DBAplus

数据库年终盘点大会-上海站

TDSQL在腾讯的研发实践

潘安群@腾讯

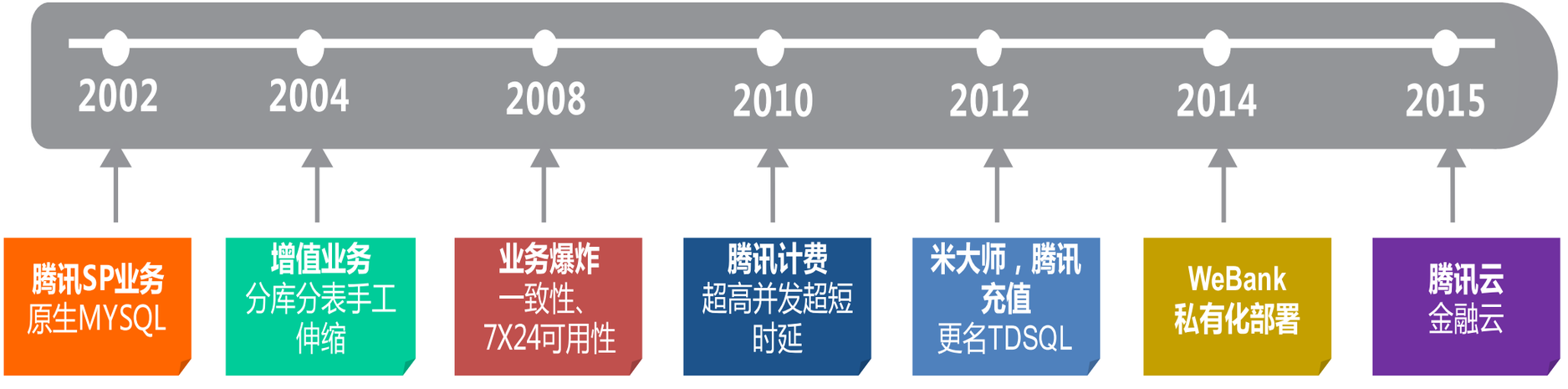
潘安群

腾讯TEG计费平台部技术总监

腾讯云金融分布式数据库TDSQL研发负责人

超10年分布式系统及数据库研发经验

TDSQL是腾讯提供的一套兼容MySQL的金融级分布式数据产品，定位为OLTP分布式数据库。



浦发银行 信用卡 SPD BANK Credit Card　WeBank　OneBank 华通银行　金谷农商银行 JINGU RURAL COMMERCIAL BANK　和泰人寿保险股份有限公司 HE TAI LIFE INSURANCE CO.,LTD　Aixin Life 爱心人寿　三一重工　米大师 腾讯移动支付Midas　Answern 安心保险　众惠相互 Public Mutual

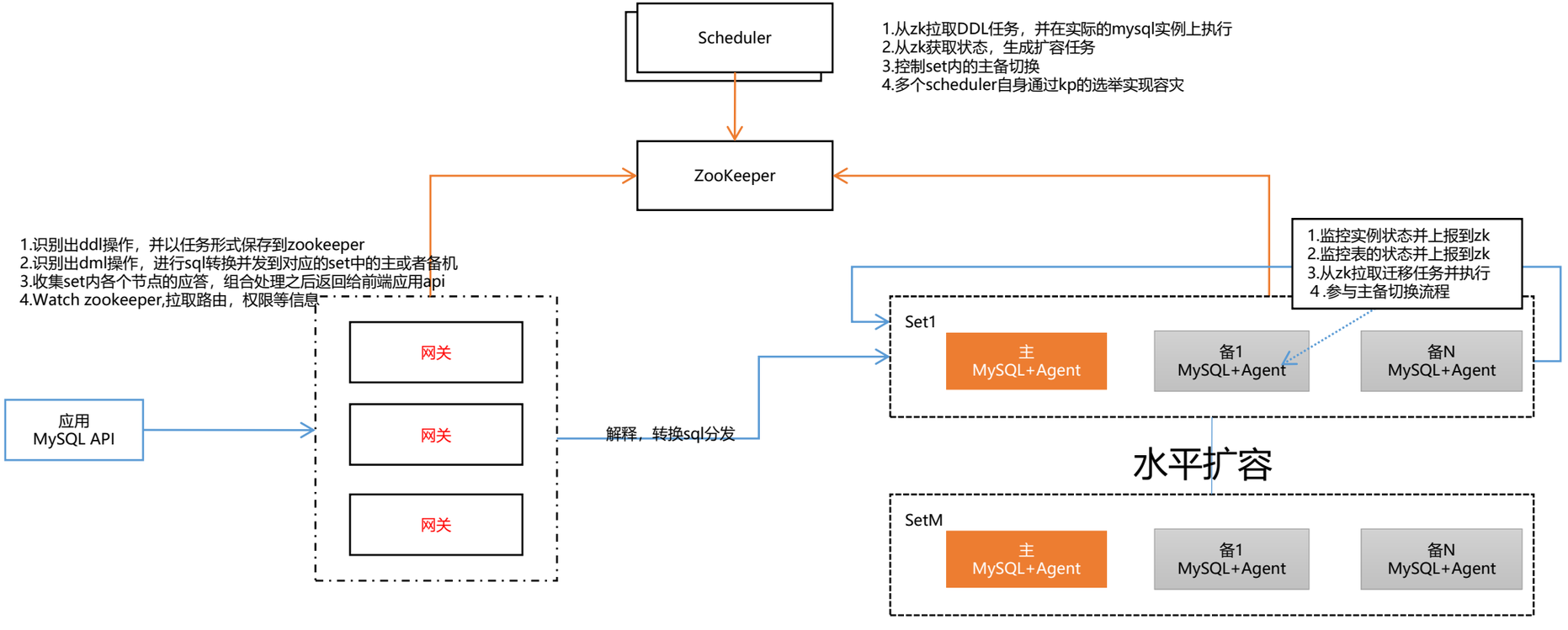
天津金城银行　荷包金融　泛华保险服务 CNINSURE INC.　Smart City 前海通　威富通 SWIFTPASS TECHNOLOGIES　高毅资产 高屋建瓴 志当恒毅　理财通　G7 汇通天下　保网 baoxian.com 比较、省钱、轻松投保

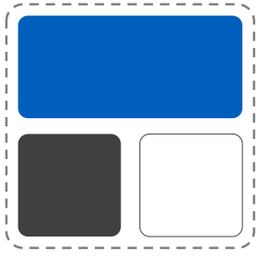
1. 核心特性
2. 分布式实践
3. 部署实践

1 核心特性

高一致性、高可用性、高性能

TDSQL核心架构


Scheduler
1.从zk拉取DDL任务，并在实际的mysql实例上执行
2.从zk获取状态，生成扩容任务
3.控制set内的主备切换
4.多个scheduler自身通过kp的选举实现容灾
ZooKeeper
1.识别出ddl操作，并以任务形式保存到zookeeper
2.识别出dml操作，进行sql转换并发到对应的set中的主或者备机
3.收集set内各个节点的应答，组合处理之后返回给前端应用api
4.Watch zookeeper,拉取路由，权限等信息
1.监控实例状态并上报到zk
2.监控表的状态并上报到zk
3.从zk拉取迁移任务并执行
4.参与主备切换流程
网关
网关
网关
应用
MySQL API
解释，转换sql分发
Set1
主
MySQL+Agent
备1
MySQL+Agent
备N
MySQL+Agent
水平扩容
SetM
主
MySQL+Agent
备1
MySQL+Agent
备N
MySQL+Agent

