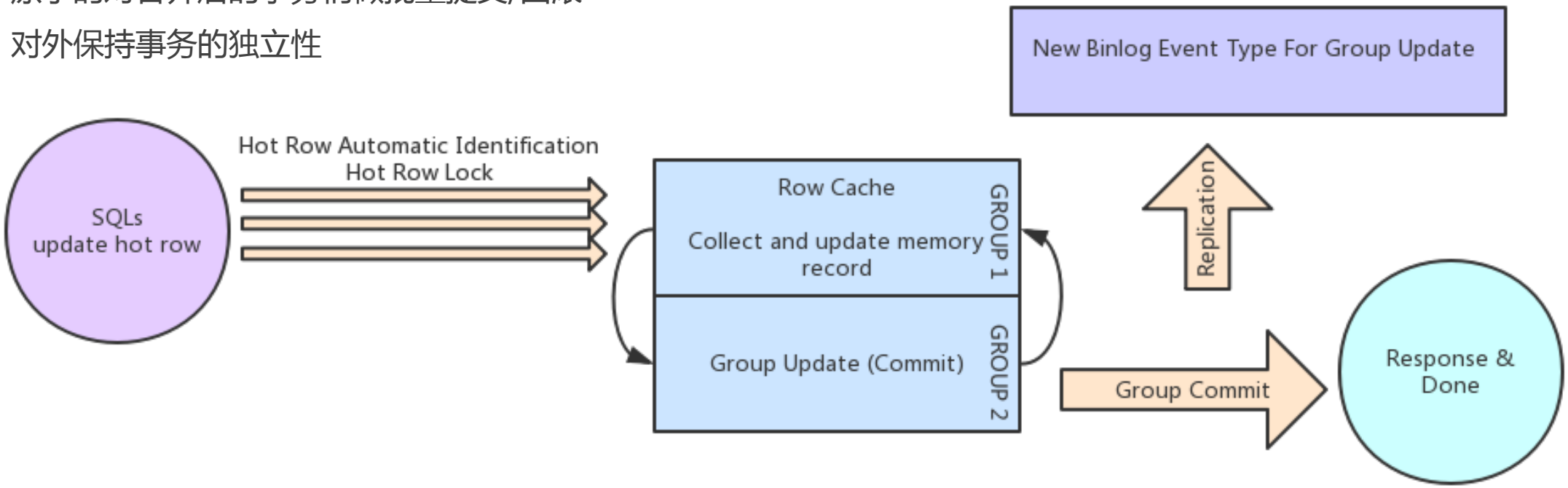
热点优化合并

✓ 动机以及挑战：

- 排队上的优化已无法支撑这两年双11的秒杀活动量
- 排队优化的上限即是单线程更新单行，想再有量级的突破很困难

✓ 解决思路

- 将多次的更新在数据库内部动态合并
- 原子的对合并后的事务们做批量提交/回滚
- 对外保持事务的独立性

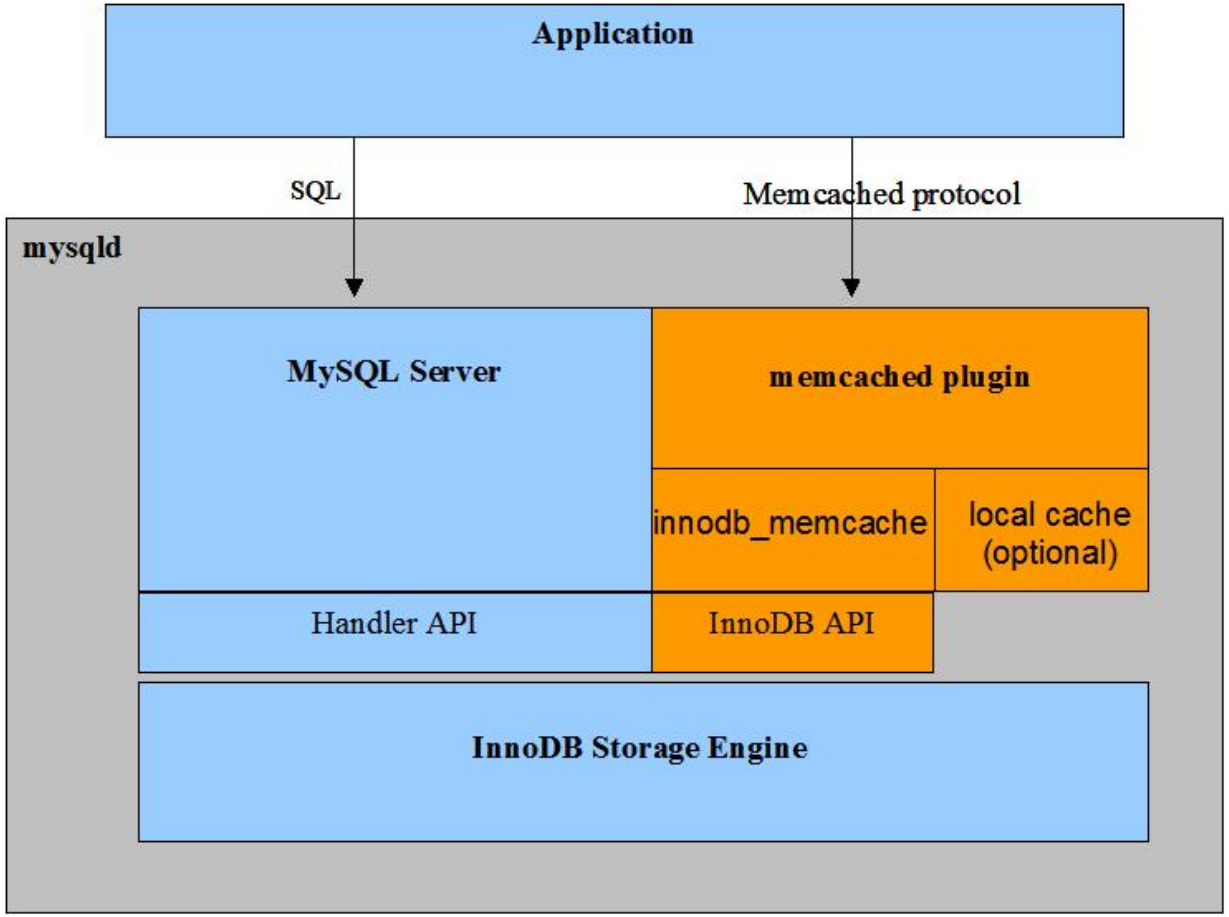


Alibaba Group
阿里巴巴集团

X-KV

回顾Memcached-plugin

- ✓ 高性能KV接口
- ✓ 省去SQL Parser/optimizer
- ✓ 数据一致性

