百度云

Raft在百度云的实践

百度云　王耀

GTLC
GLOBAL
TECH LEADERSHIP
CONFERENCE

全球技术领导力峰会

通往**年薪百万**的CTO的路上，
如何打造自己的技术**领导**力？

扫描二维码了解详情

个人简介

百度云

- 王耀
- 百度云IaaS主任架构师
- braft开源项目负责人
- 分布式存储系统
- 公有云网络虚拟化

分布式系统日常Keyword

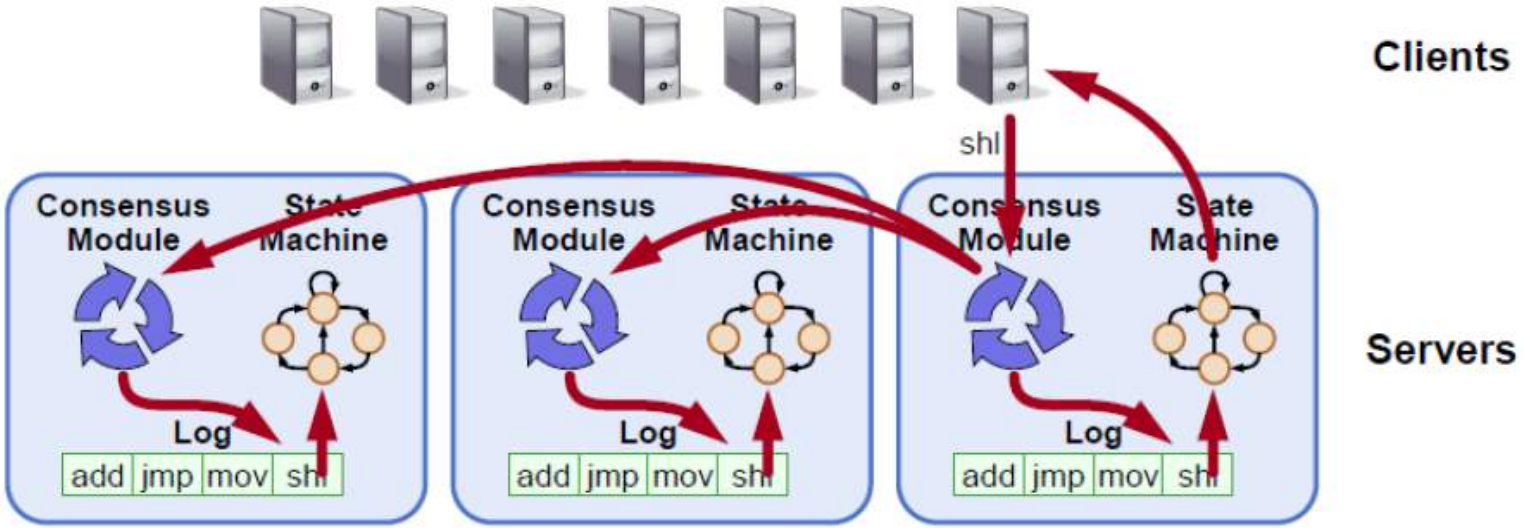
内容提要

百度云

- Raft协议简介
- braft实现简介
- 基于Raft的存储模型
- 百度云CDS存储设计

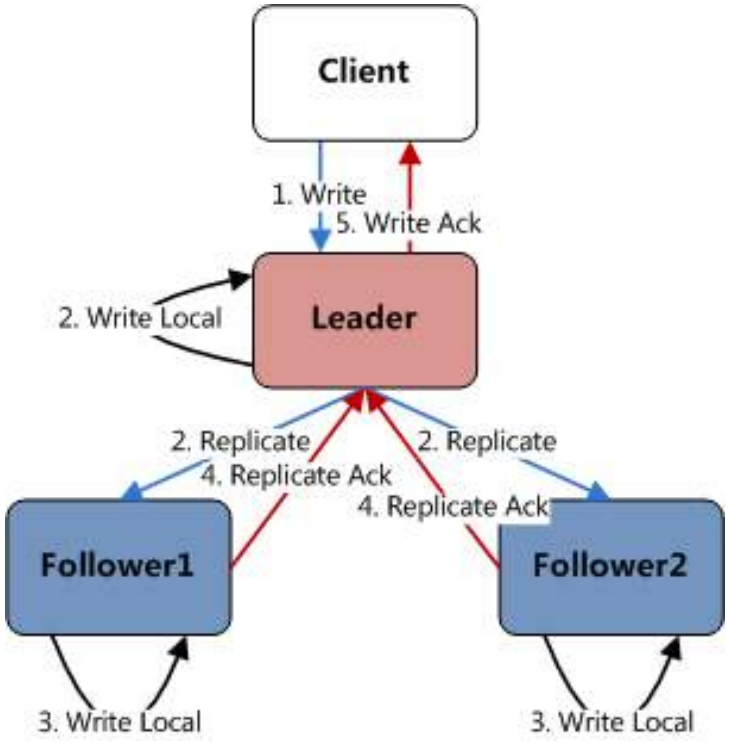
Raft复制状态机

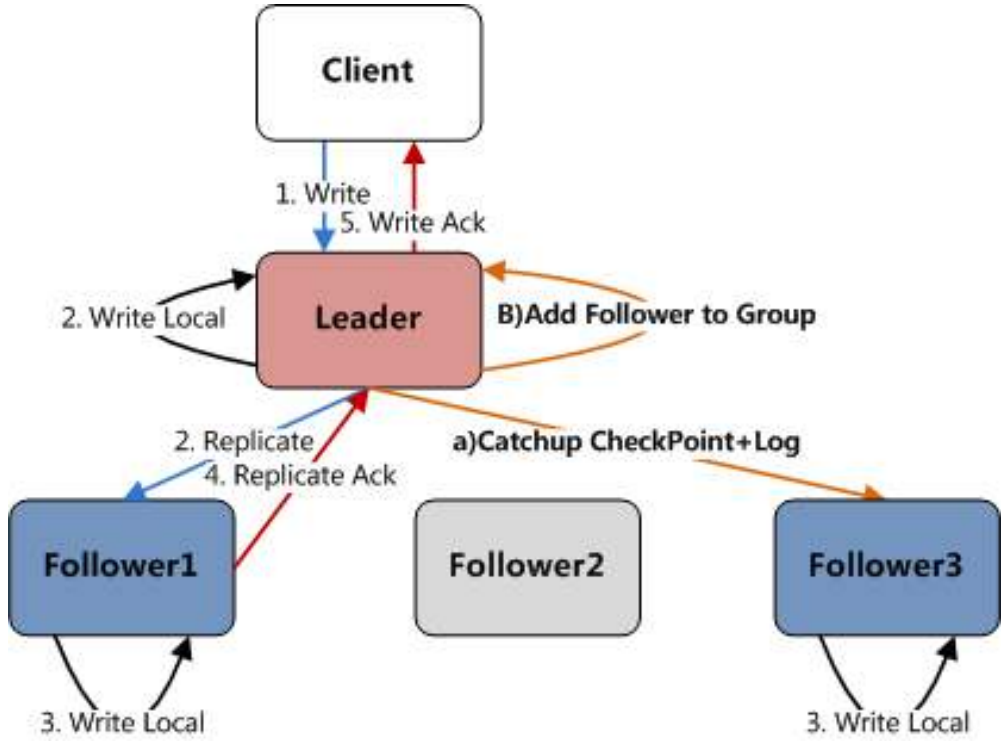
- Leader Election
- Log Replication
- Membership Change
- Log Compaction