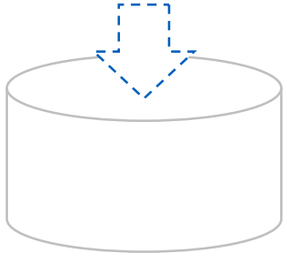


复制

Replica

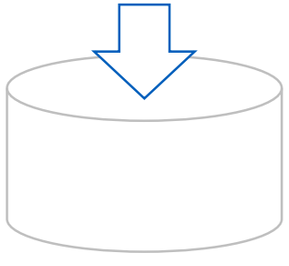
主备数据复制方式

DBAplus



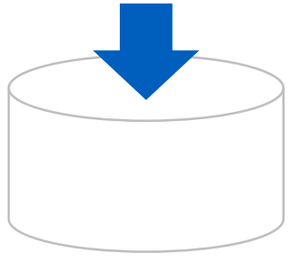
异步复制

Async replication



半同步复制

Semi-Sync replication

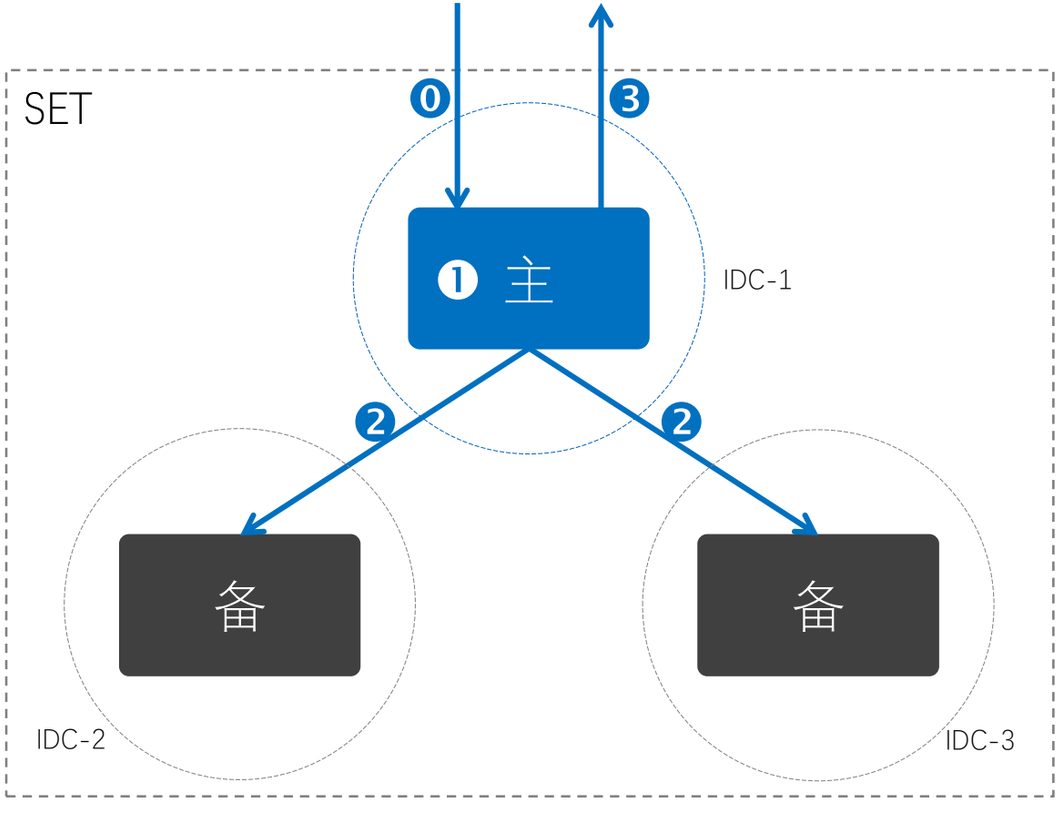


强同步复制

Sync replication

强同步更新流程

DBAplus



半同步复制的不足

DBAplus

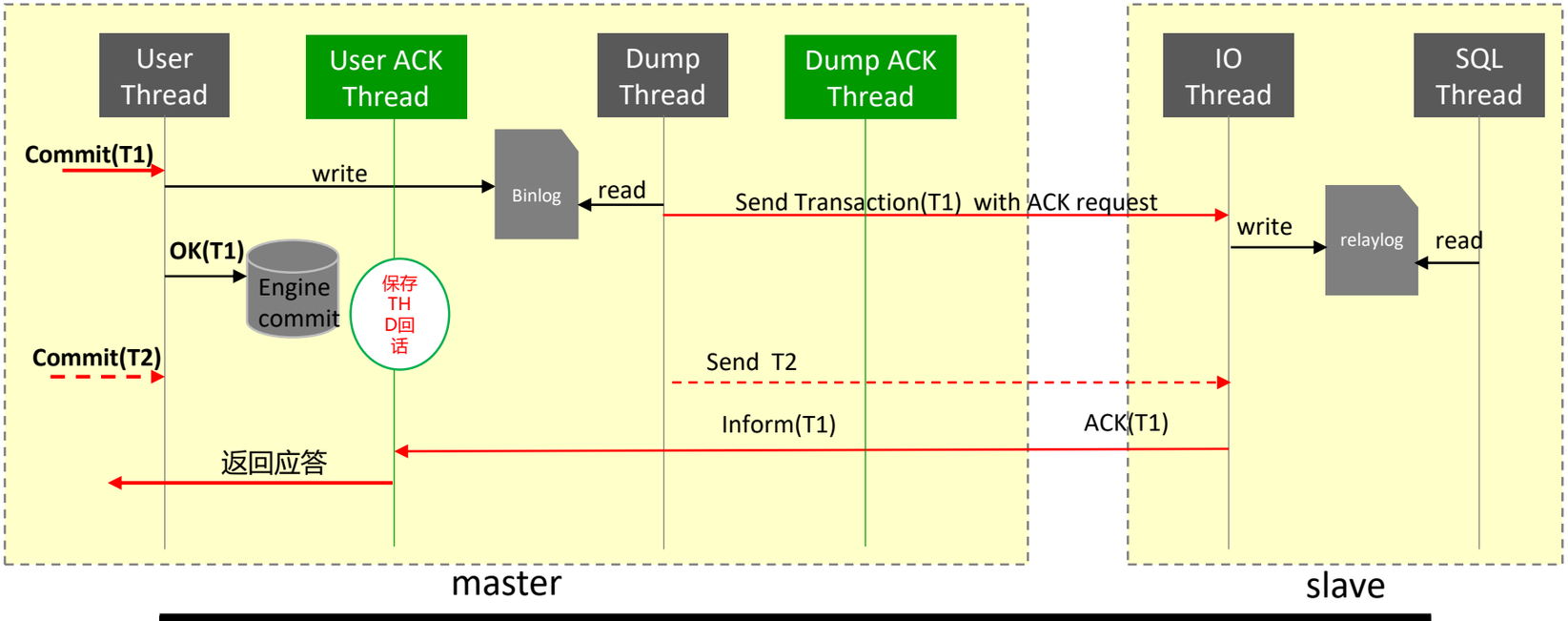
1.超时后蜕化成异步，金融场景不合适

2.跨IDC的情况下性能不乐观

主备复制方案（同城跨IDC）	TPS	时耗(ms)
异步	20,000	<10
半同步	2,200	4~600
MariaDB Galera Cluster	6,000	4~10000