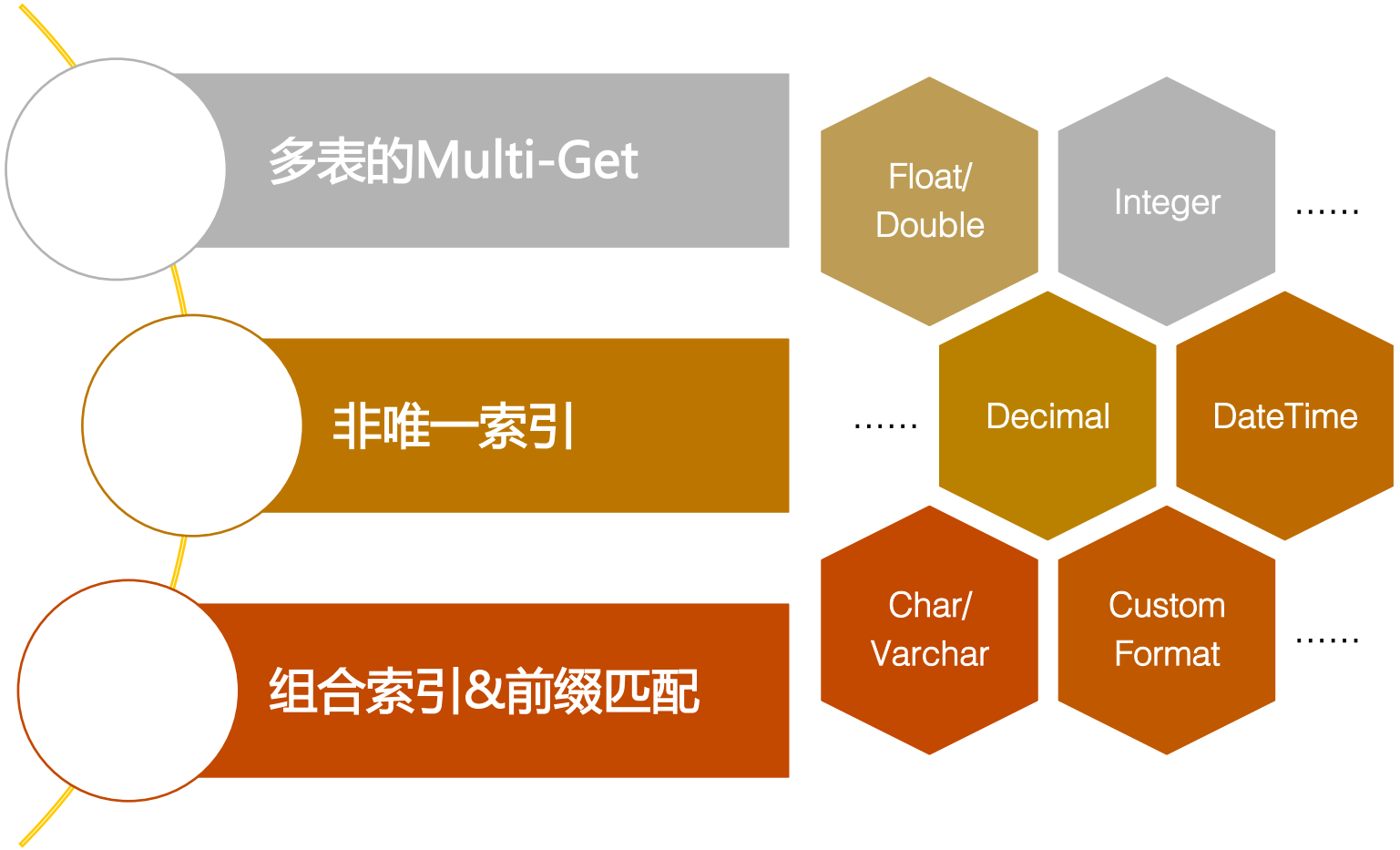


X-KV的概述

✓AliSQL高性能K-V接口，InnoDB Memcached Plugin的扩展

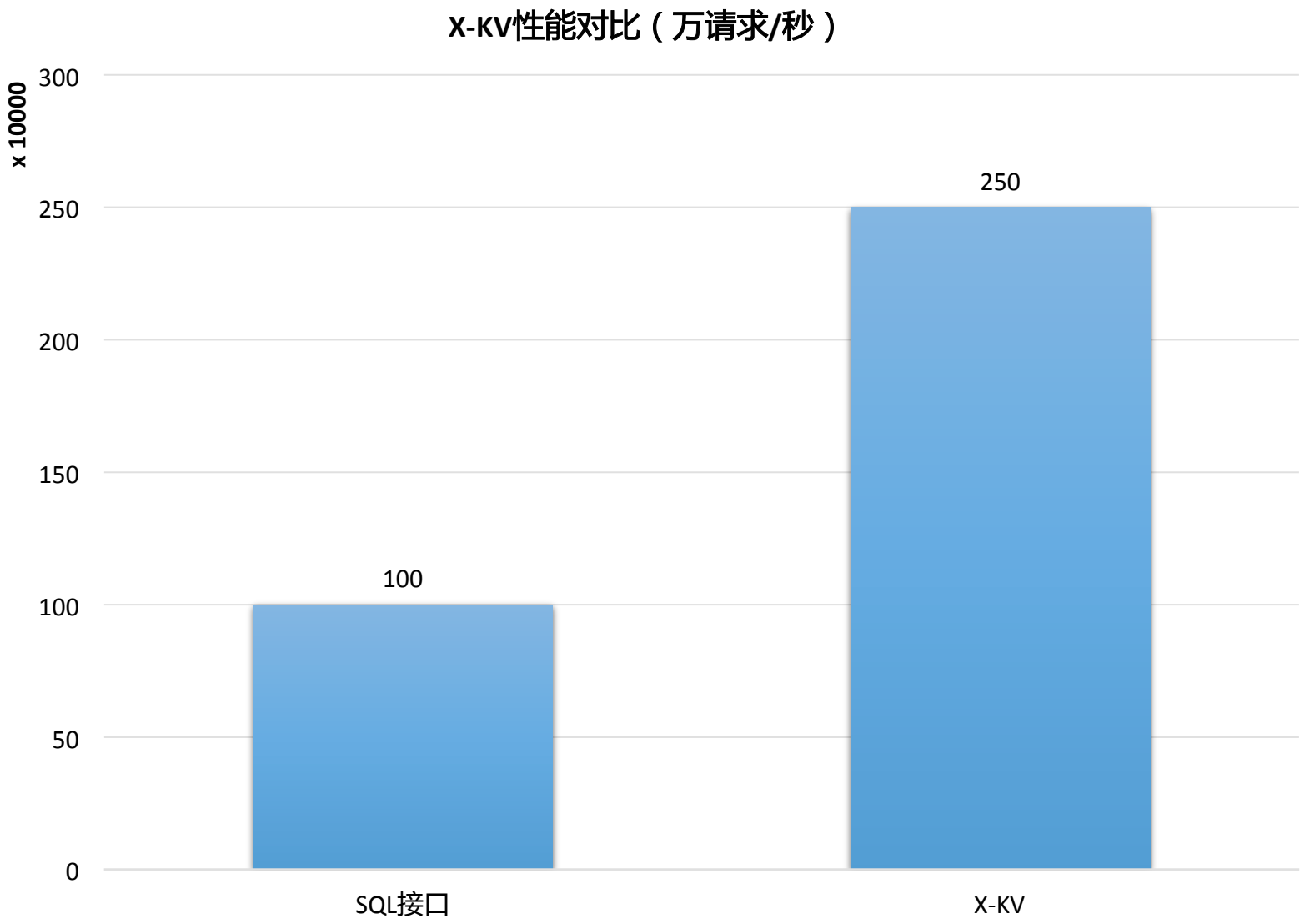
功能增强	高可运维性	极致性能
• 丰富数据类型 • 范围查询支持 • 新协议支持	• 在线配置修改生效 • 在线DDL自动加载	• 重写核心转换函数

X-KV功能增强


多表的Multi-Get
非唯一索引
组合索引&前缀匹配
Float/
Double
Integer
.....
.....
Decimal
DateTime
Char/
Varchar
Custom
Format
.....