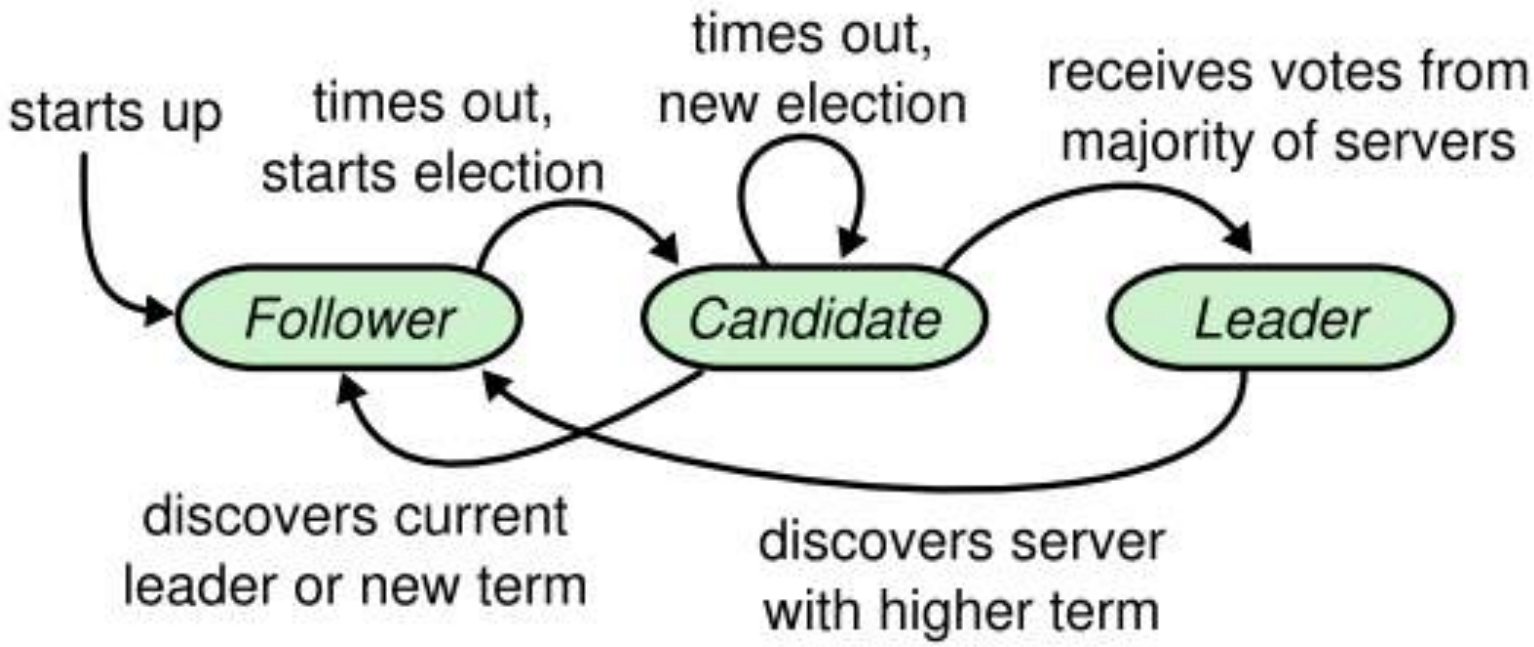
Raft复制结构

- 树形结构
- 多数复制
- 写时修复
- 断点续传





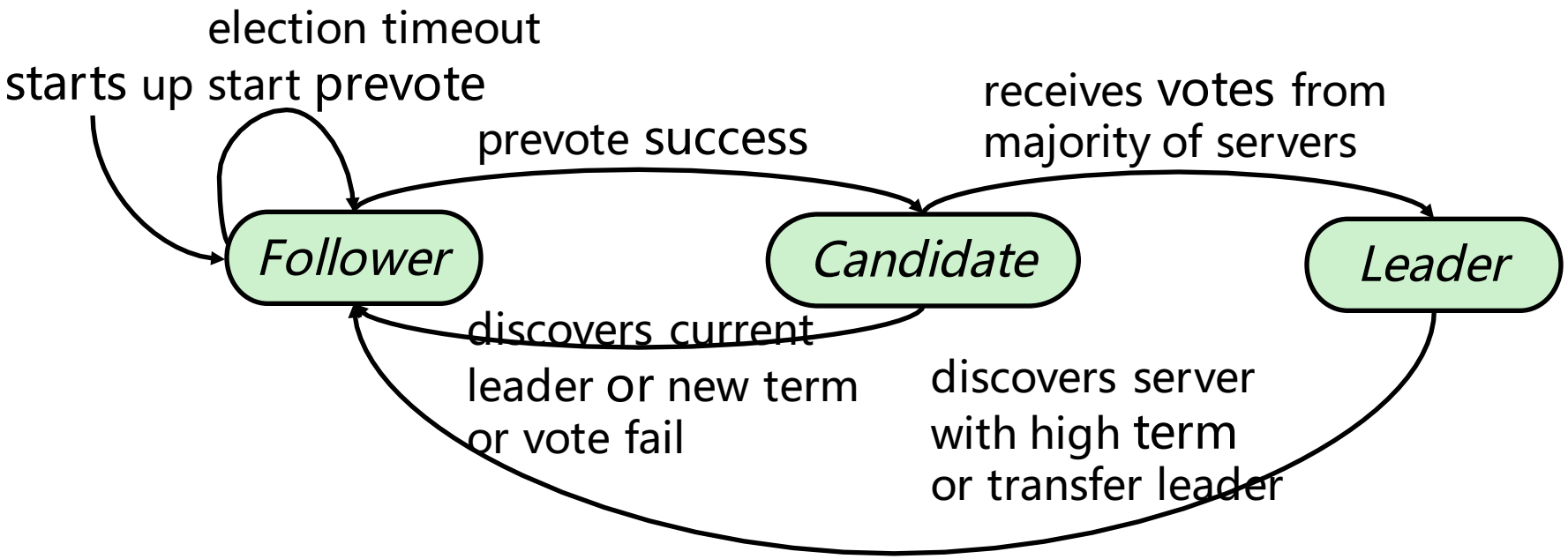
Raft节点状态机



- 捣乱的Candidate
  - 网络划分
  - 节点负载高
- 指定节点为Leader

- 功能完备
  - PreVote
  - Leader Transfer
- 高灵活性
  - 自定义Storage
  - 两阶段InstallSnapshot
- 高性能
  - Append Log Batch
  - Replicate Batch and Pipeline
  - Cache Last LogEntries
  - Apply Async and Batch