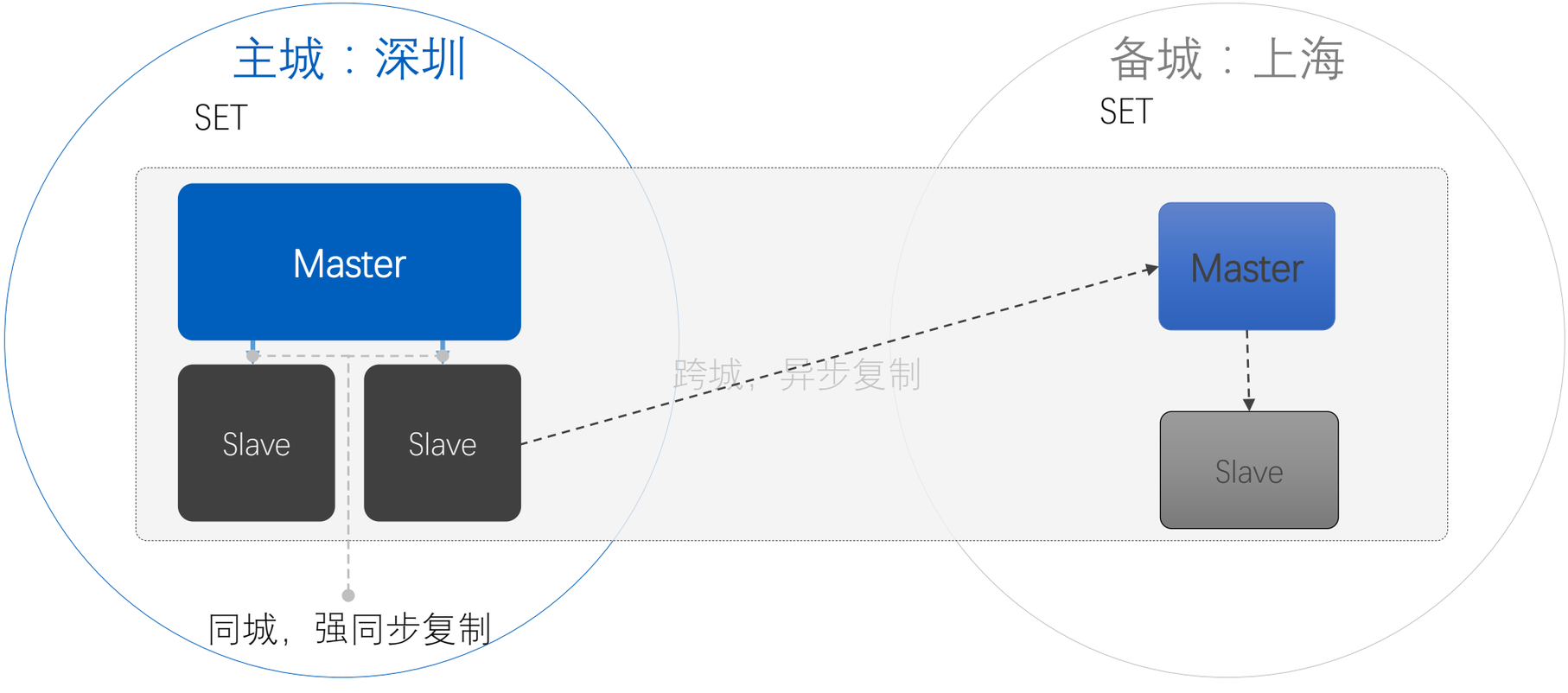
备注：以上测试数据仅有对比意义

用户线程异步化



主备复制方案（跨IDC）	TPS	时耗(ms)
异步	20,000	<10
半同步	2,200	4~600ms
强同步	20000	<10
MariaDB Galera Cluster	6,000	4~10000ms