性能表现

- ✓ 32 core
- ✓ Sysbench 只读场景



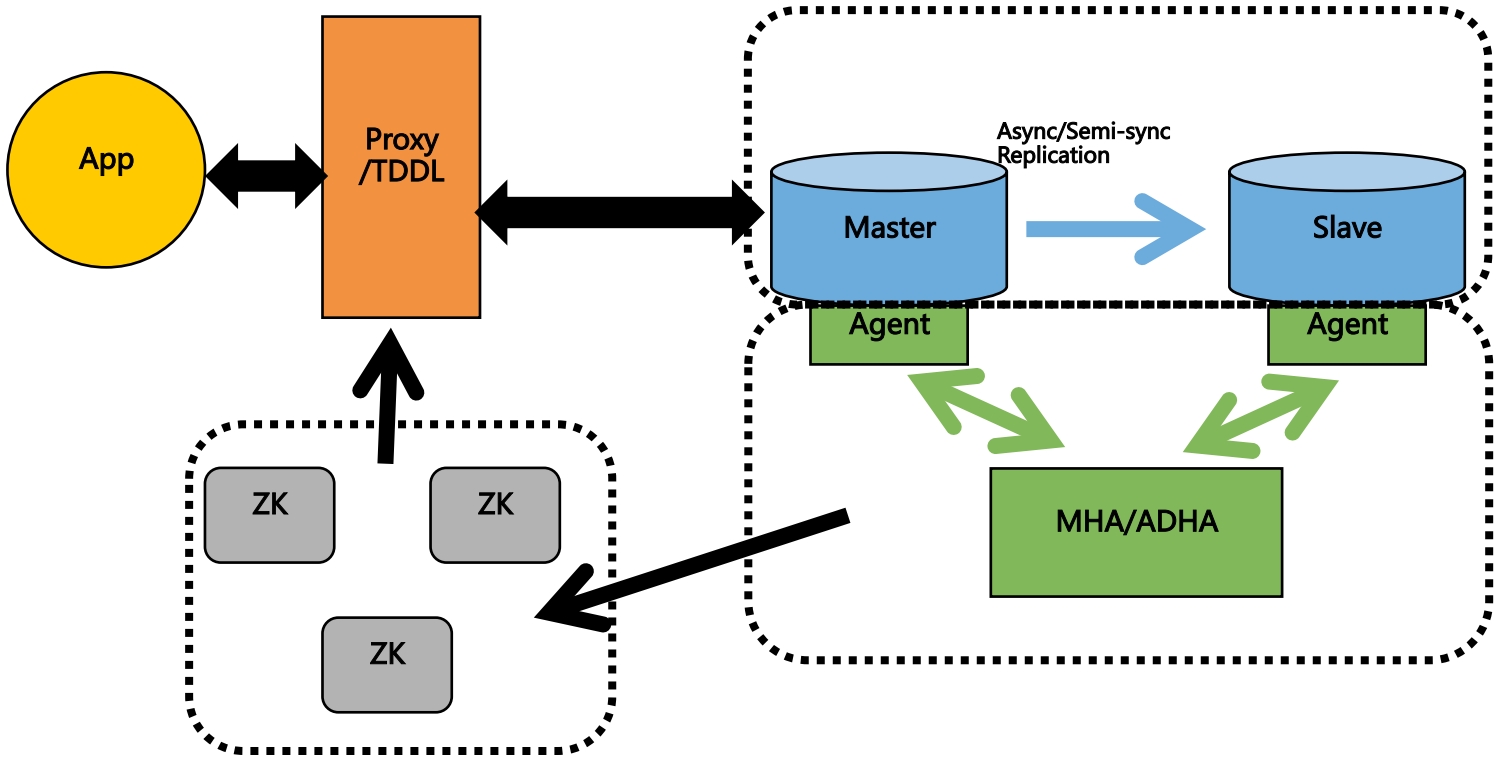
Alibaba Group
阿里巴巴集团

X-DB

经典的MySQL主备架构

✓不足

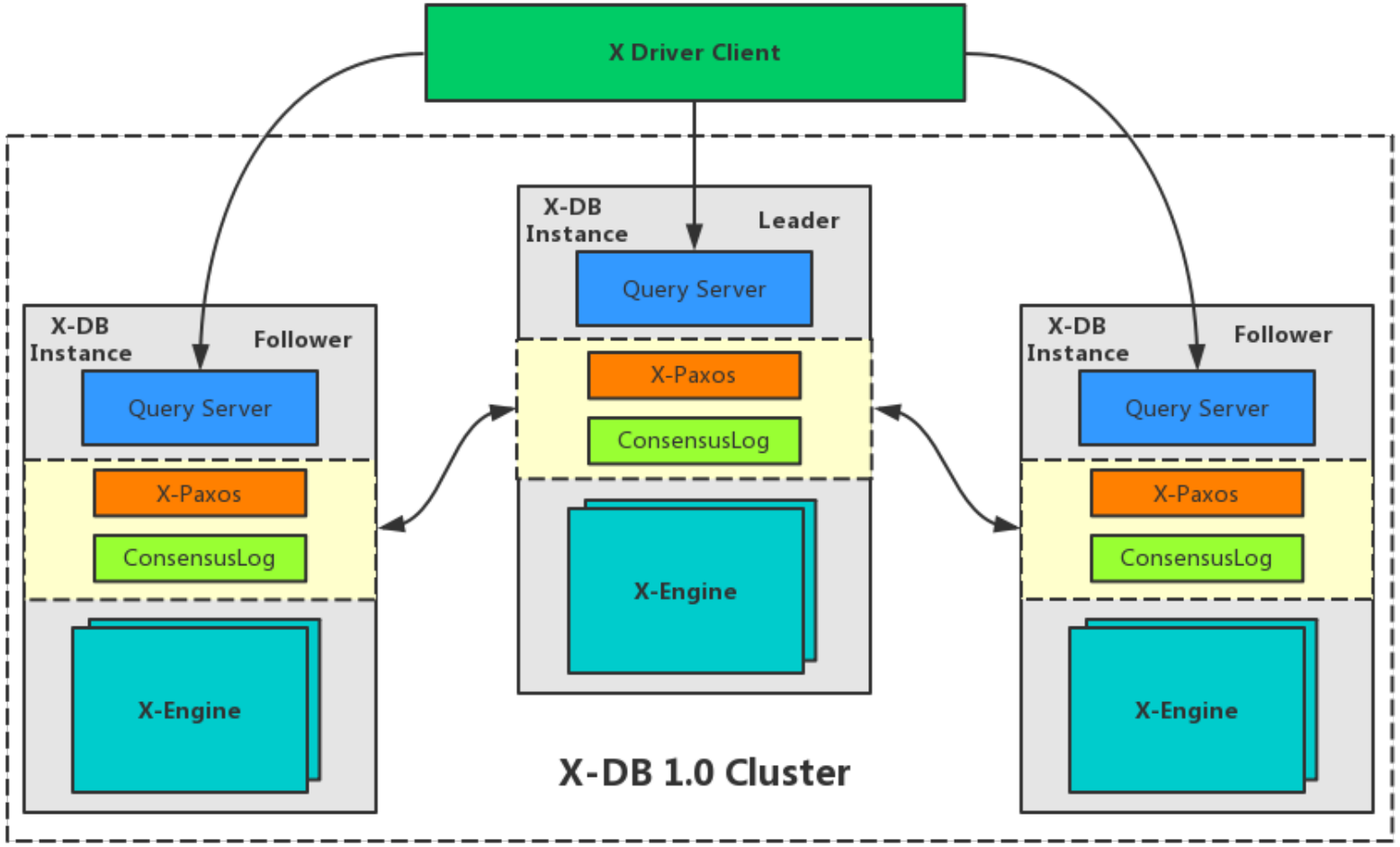
- 多系统耦合
- 强一致
- 持续可用性
- 跨Region的性能
- 可运维性