braft协议状态机改进



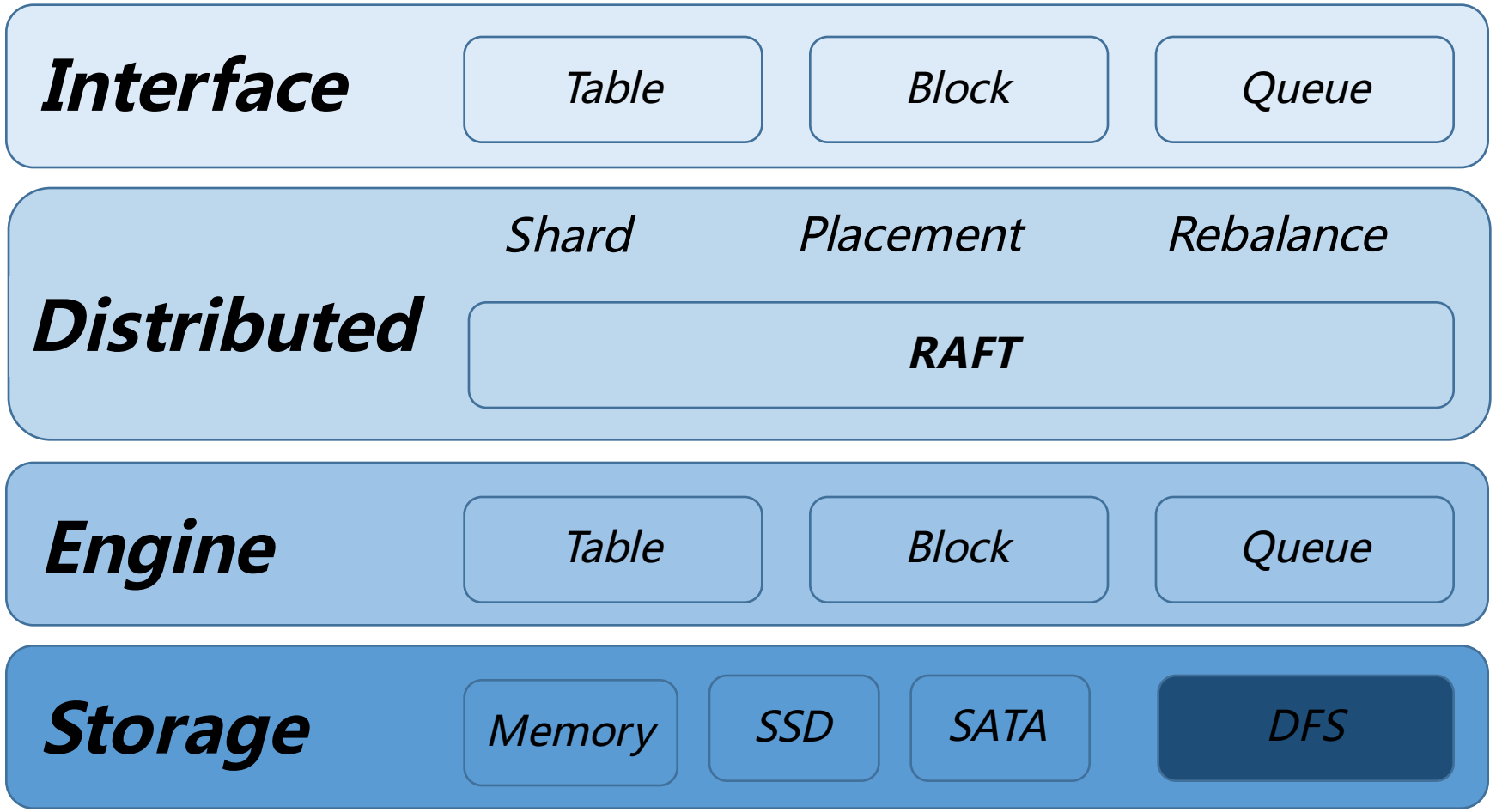
braft使用tips

- on_apply保证主从执行结果一致
- on_snapshot_load要先清空状态机
- on_leader_stop保证leader相关任务cancel
- apply task间调用的结果都是独立的
- apply task和configuration_change存在false negative

braft在百度内的应用

- 元信息管理
  - 容器系统Master
  - 虚机系统Master
  - 流式计算系统Master
- 存储系统
  - 强一致性MySQL
  - 分布式块存储CDS
  - 分布式文件系统CFS
  - 分布式NewSQL TafDB