SET结构


主城：深圳
SET
Master
Slave
Slave
同城，强同步复制
跨城，异步复制
备城：上海
SET
Master
Slave

主备高一致性保障

Consistency guarantee

高一致性容灾 — 如何保证没有脏数据

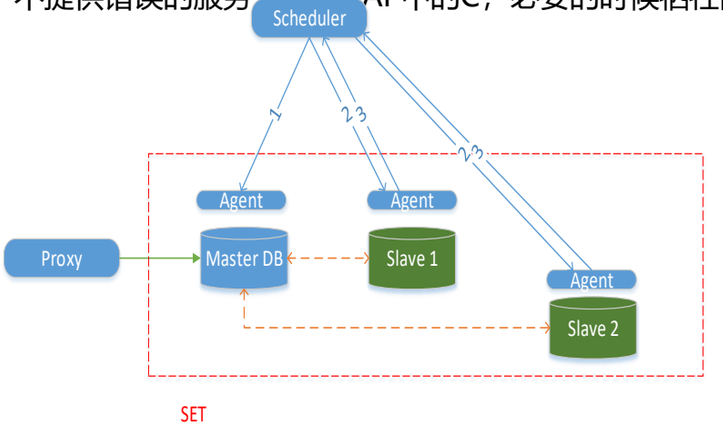
DBAplus

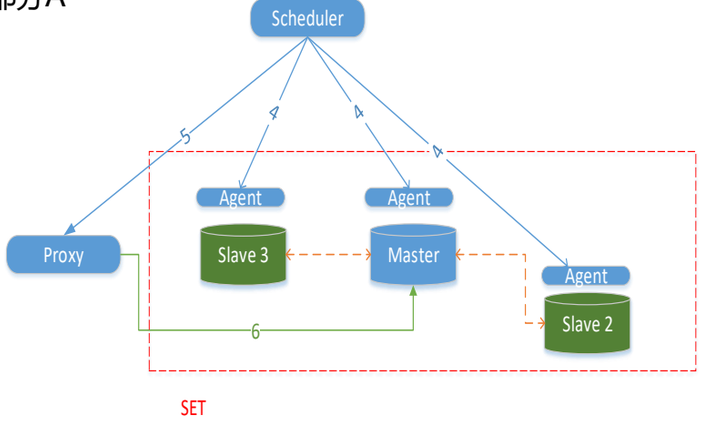
原则：

1、主机可读可写，备机只读，备机可以开放给业务查询使用

2、任何时刻同一个SET不能有两个主机

3， 宁愿拒绝服务，不提供错误的服务，追求CAP中的C，必要的时候牺牲部分A





1、主DB降级为备机

2、参与选举的备机上报最新的binlog点

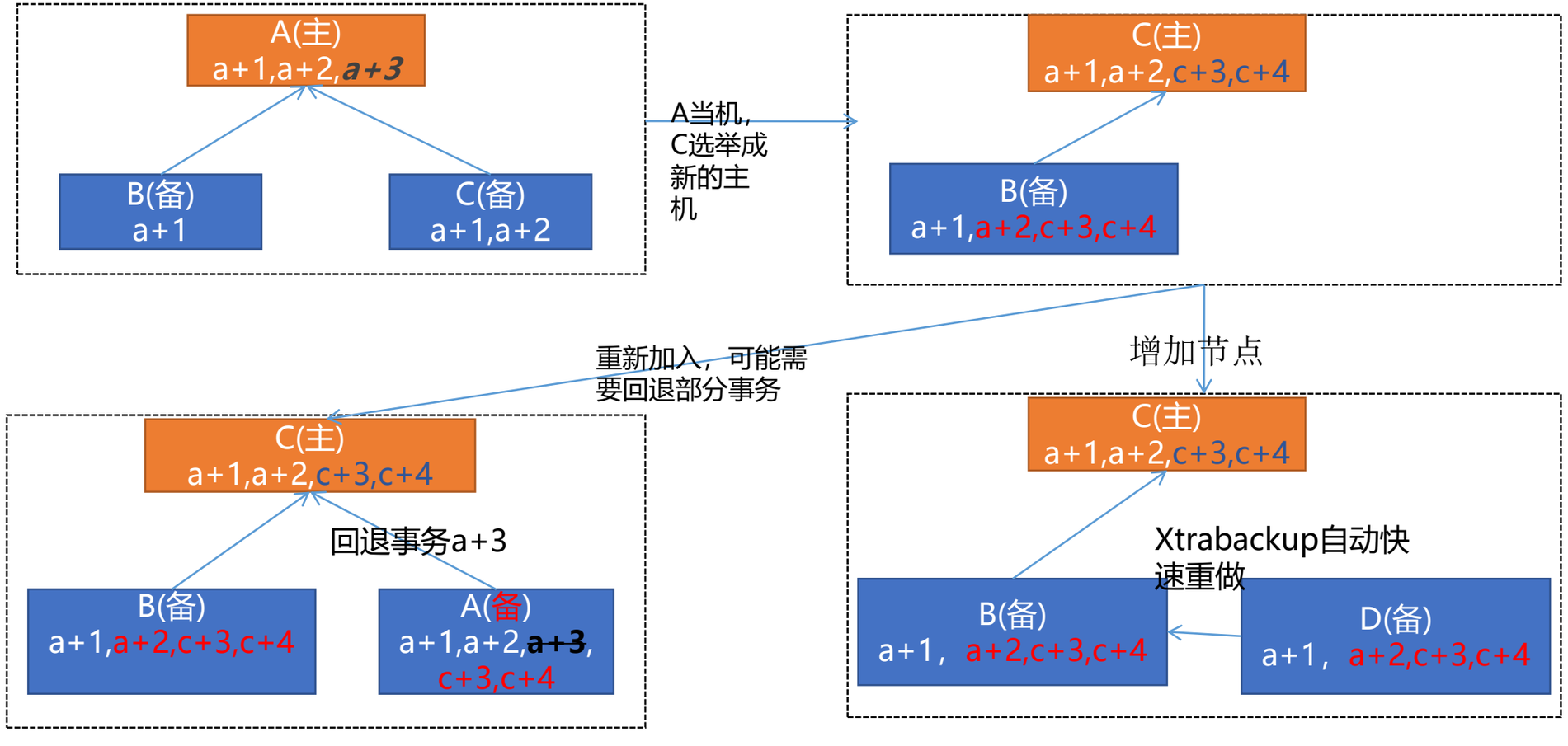
3、scheduler收到binlog点之后，选择出binlog最大的节点

4、重建主备关系

5、修改路由

6、请求发给新的主机

数据高可用性的保障机制（恢复）

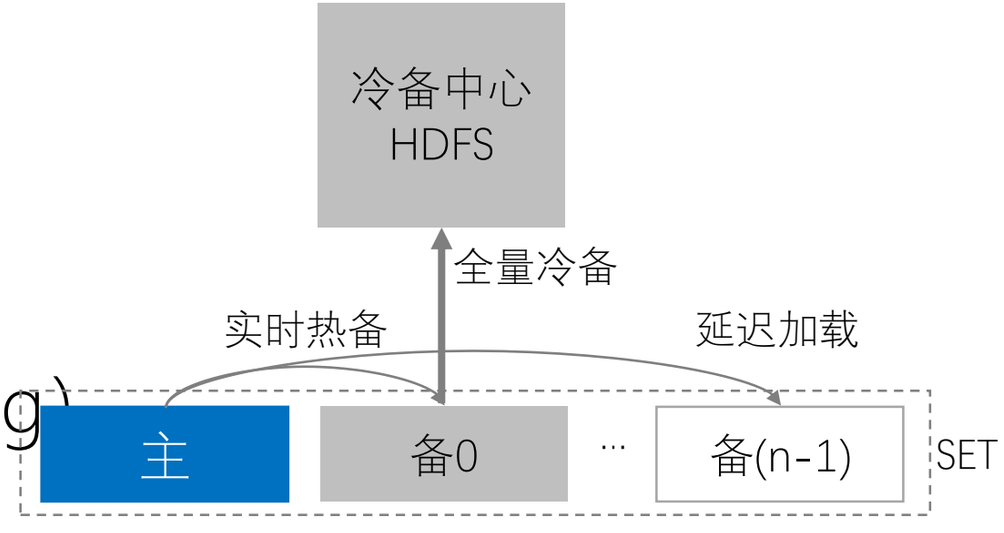


可靠性保障体系

DBAplus

- 数据备份
  - 热备：实时同步，实时加载
  - 冷备：快照 + binlog

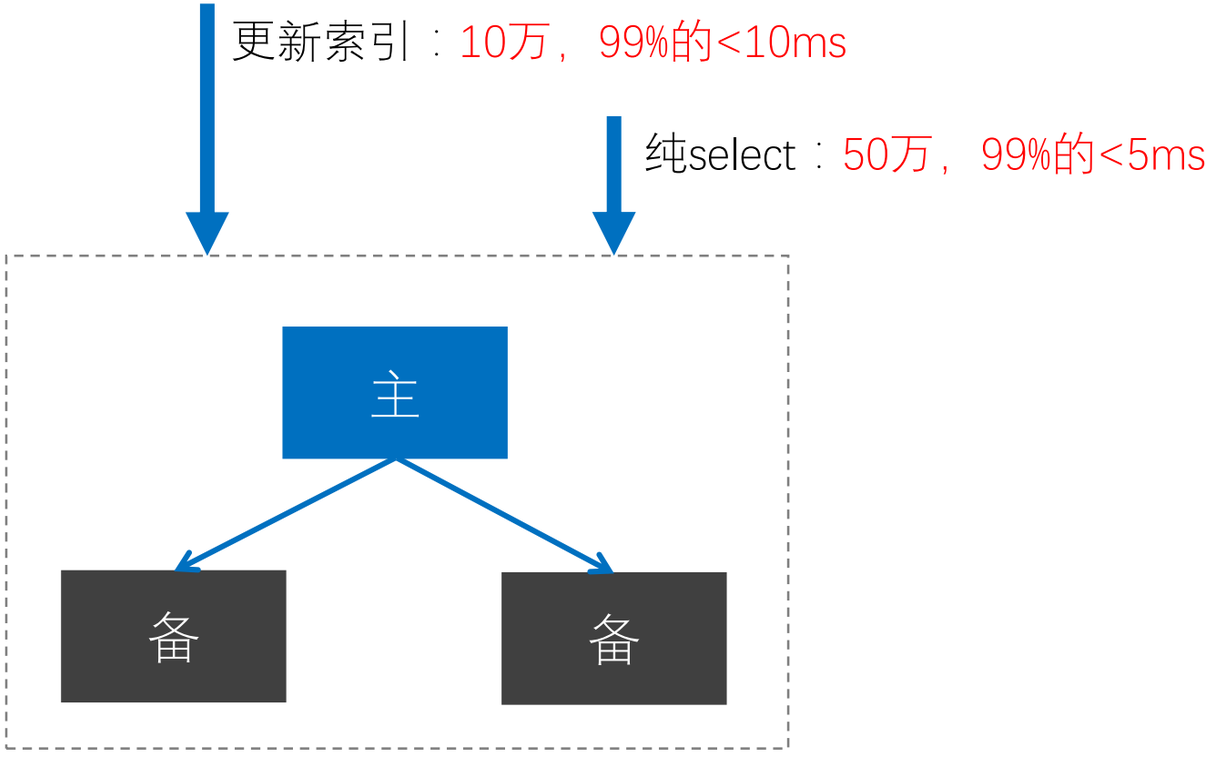
- 数据恢复
  - 就地恢复（闪回/补录）
  - 新节点重建（冷备+binlog）
  - 定点回退（冷备+binlog）