X-DB 1.0 架构

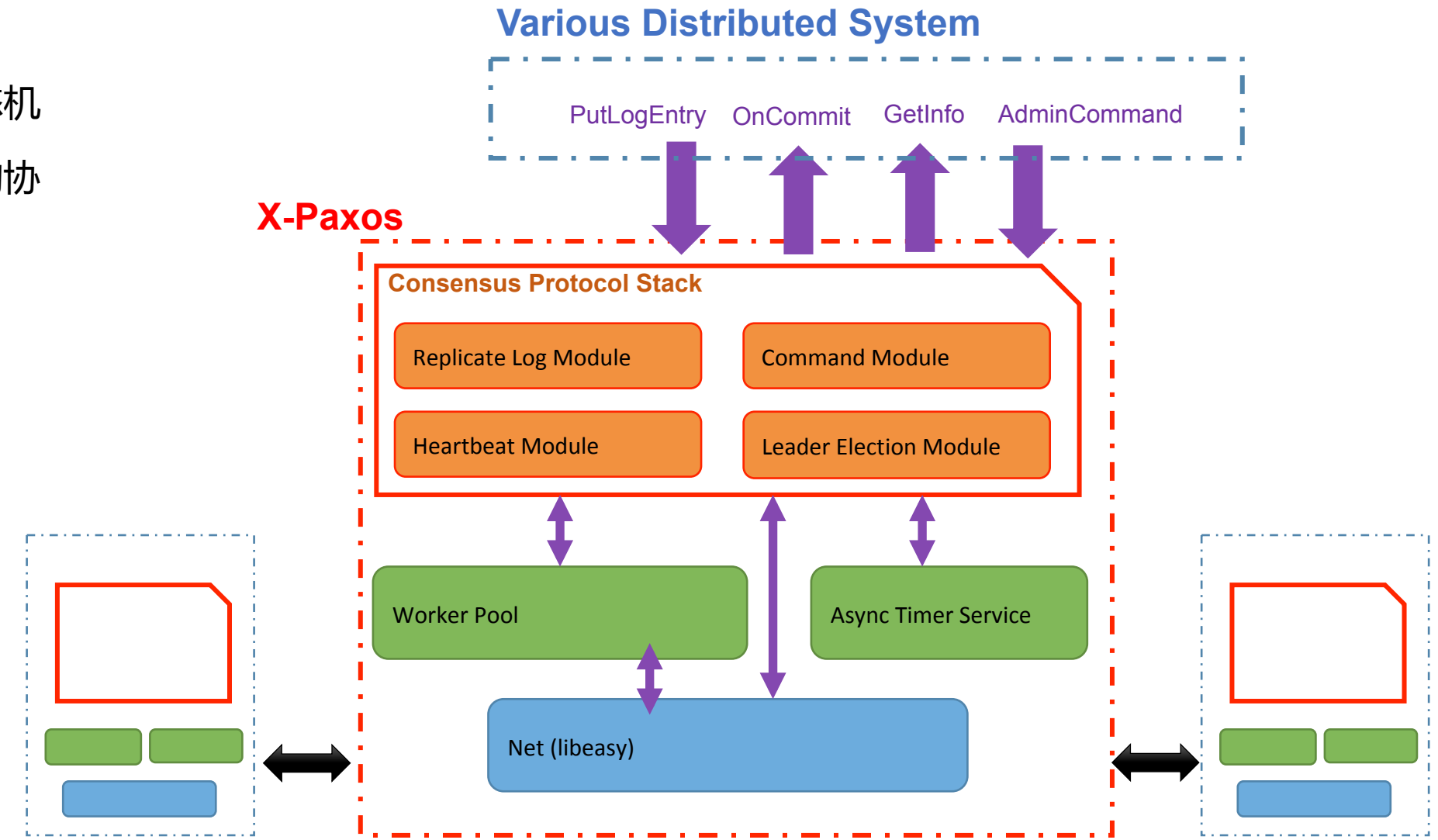
✓ X-DB 1.0 架构

- 一体化架构，运维友好
- 高性能，三副本最高相对于单机10%的损耗
- 可跨region部署，保持高吞吐
- 稳定性：网络抖动高容忍
- 兼容性，业务无需改造