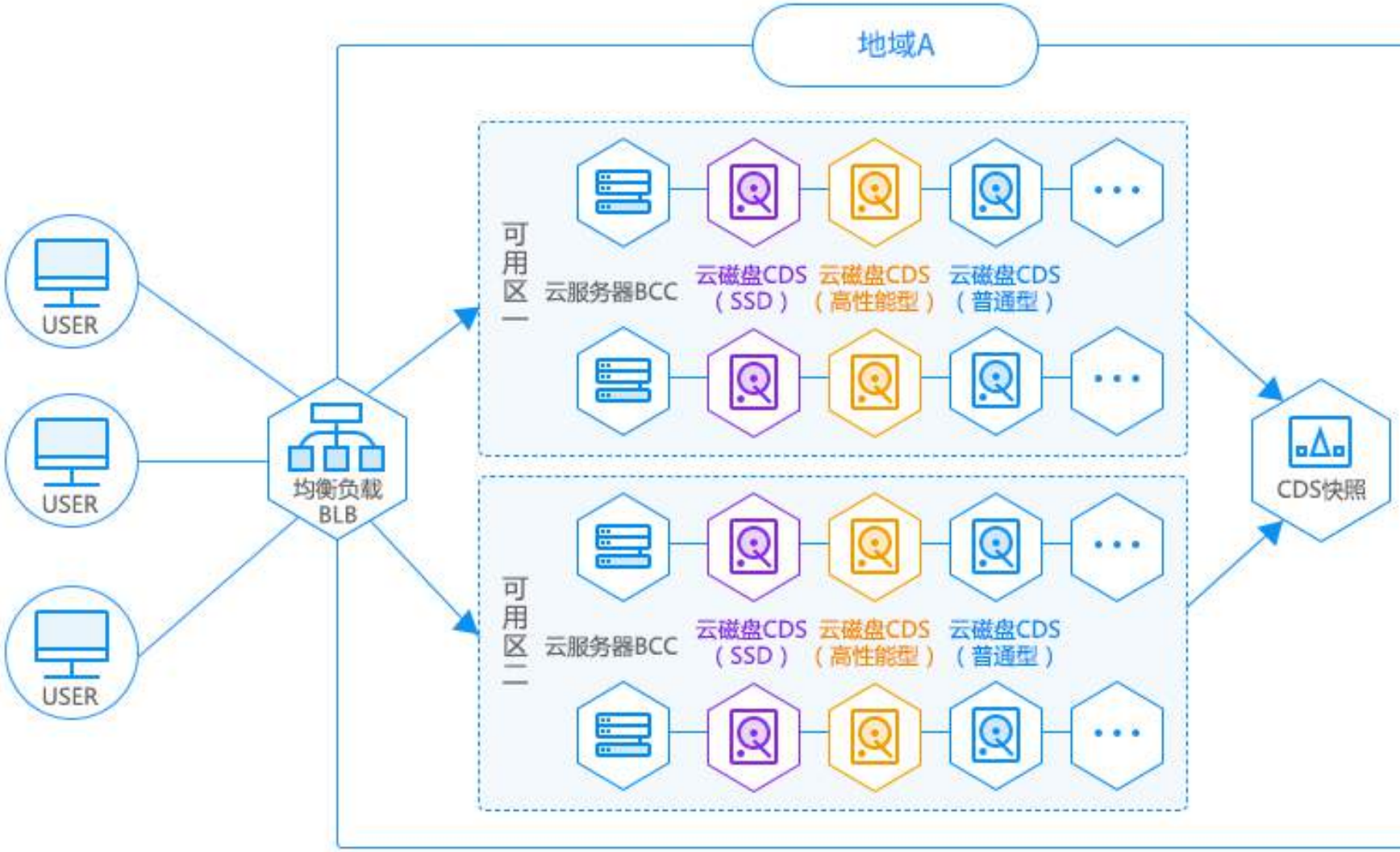
基于Raft的存储模型



存储系统实现步骤

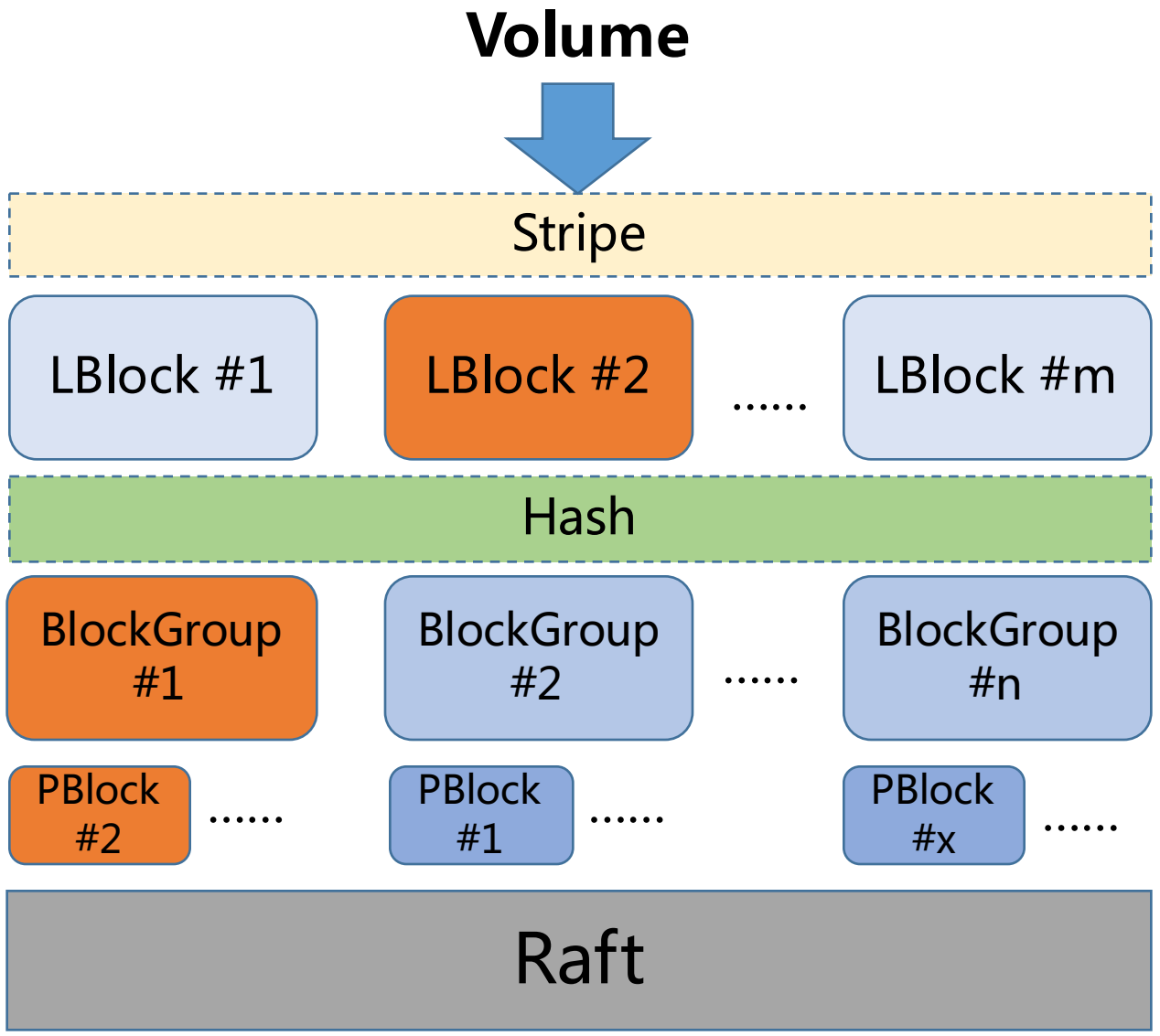
- 模型分析
  - 接口
  - 分片
  - 引擎
- 系统实现
  - 放置
  - 选主、复制、修复
  - 负载均衡
- 系统测试与上线
  - 异常注入
  - 平滑数据迁移

