谭俊青@途牛，资深架构师，技术总监

跨数据中心状态同步
两地三中心的理论基础

分布式协议/概念

- 分布式一致性算法
  - Paxos (zookeeper, chubby,doozer)
  - Raft (etcd)
  - 2PC
- CAP
- Sharding

背景

- 核心系统在南京：BOSS，呼叫中心
- 南京没有BGP机房
- 网站、无线要照顾到用户体验
- WEB服务器放在北京BGP机房
- 南北机房跨专线调用

远距离跨机房同步问题

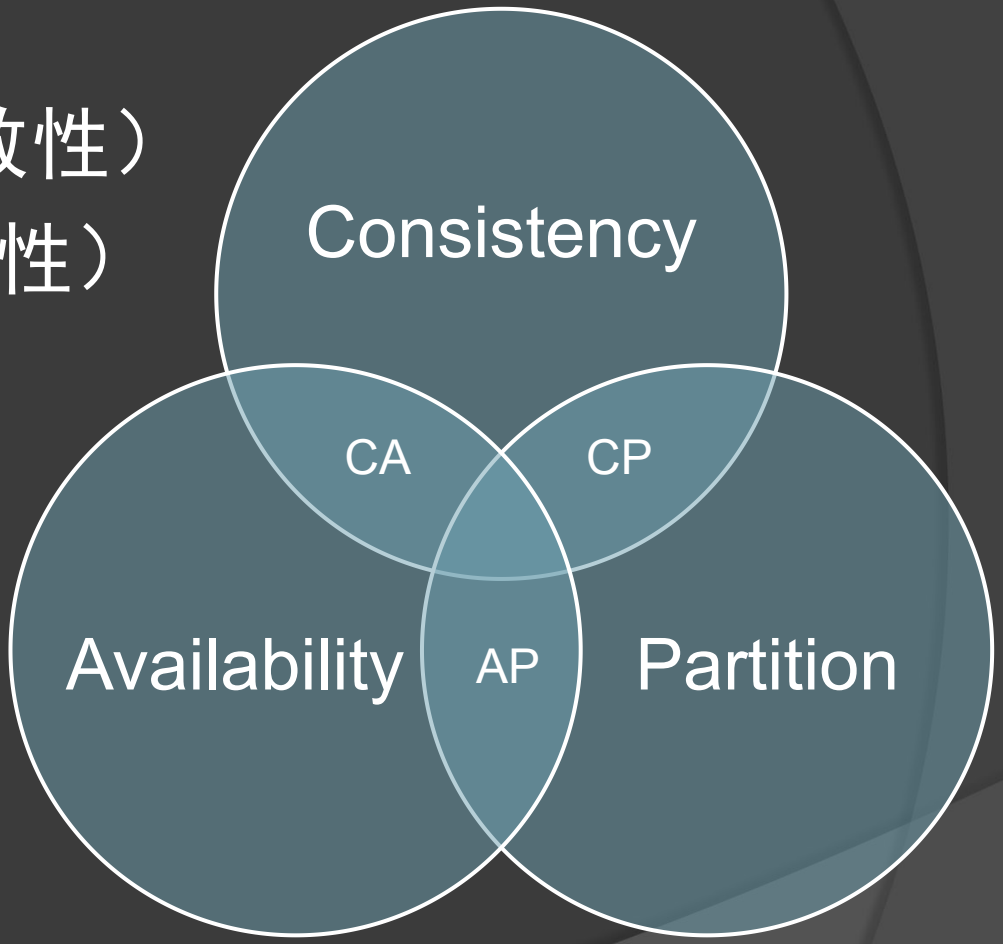
- 专线稳定性影响服务质量
- 带宽挤占(成本高，带宽小)