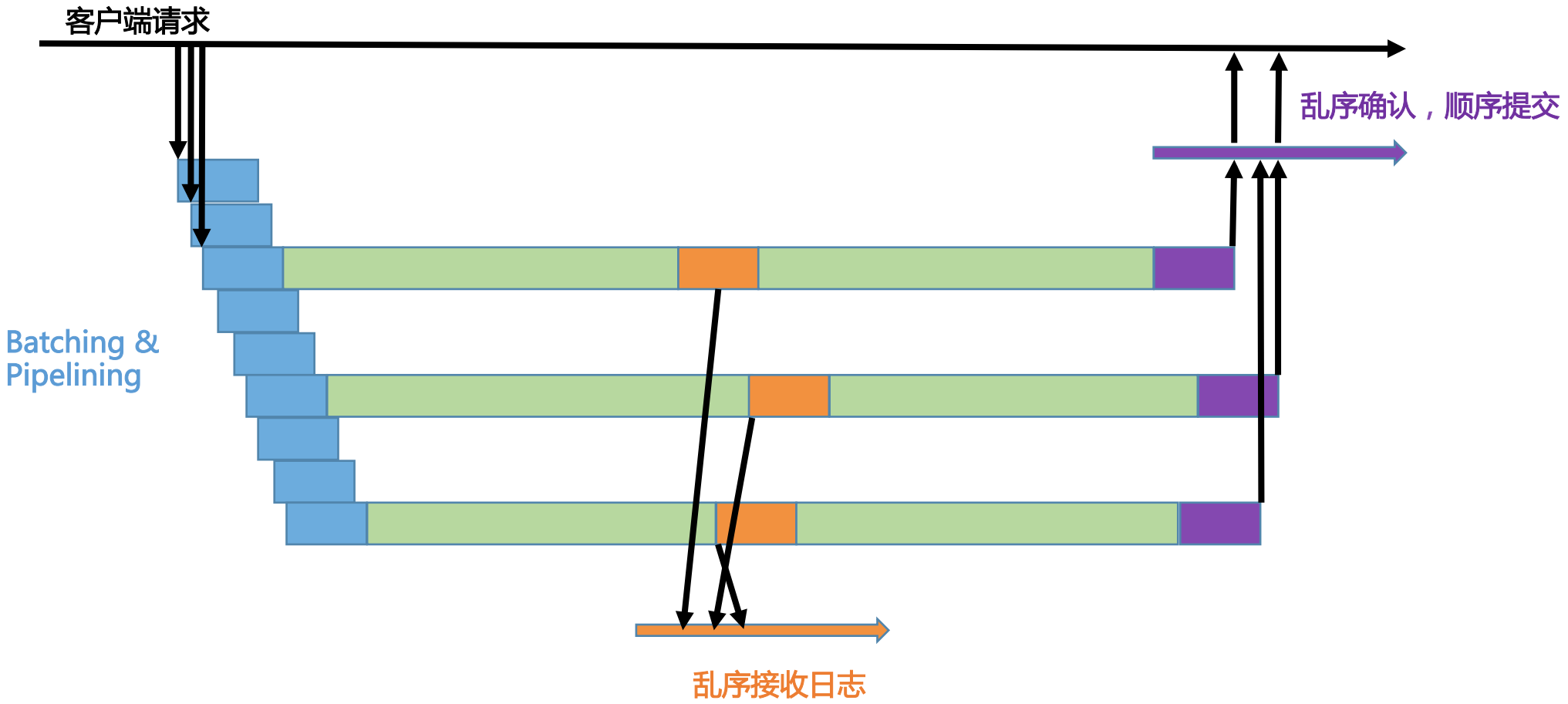
高性能一致性协议库X-Paxos

- ✓ 高性能网络库libeasy
- ✓ 插件式的日志及状态机
- ✓ 经过生产环境考验的协议库

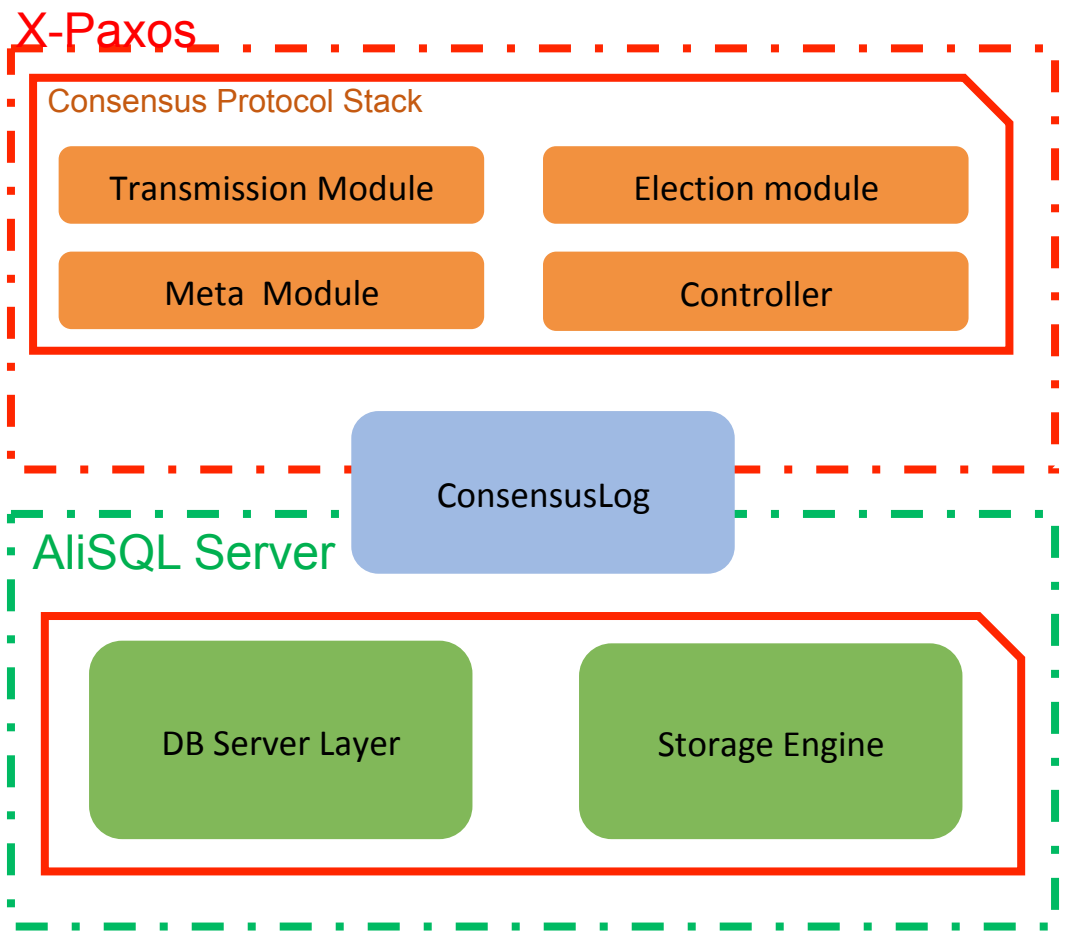


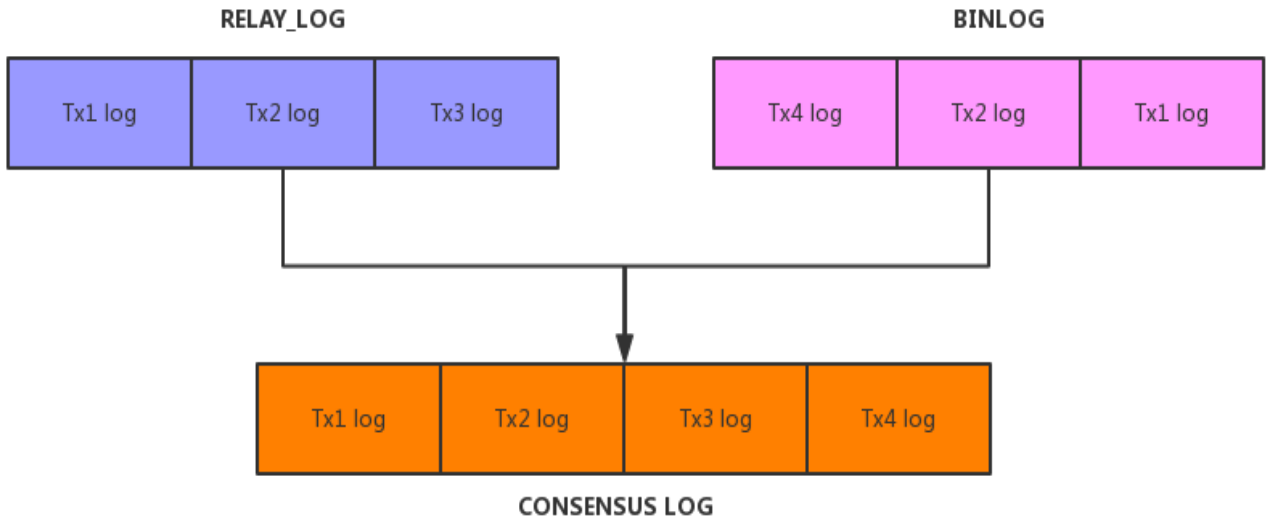
Batching&Pipelining优化

✓ 跨Region 长传网络下有效优化吞吐的手段



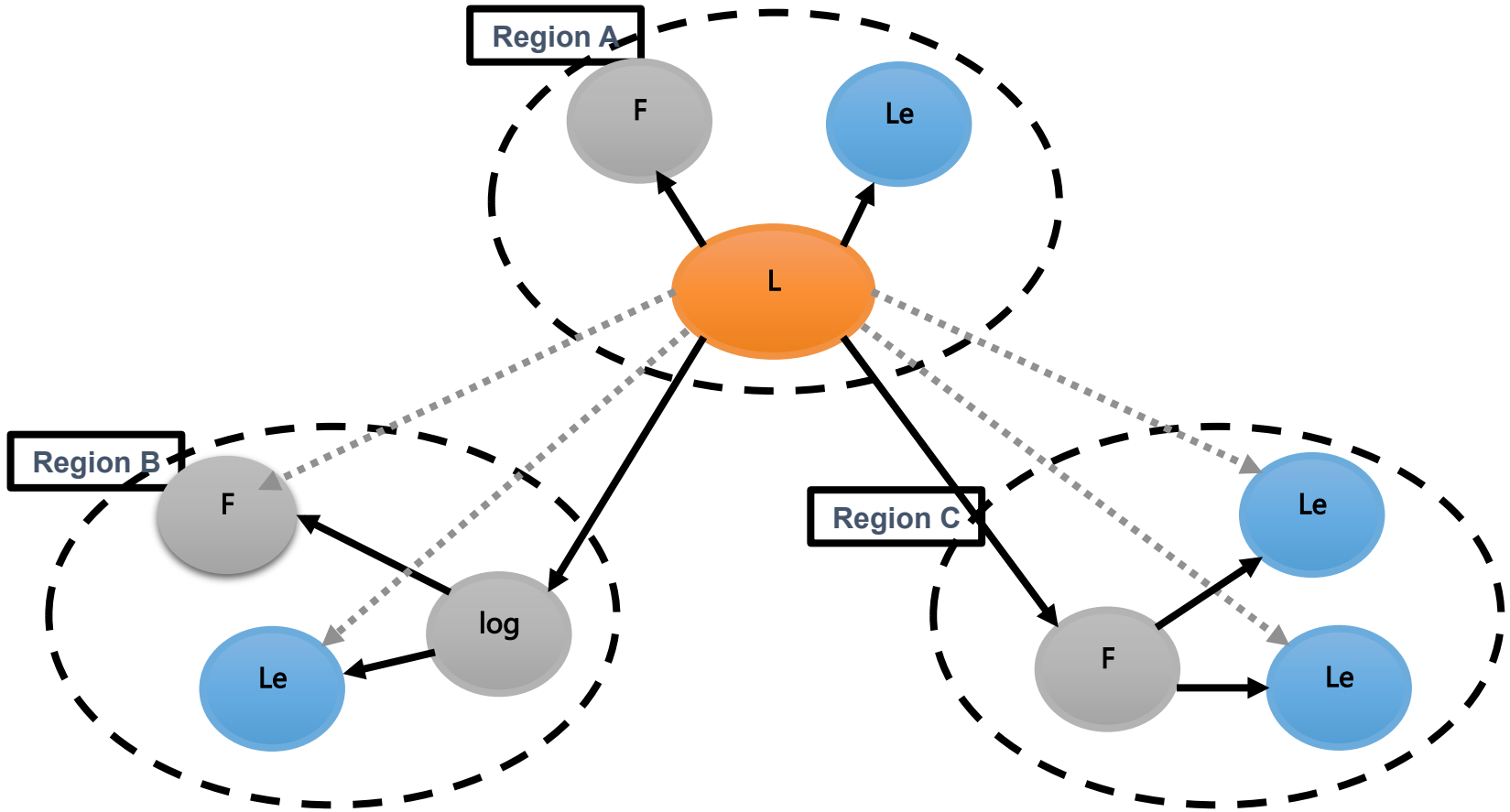
X-DB 1.0 的日志


X-Paxos
Consensus Protocol Stack
Transmission Module
Election module
Meta Module
Controller
ConsensusLog
AliSQL Server
DB Server Layer
Storage Engine


RELAY_LOG
BINLOG
Tx1 log
Tx2 log
Tx3 log
Tx4 log
Tx2 log
Tx1 log
Tx1 log
Tx2 log
Tx3 log
Tx4 log
CONSENSUS LOG