百度云块存储CDS


地域A
USER
USER
USER
均衡负载
BLB
可用区一
云服务器BCC
云磁盘CDS（SSD）
云磁盘CDS（高性能型）
云磁盘CDS（普通型）
可用区二
云服务器BCC
云磁盘CDS（SSD）
云磁盘CDS（高性能型）
云磁盘CDS（普通型）
CDS快照

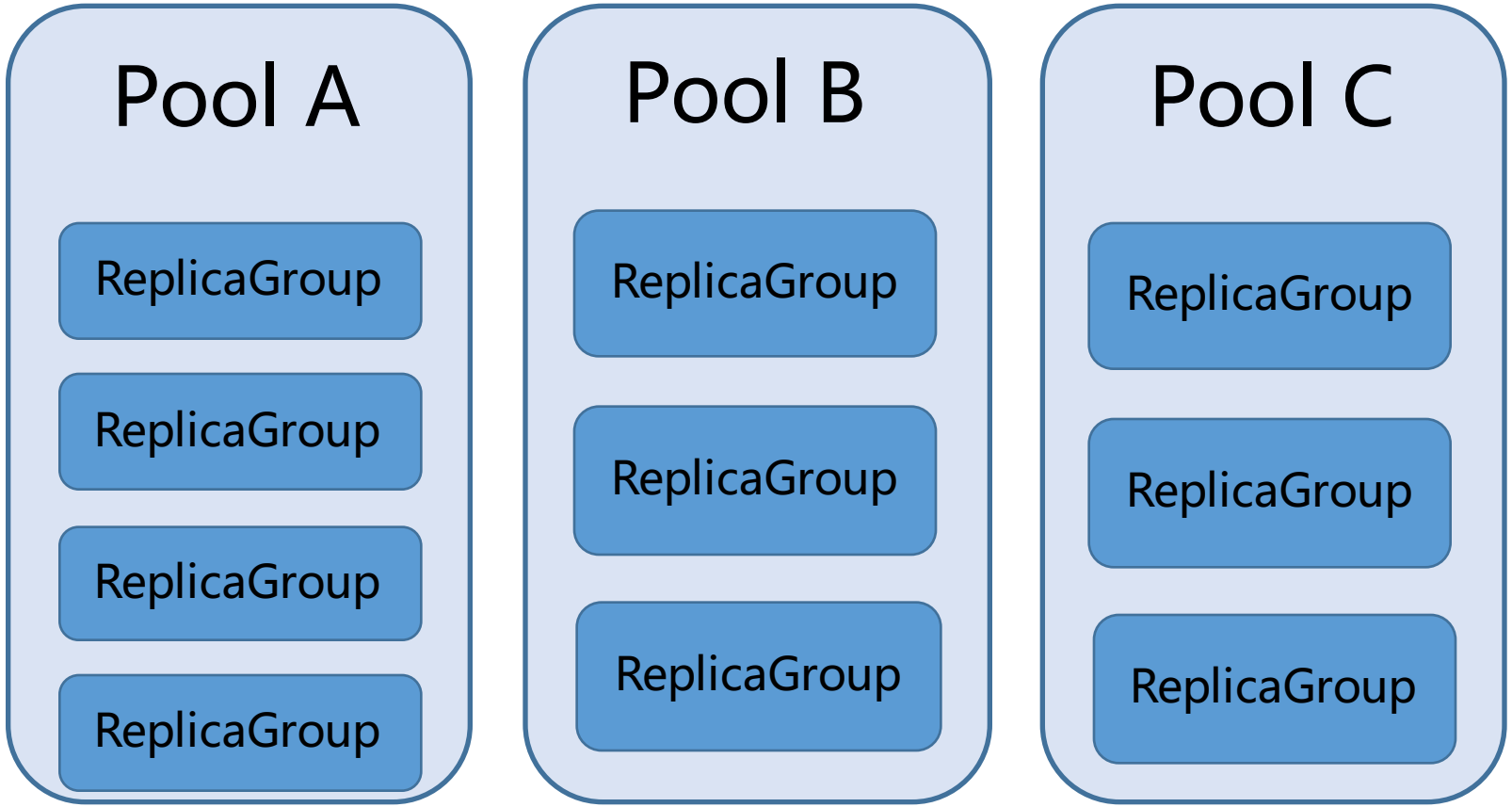
CDS存储模型

- Volume拆Block