- 数据库同步延时
- 会员注册、修改密码登陆出错
- 不能做强一致性高可用（丢失数据）

CAP

- Consistency（一致性）
- Availability（可用性）
- Partition tolerance（分区容错性）

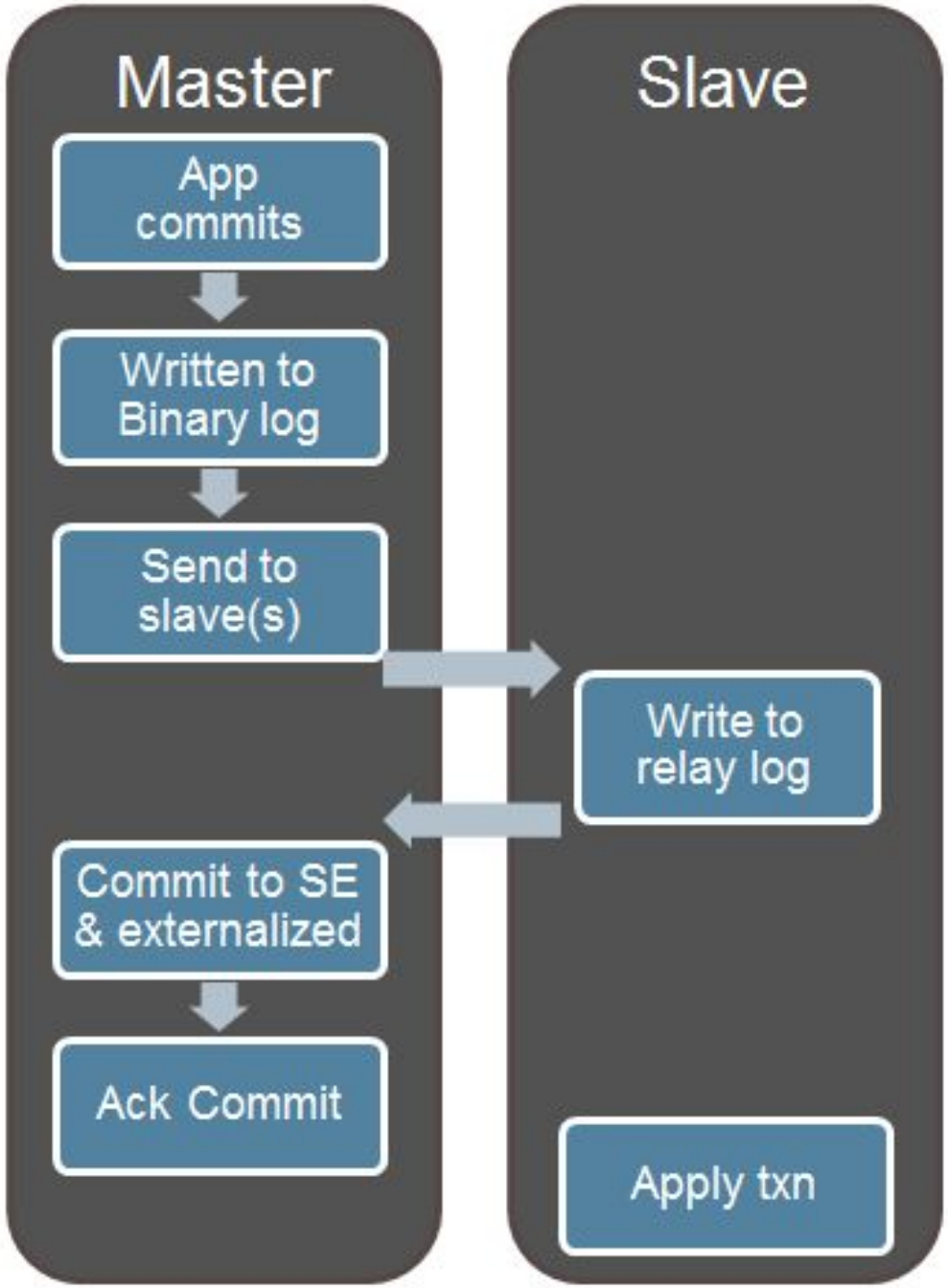
- 三者不可得兼



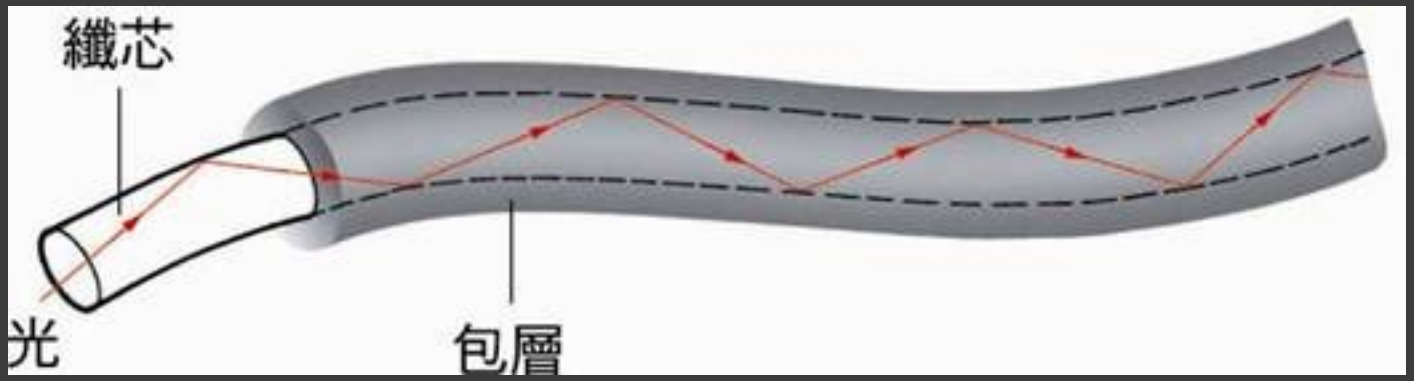
我们的选择 (CP)

- 主从复制、用来做高可用(Partition)
- 不能丢失数据(Consistency)
- 只能牺牲一定的性能(等待数据被同步)

  - 但远距离机房延时太大
  - 从而影响了系统吞吐量
  - 所以我们做了机房搬迁


Master
App commits
Written to Binary log
Send to slave(s)
Commit to SE & externalized
Ack Commit
Slave
Write to relay log
Apply txn

一条专线通南北



理论传输性能

- 真空中光速：30万千米/S, 30km/0.1ms
- 光纤材质折射率1.45左右
- 全反射传输，路径大于光纤长度
- 折算：<20km/0.1ms

- 同城：1000ms/0.2ms=5000(TPS)
- 南北：1000ms/50ms=20(TPS)

HA组件及双中心HA缺陷

- Heartbeat
- Keepalived
- etc.
- 缺少第三方仲裁，会导致脑裂，从而可能引起数据不一致，进一步造成数据丢失等

- 2F+1 (F=1, 2, …)，一般F=1即可

三中心HA

- 同城数据中心降低了同步延时
- 低延时极大提升了吞吐量
- 第三数据中心参与选举仲裁及灾备恢复

- 为了安全，第三中心一般距离会较远

吞吐量提升

- 2015双11期间，峰值达8.5万TPS
- Oceanbase采用paxos多活是如何实现的（猜测）？
  - Sharding 分区
  - 发号器发号避免冲突，实现高并发
  - 分区间用2pc实现分布式事务
  - 部分采用队列异步处理

总结

- 两地三中心
- 三中心对应状态同步的三节点
- 两地是照顾到安全(异地)和性能(同城)

Q/A

?

谢谢

http://www.mysqlab.net/blog/