性能

Performance

性能指标：单节点



环境：ts85机型(x86，24核(48超线程)，512G内存 ,6T SSD)

读写分离

DBAplus

- 基于数据库账号的读写分离：

- 基于Hint的读写分离：

```
//主机读//
select * from emp order by sal, deptno desc;
//从机读//
/*slave*/ select * from emp order by sal, deptno desc;
```

2 分布式实践

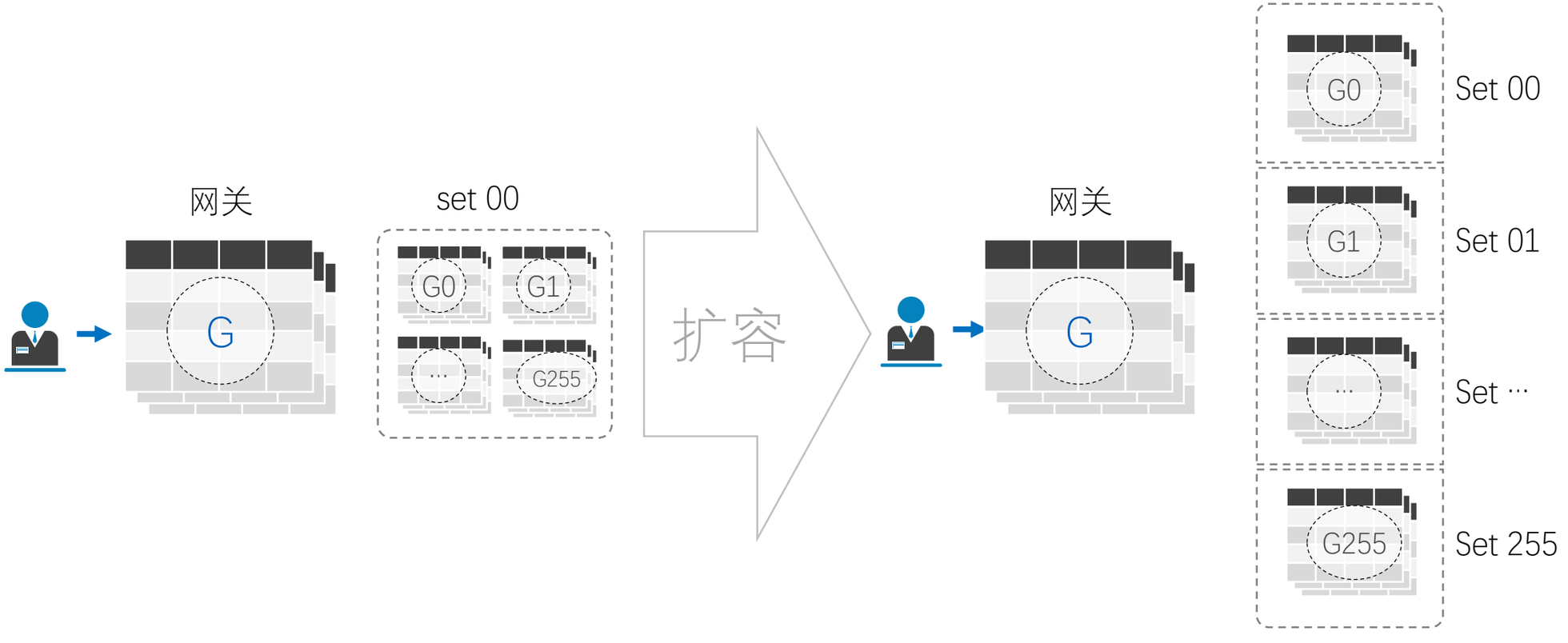
扩展性、分布式事务

水平扩展性

Scalability

GroupShard水平扩容

DBAplus



三种数据Sharding方式

DBAplus

- **Sharding Key方式**
  - create table account(user int , payamt int, c char(20) ,PRIMARY KEY (user)) **shardkey=**user;
  - create table bill(user int , billno int, c char(20) ,PRIMARY KEY (user)) **shardkey=** user;
  - create table dummytable(seqno int , c char(20) ,PRIMARY KEY (seqno)) **shardkey=** seqno;
- **No Sharding方式**，如一些简单的配置表
  - create table noshard_table (a int, b int key, PRIMARY KEY (a));
- **广播小表方式**，支持全局广播
  - create table global_table (a int, b int key, PRIMARY KEY (a)) shardkey=noshardkey_allset;

SQL支持

DBAplus

- group by, order by
- max, min, sum, avg等聚合函数
- distinct, count
- Join (有限支持)
- Transaction (分布式事务)