- Block聚BlockGroup
- BlockGroup braft复制
- Block多版本引擎



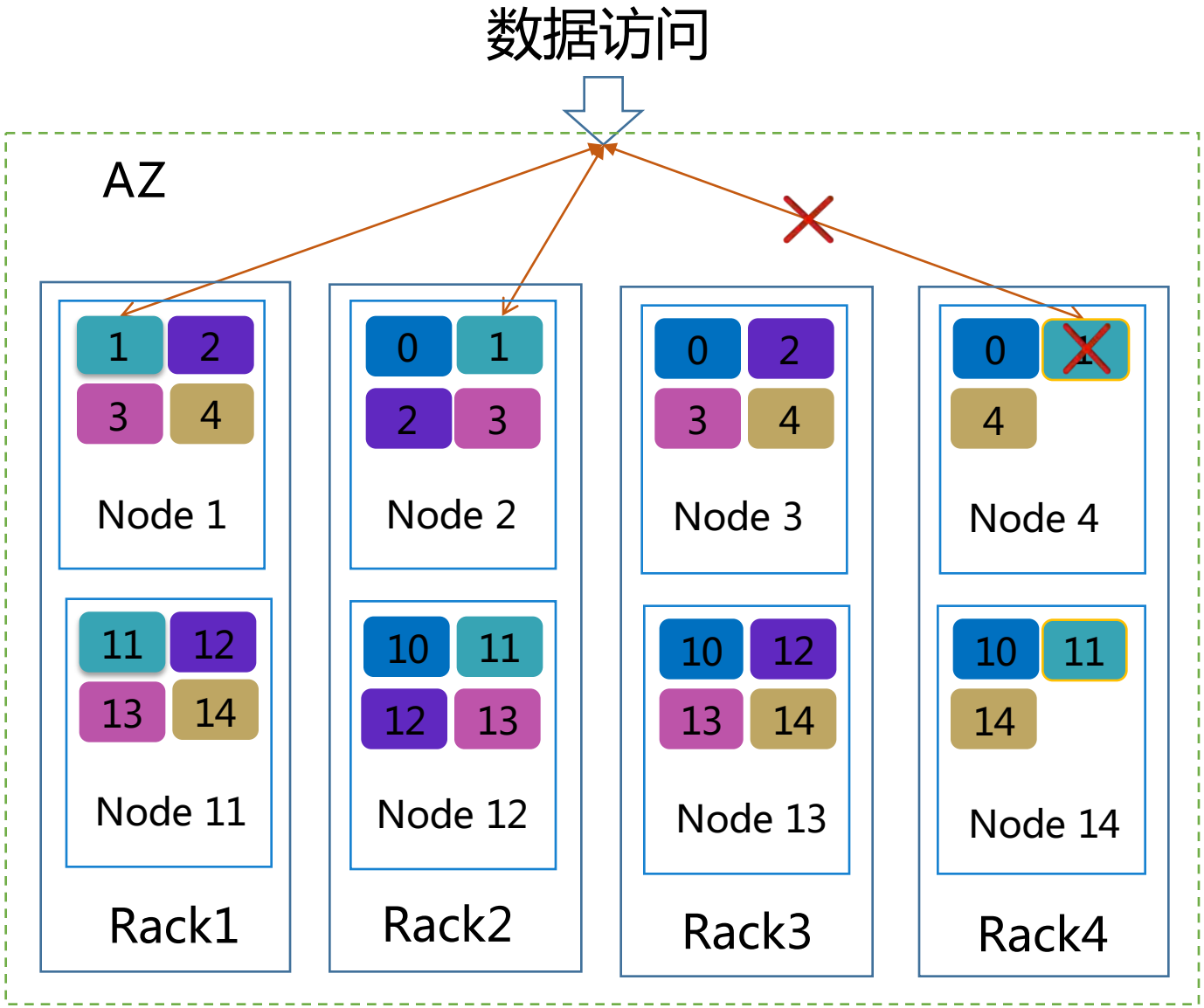
CDS逻辑数据分布

- 两级分布
  - Pool
  - ReplicaGroup



CDS物理数据分布

- 五级隔离
  - Region
  - Zone
  - Rack
  - Node
  - Disk