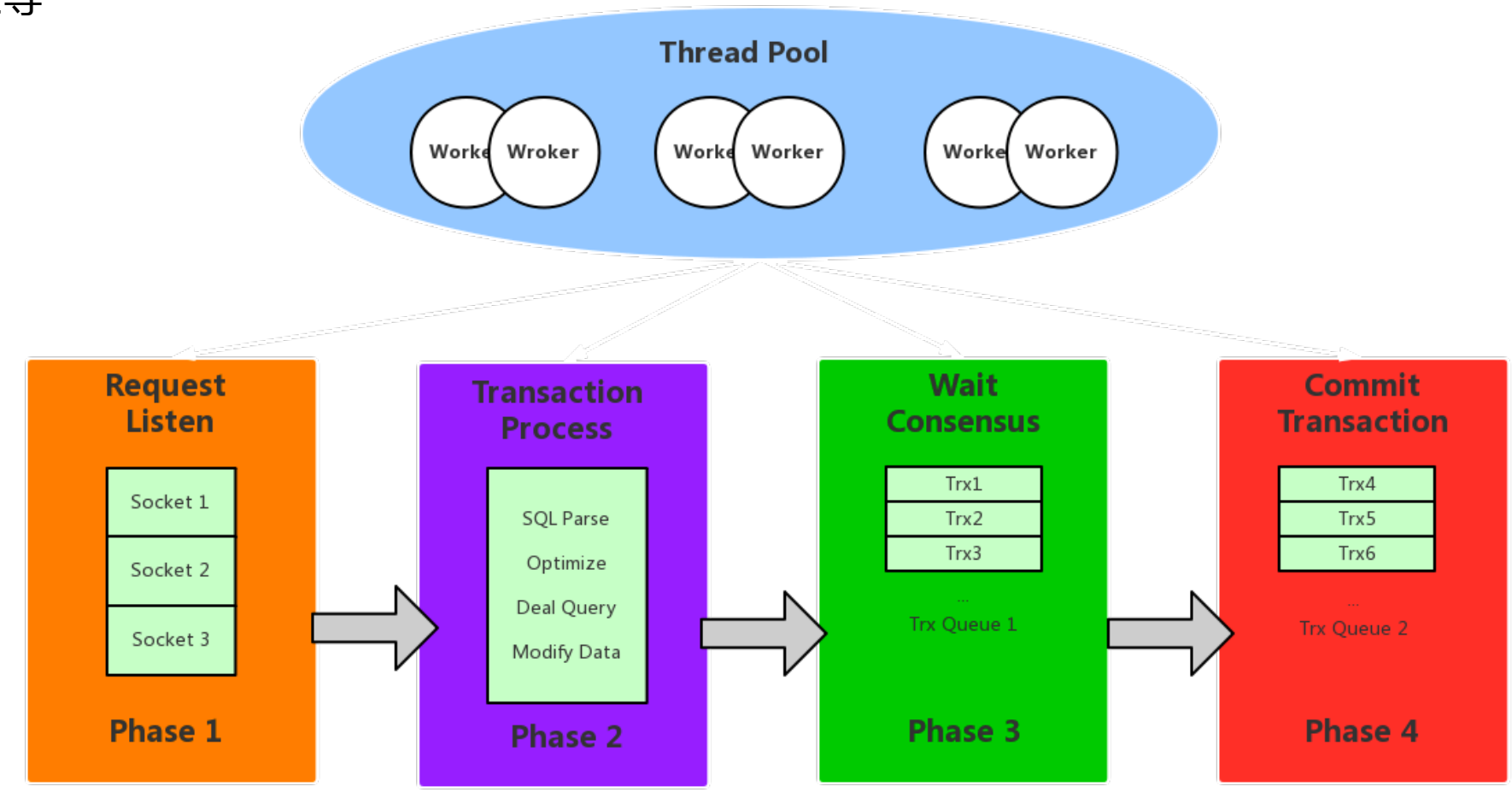
灵活的角色和部署方式

- ✓ 降低单实例网络瓶颈
- ✓ 降低广域网网络带宽压力
- ✓ 可选的压缩传输
- ✓ 可选是否带状态机



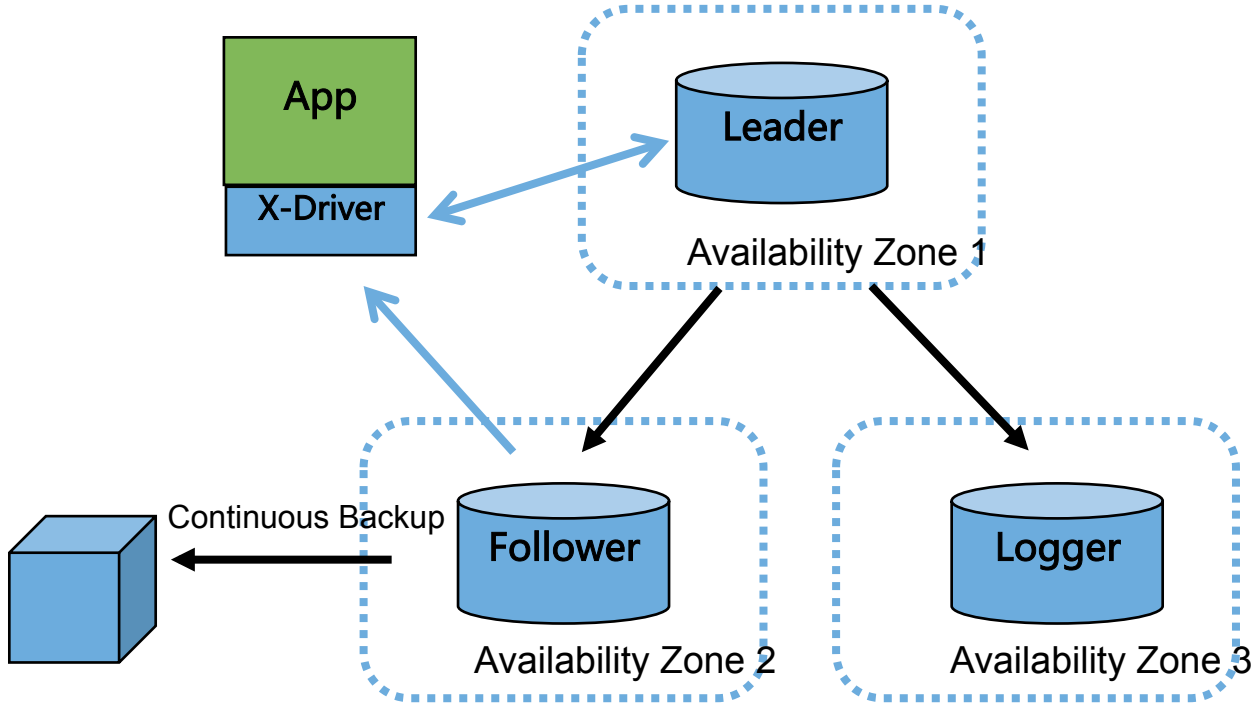
异步事务提交

- ✓ 提高CPU利用率，降低空等
- ✓ 单线程处理多事务请求



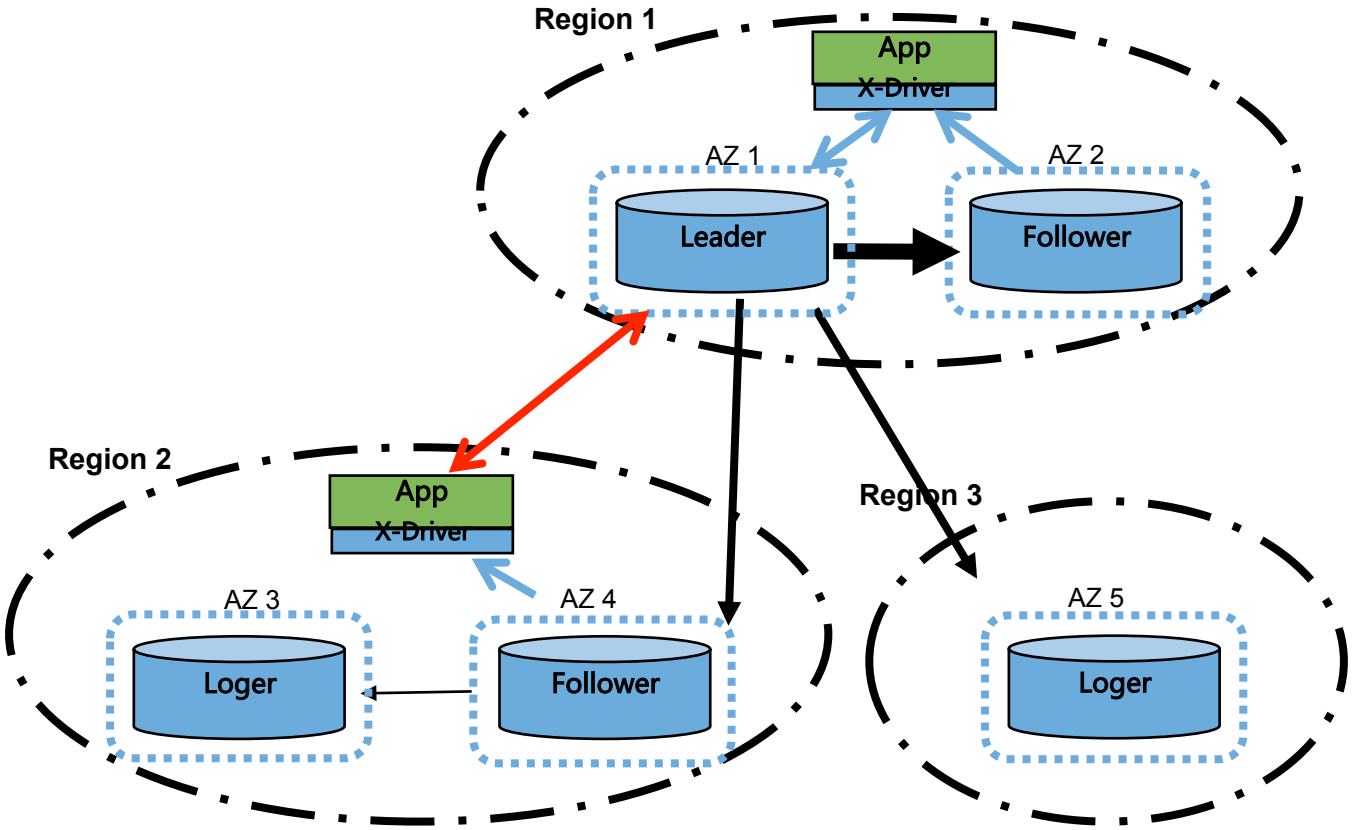