分布式事务

XA

分布式事务

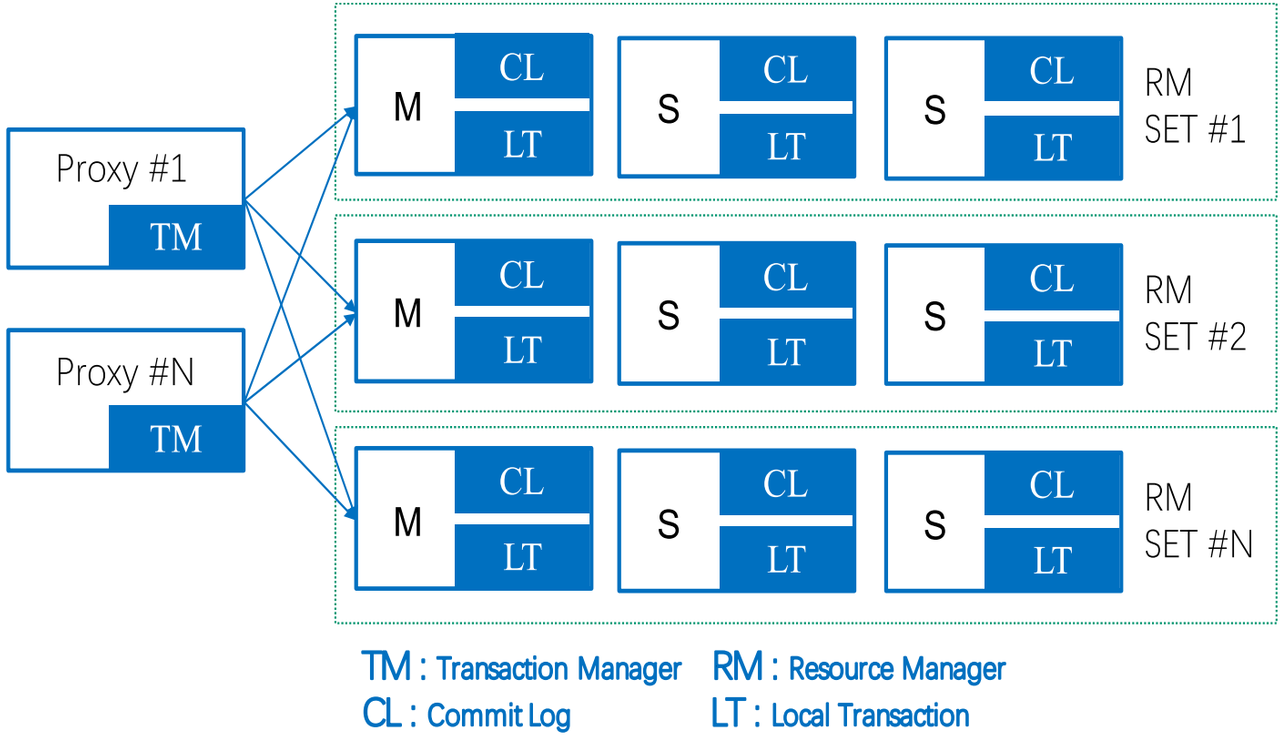
DBAplus

完全去中心化、性能线性增长

健壮的异常处理

全局死锁检测机制

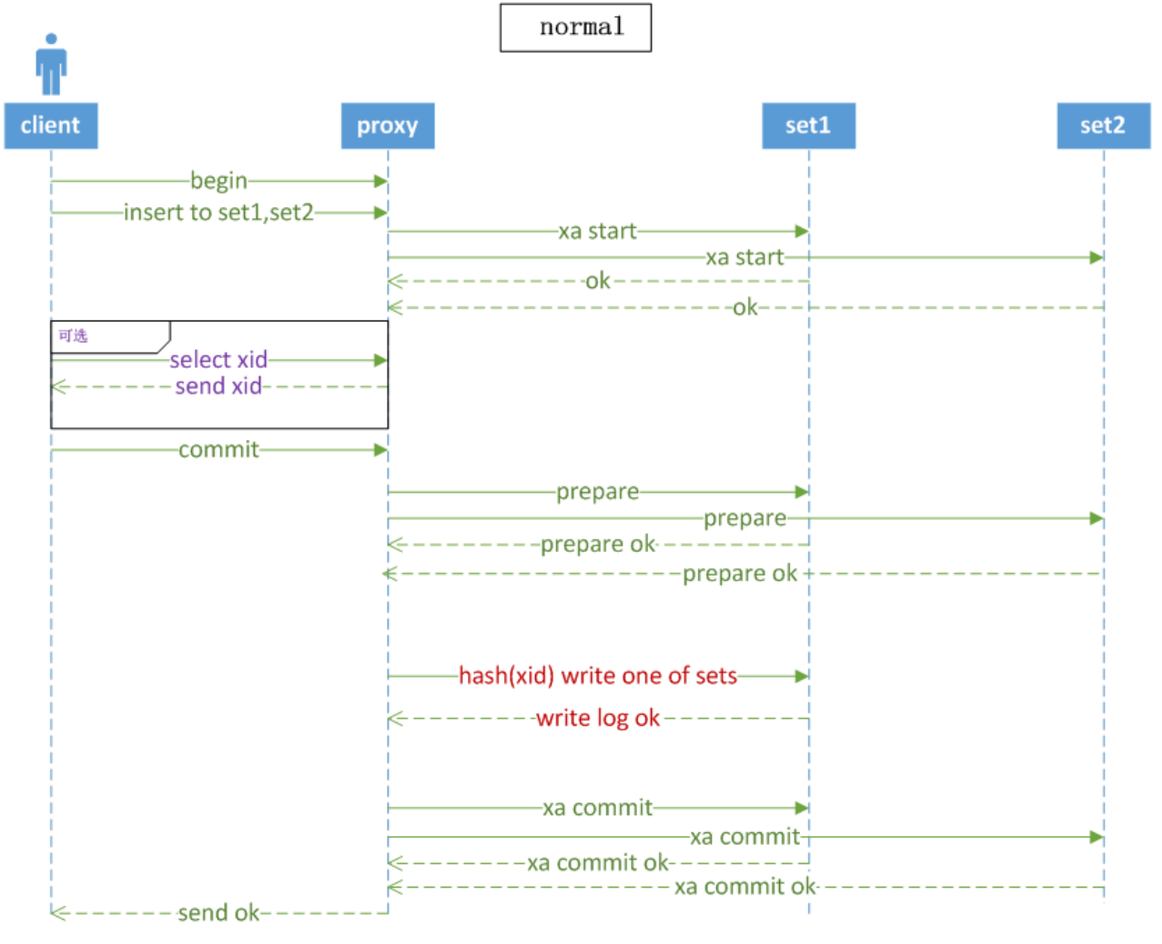
TPCC标准验证



分布式事务

DBAplus

- Prepare 超时或者失败
- Commit log写失败
- Commit log写超时
- Commit超时或者失败
- 异常的总结



分布式事务 -- 异常总结

DBAplus

- 调用commit之前遇到超时或者失败可以直接回滚

- 调用commit的结果:
  - 1)成功，所有涉及到事务的节点都更新成功
  - 2)失败，所有的操作会回滚
  - 3)超时，不能立即确认是成功还是失败，但是通过select xid可以查询到事务的id,后面可以通过这个xid查询事务最终的执行结果

分布式事务性能对比

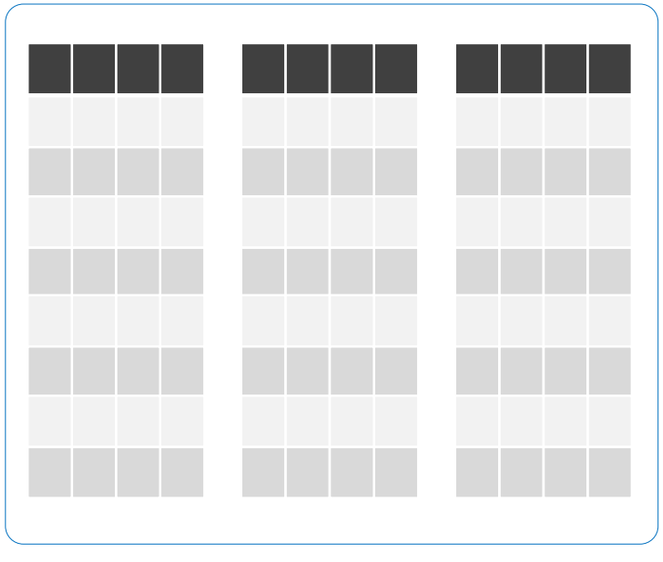
DBAplus

- 采用Sysbench进行测试（每个数据库实例10张表，每个表1000000条记录,每个事务2个update,一个insert,一个delete）
- 256个连接，每个用例测试30分钟
- 测试结果：
  - 单个proxy,单个mysql实例 tps：30000
  - 2个proxy,2个mysql实例 tps：43500
  - 分布式事务下性能是非分布式的 43500/(30000*2)=72%