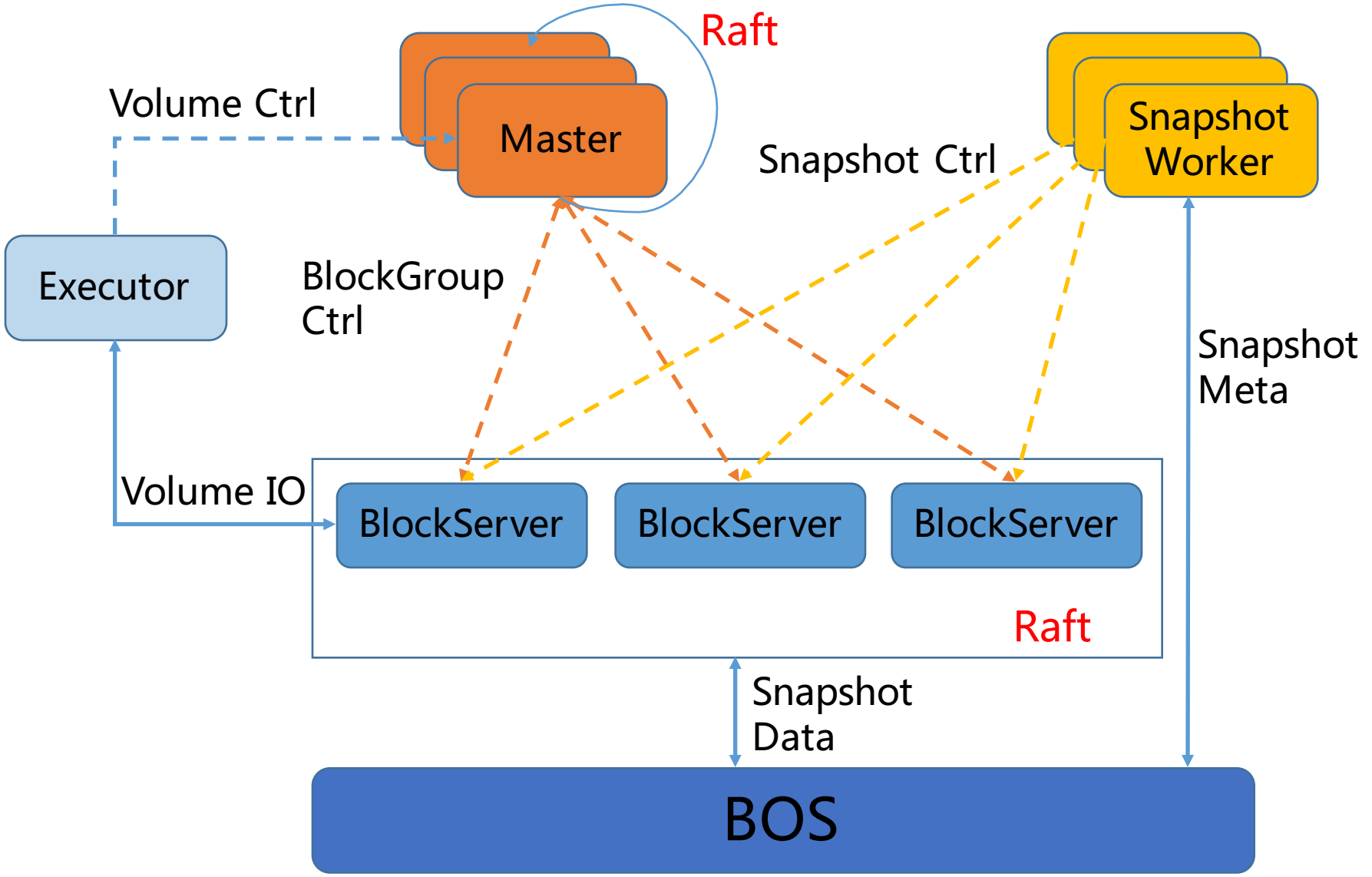
CDS中的副本管理

- Node
  - 定期汇报状态
  - 定期GC垃圾Replica

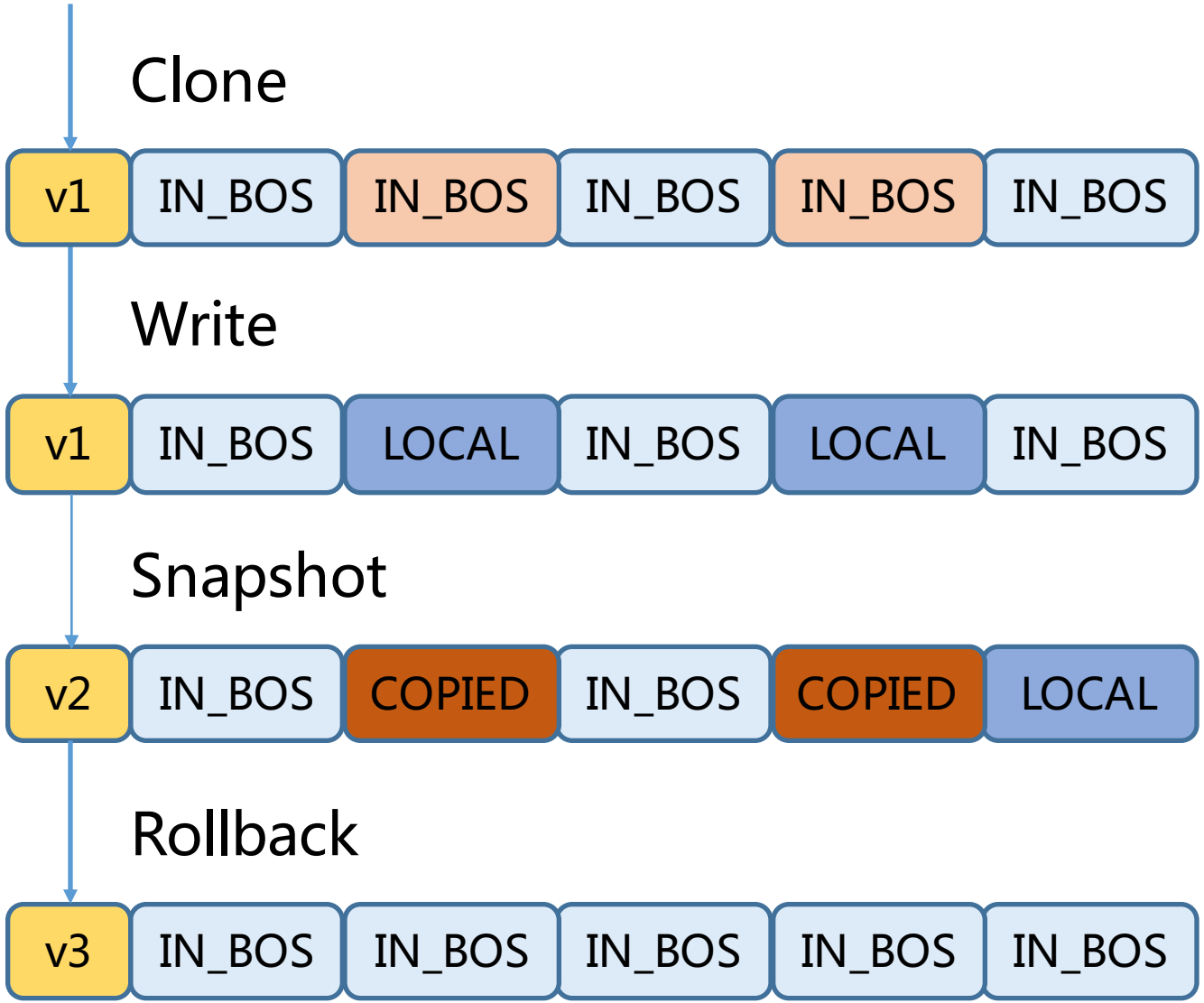
- Master
  - 维护Node和Replica间映射
  - 修复Node/Disk故障
  - 定期Disk容量均衡
    - Replica数量
  - 定期IO负载均衡
    - Leader数量