X-DB 典型部署架构 1

- ✓ AZ级别无损容灾
- ✓ 相对主备只多存一份Consensus 日志
- ✓ 持续备份 RPO < 1s



X-DB 典型部署架构2

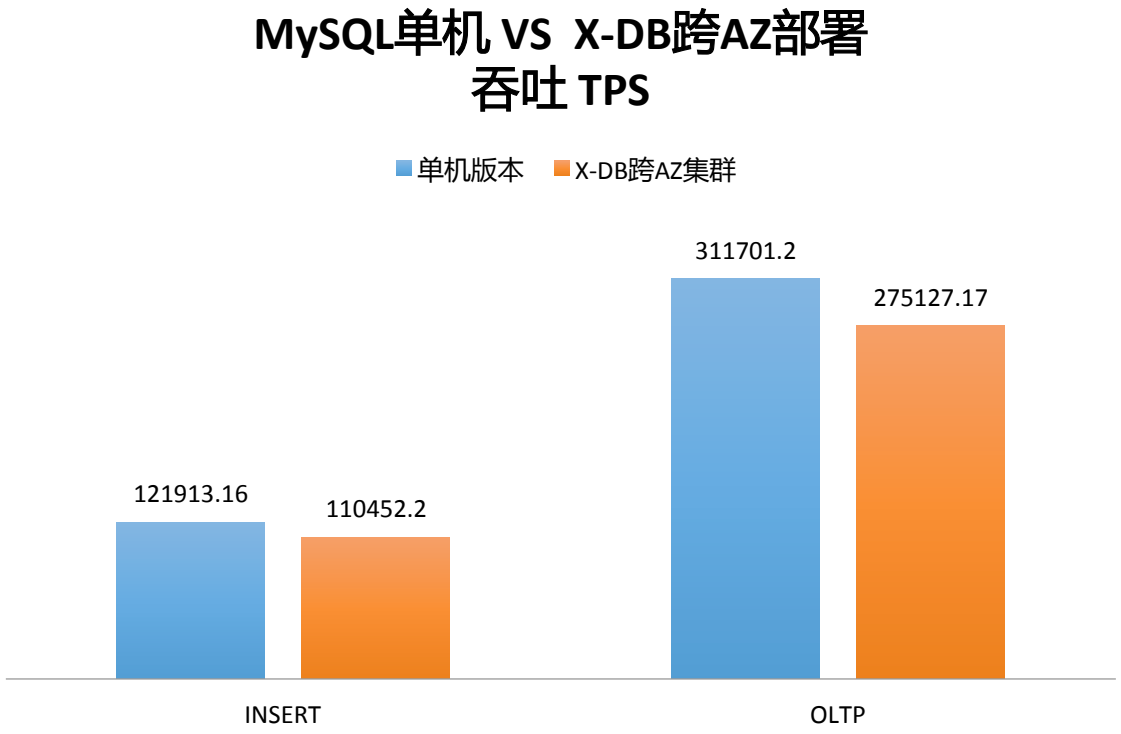
- ✓ Region级无损容灾
- ✓ 跨Region强同步下，依然保持较高吞吐
- ✓ 优先级选主
- ✓ 可与多AZ架构之间无缝切换