两种模式

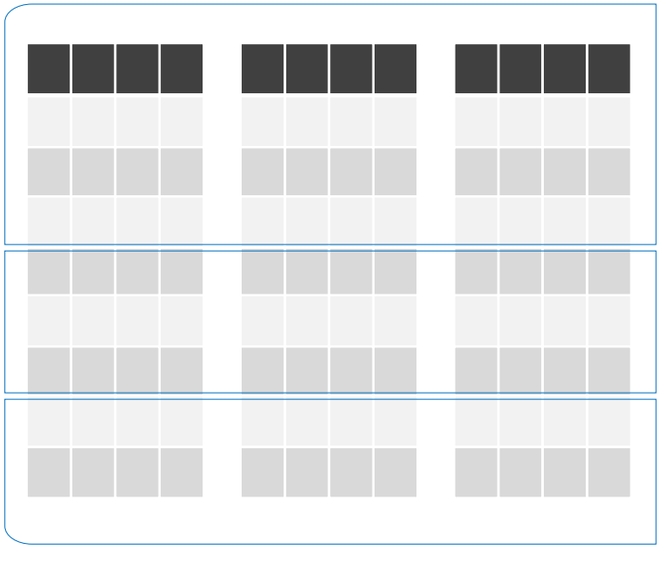
DBAplus

No_sharding



完全兼容	高一致性	自动容灾	水平伸缩

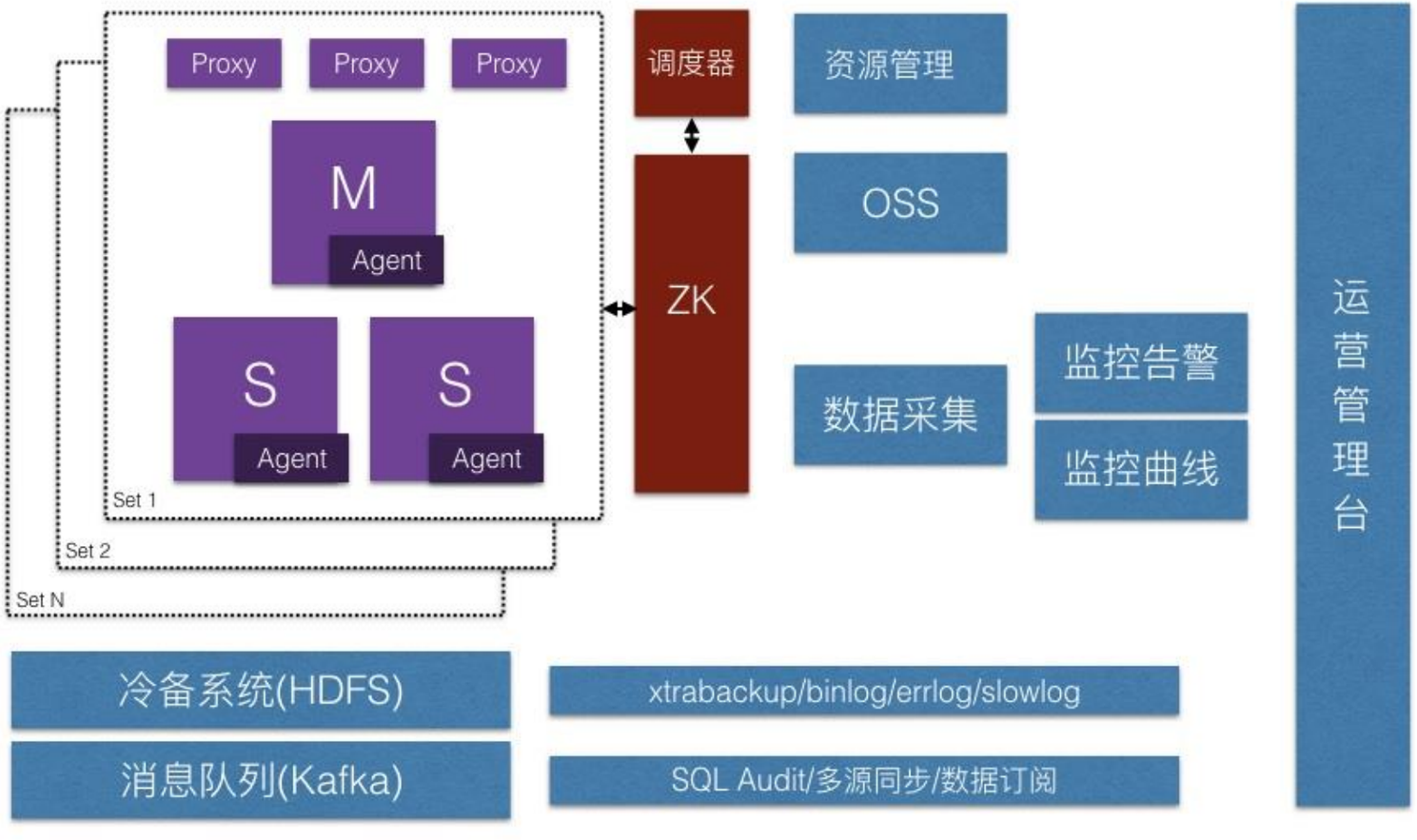
Group_sharding



完全兼容	高一致性	自动容灾	水平伸缩

TDSQL整体视图

DBAplus

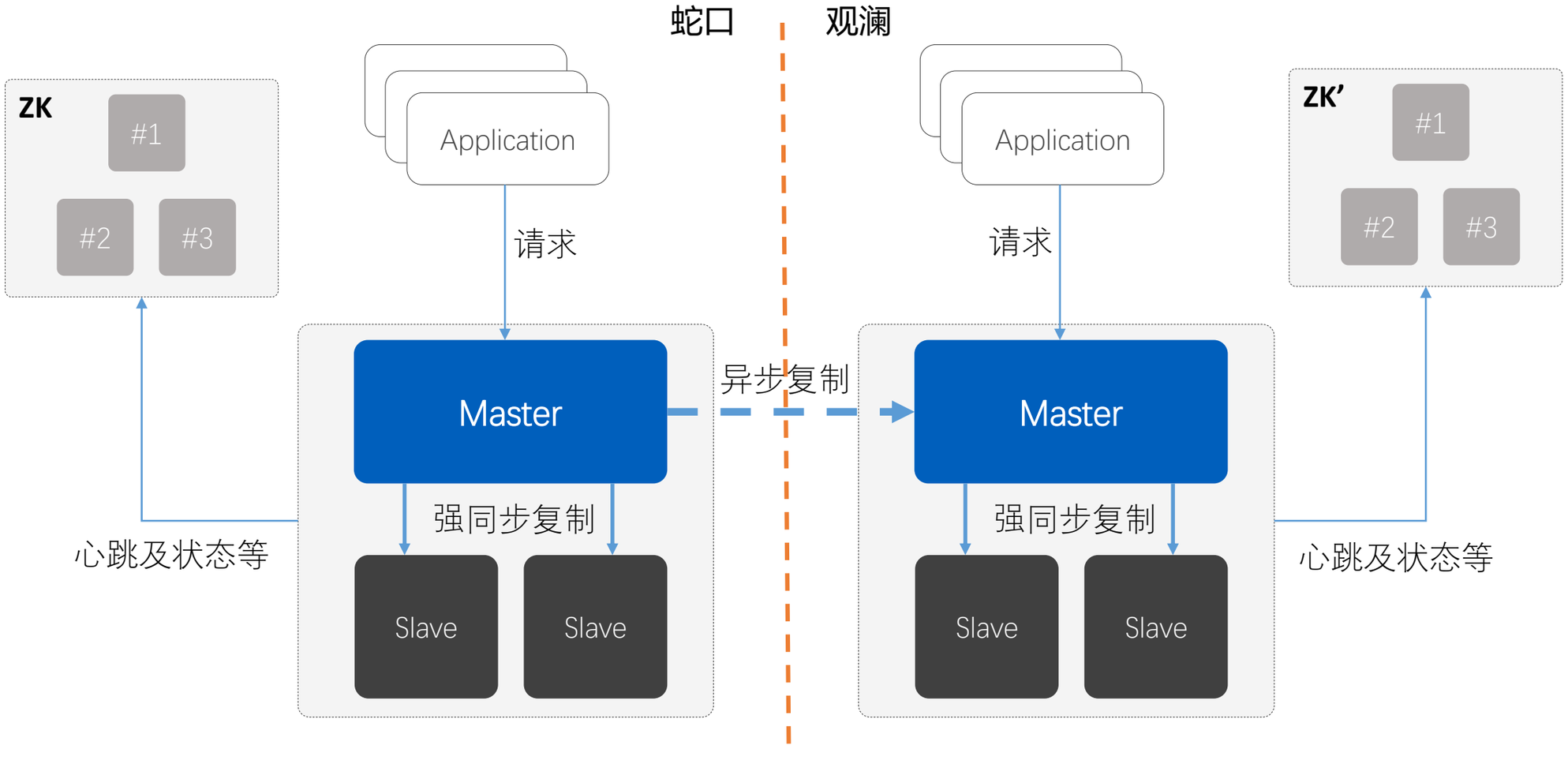


3 部署实践

多地多中心、强同步异步灵活部署