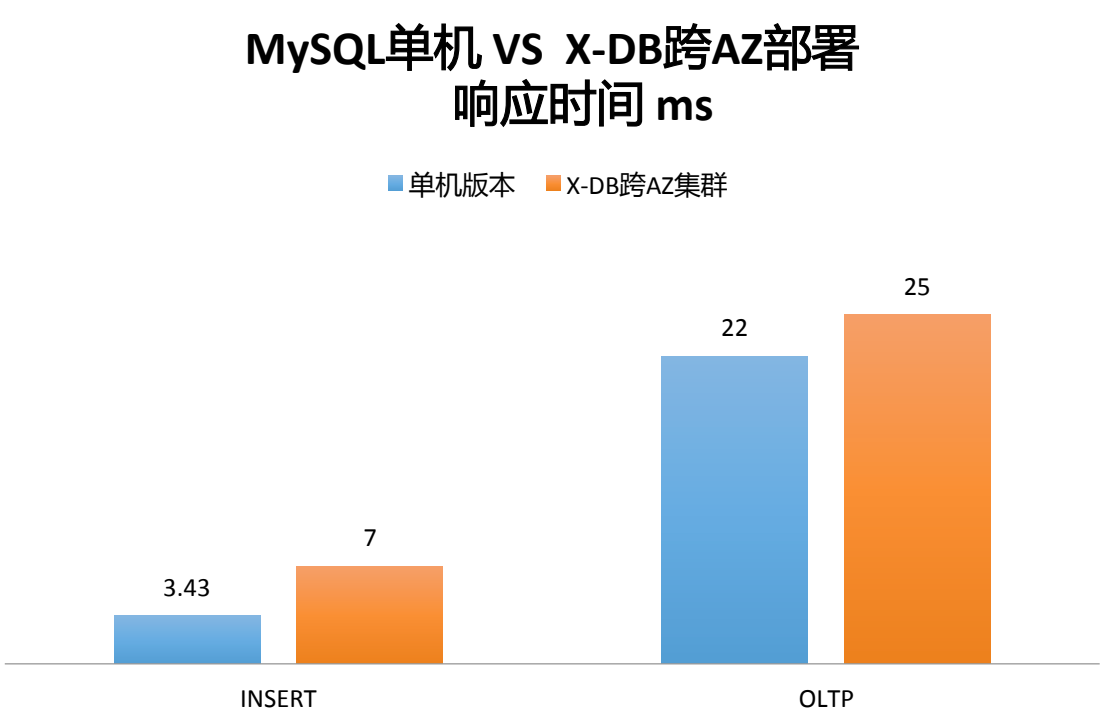


X-DB性能表现

✓ 64 core / ssd

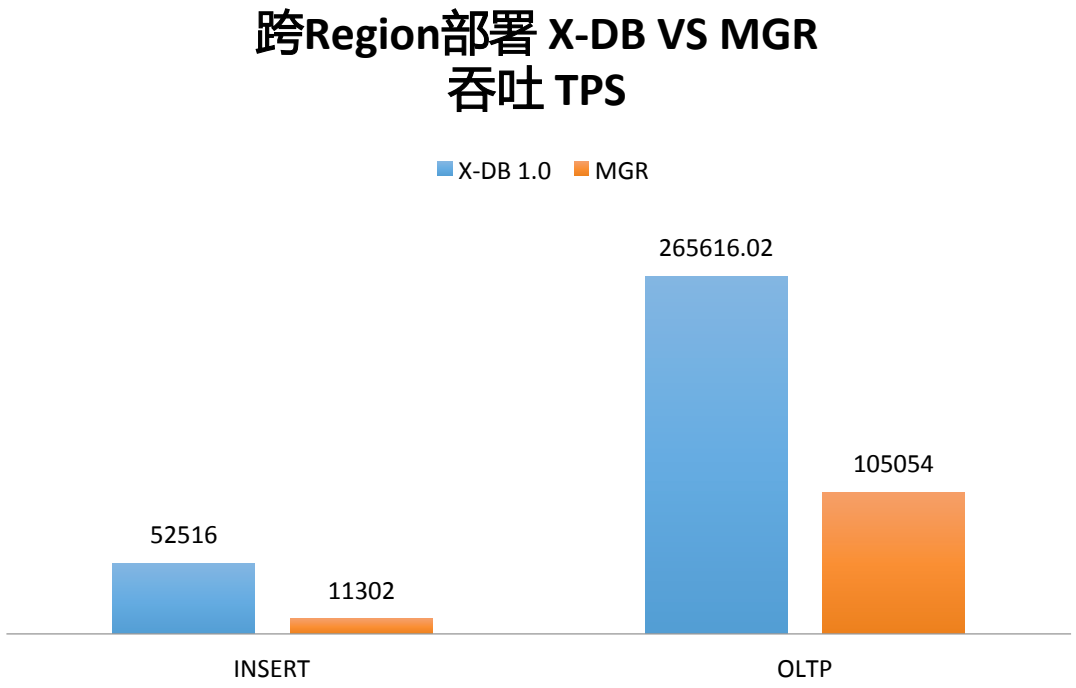
✓ 单机 VS 跨AZ部署形态

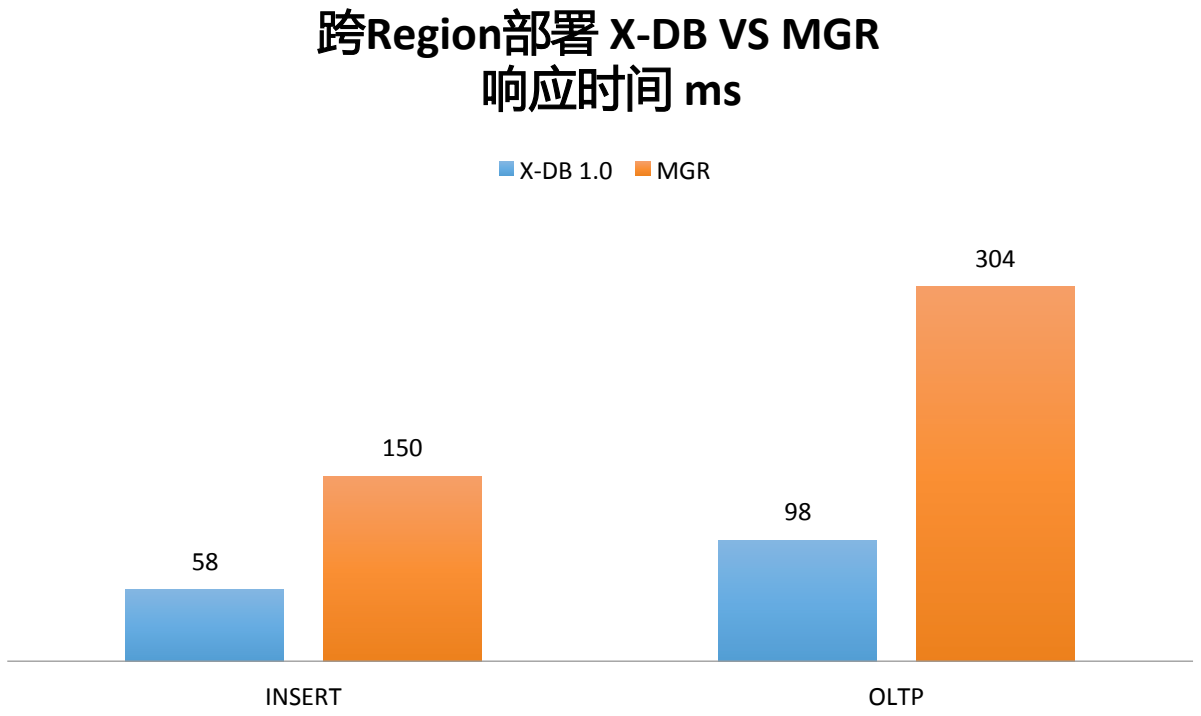




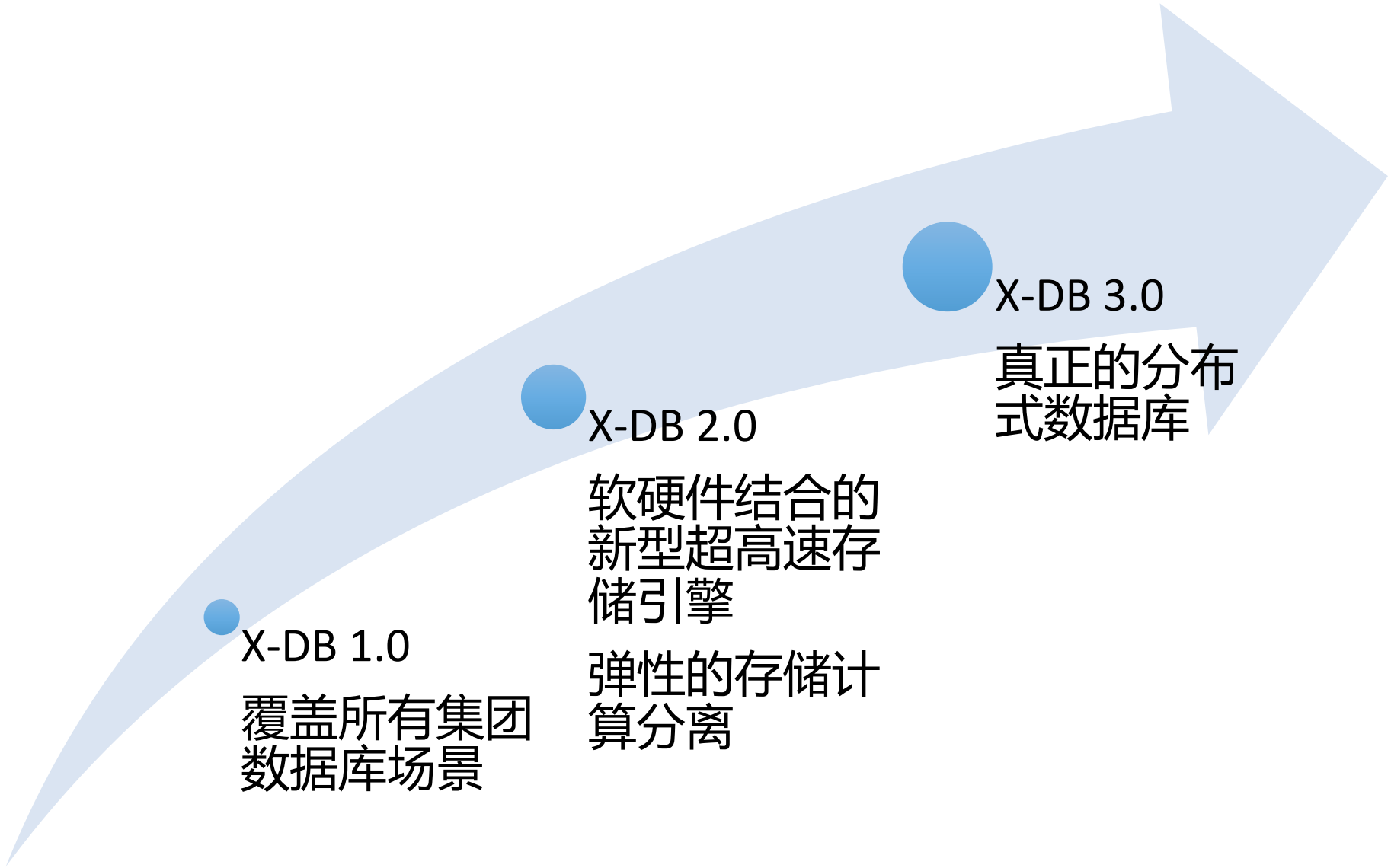
X-DB性能表现

- ✓ 64core / ssd
- ✓ 跨Region部署下 X-DB VS MGR
- ✓ 网络rtt 约30ms





X-DB的Roadmap



Alibaba Group 阿里巴巴集团

THANKS/ 欢迎加入阿里数据库事业部

--------- Q&A Section --------