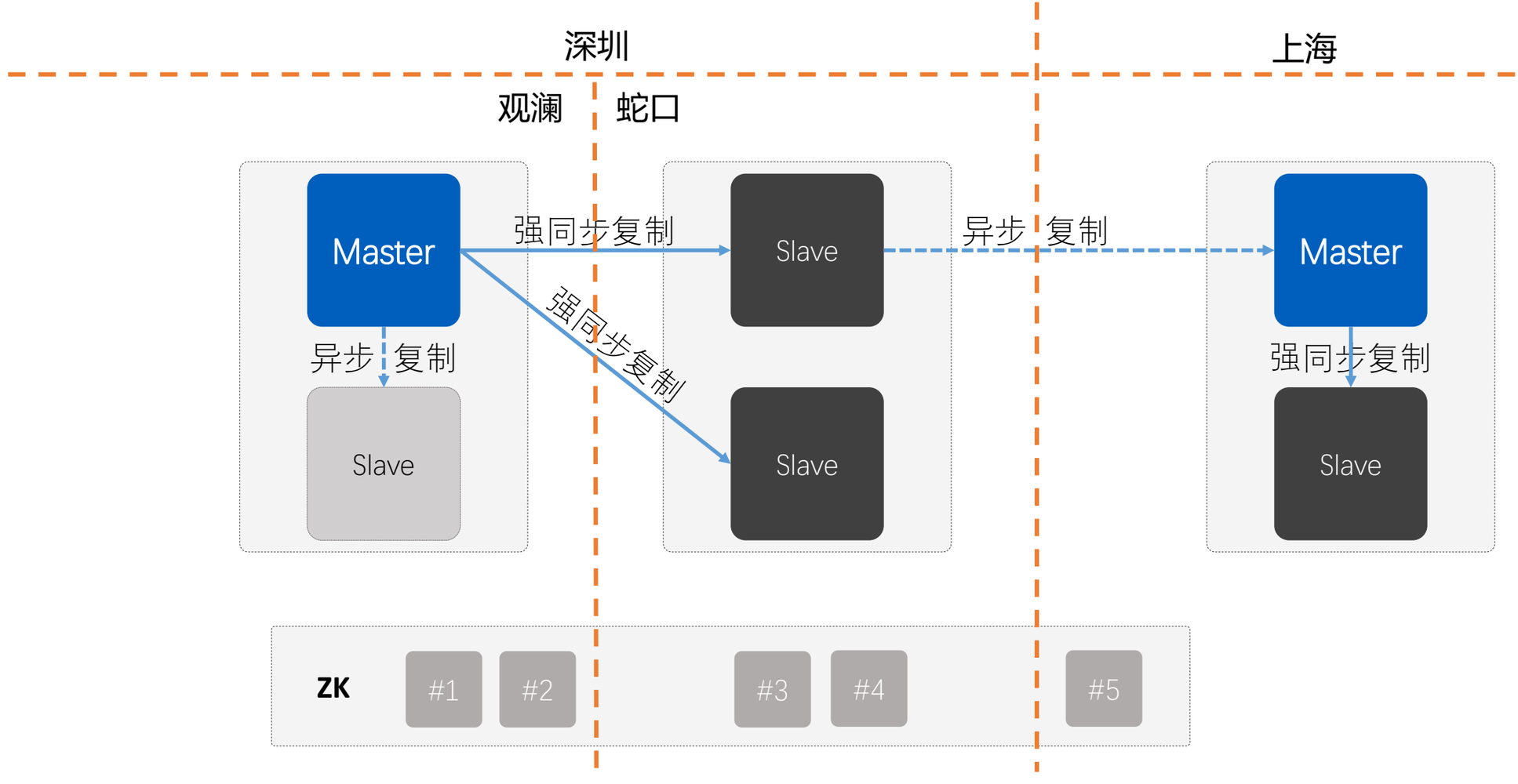
同城主从双中心

DBAplus

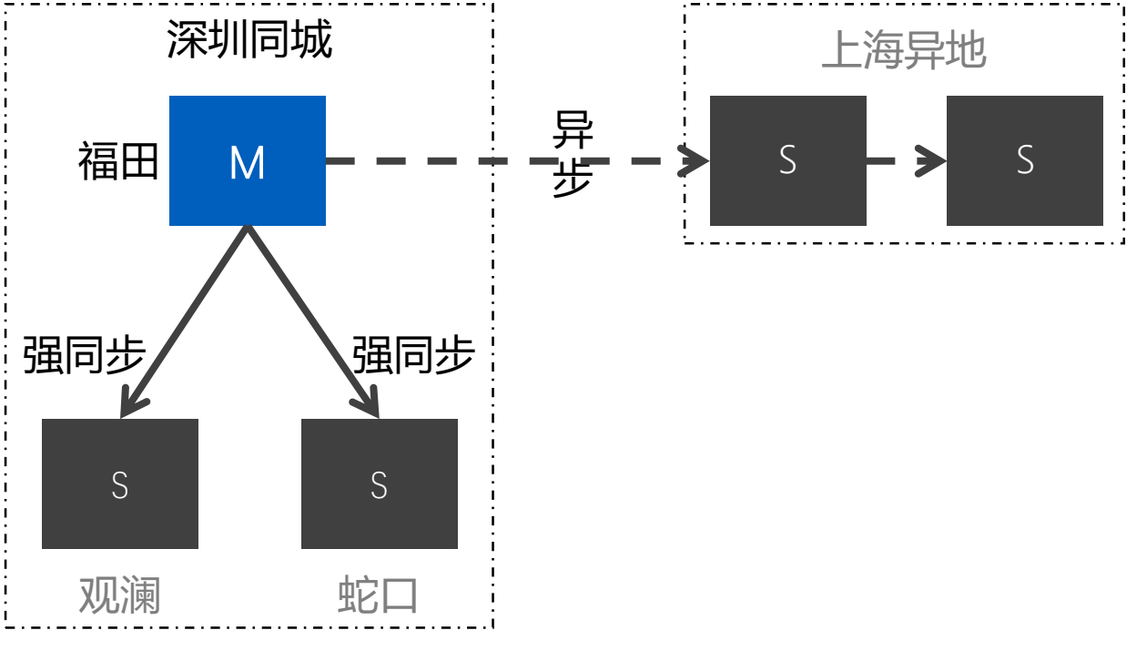


两地三中心

DBAplus



两地四中心 ---(自动化切换的强同步架构)



- 同城三中心集群化部署，简化同步策略，运营简单，数据可用性、一致性高
- 单中心故障不影响数据服务
- 深圳生产集群三中心多活
- 整个城市故障可以人工切换


DBAplus
www.dbaplus.cn
THANK YOU !