CDS架构

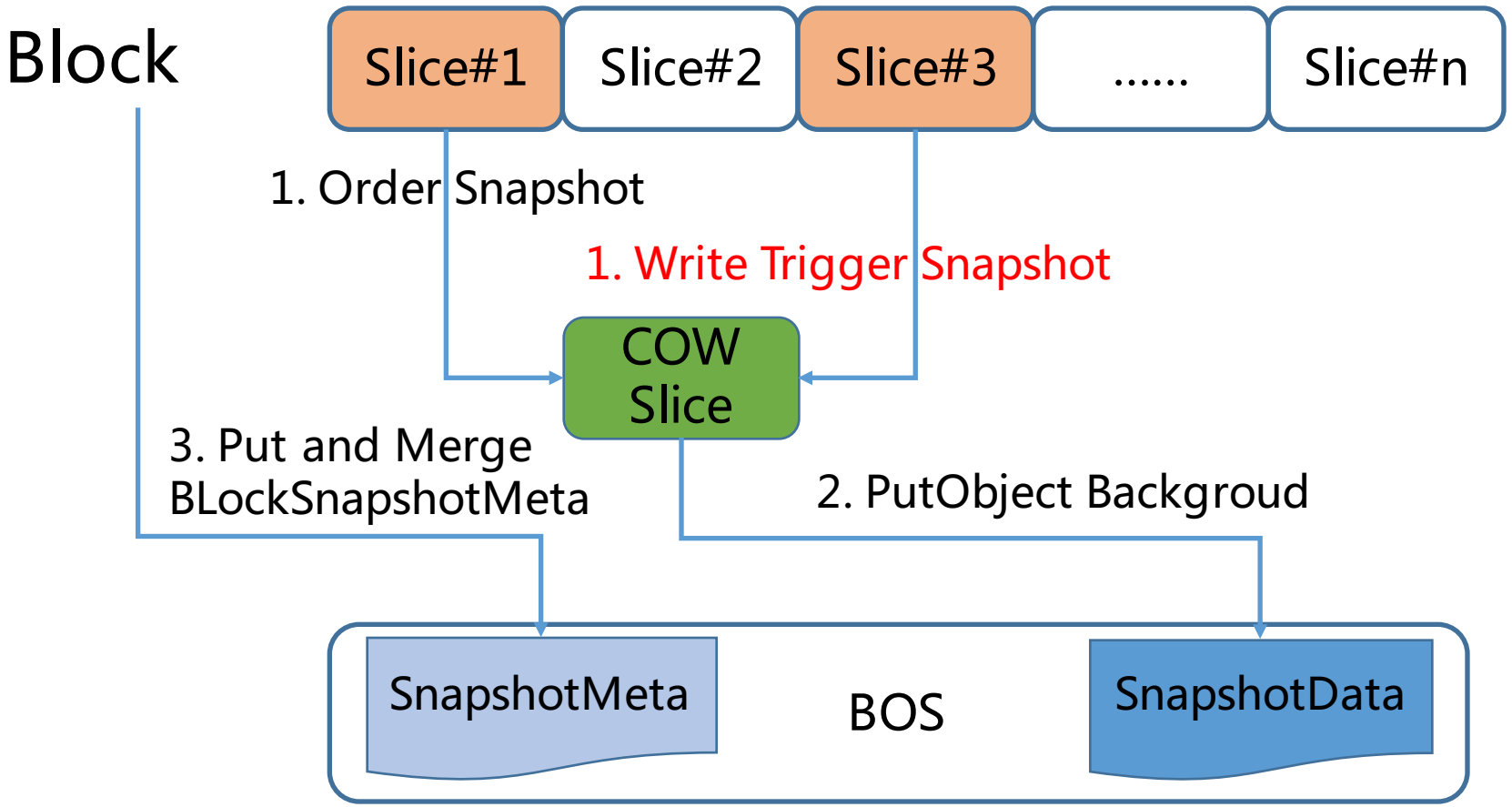
- Master
- BlockServer
- SnapshotWorker
- Executor



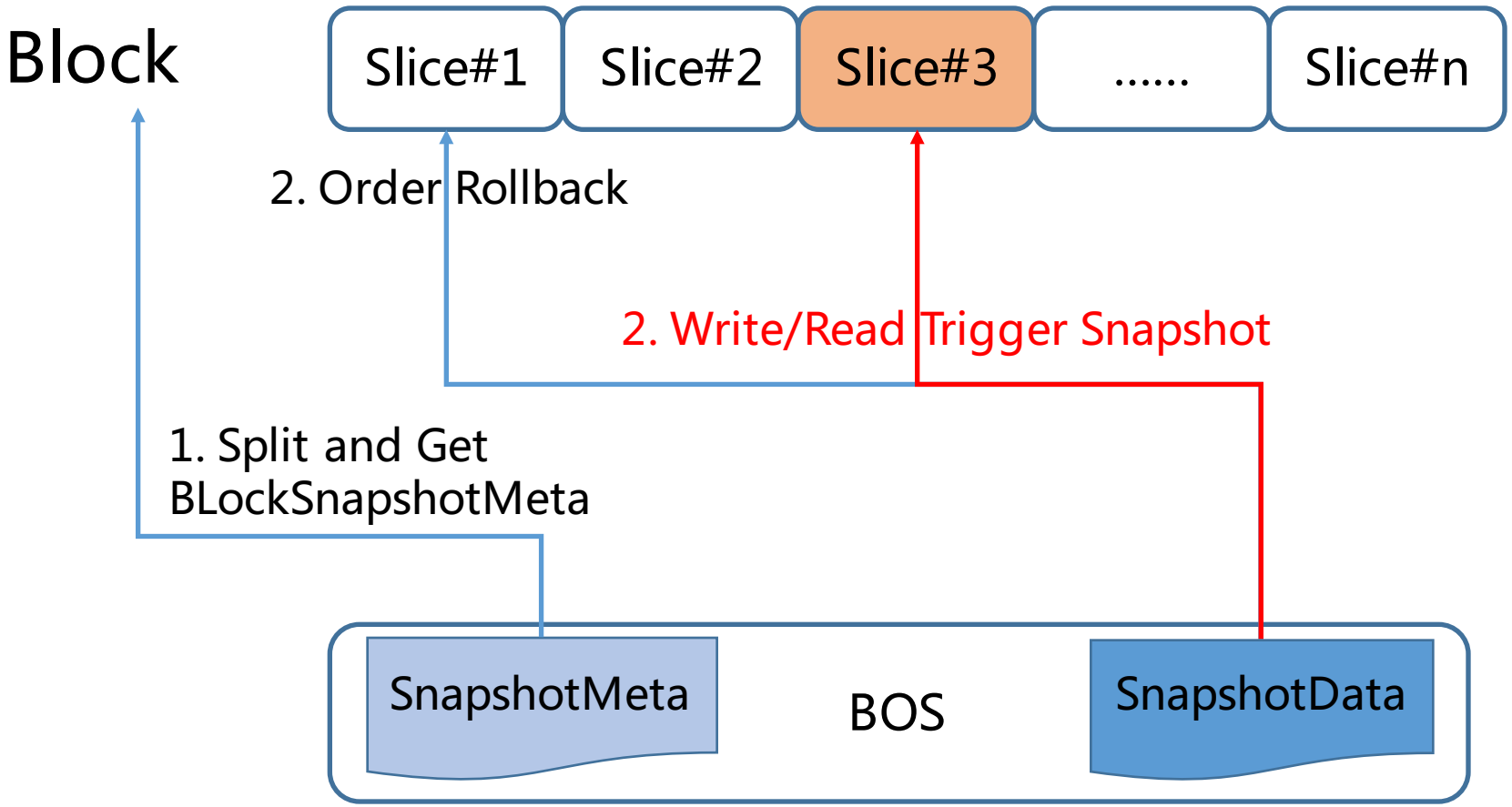
- Block分Slice
- 多版本Slice
  - 异步Snapshot
  - 异步Rollback



CDS快照

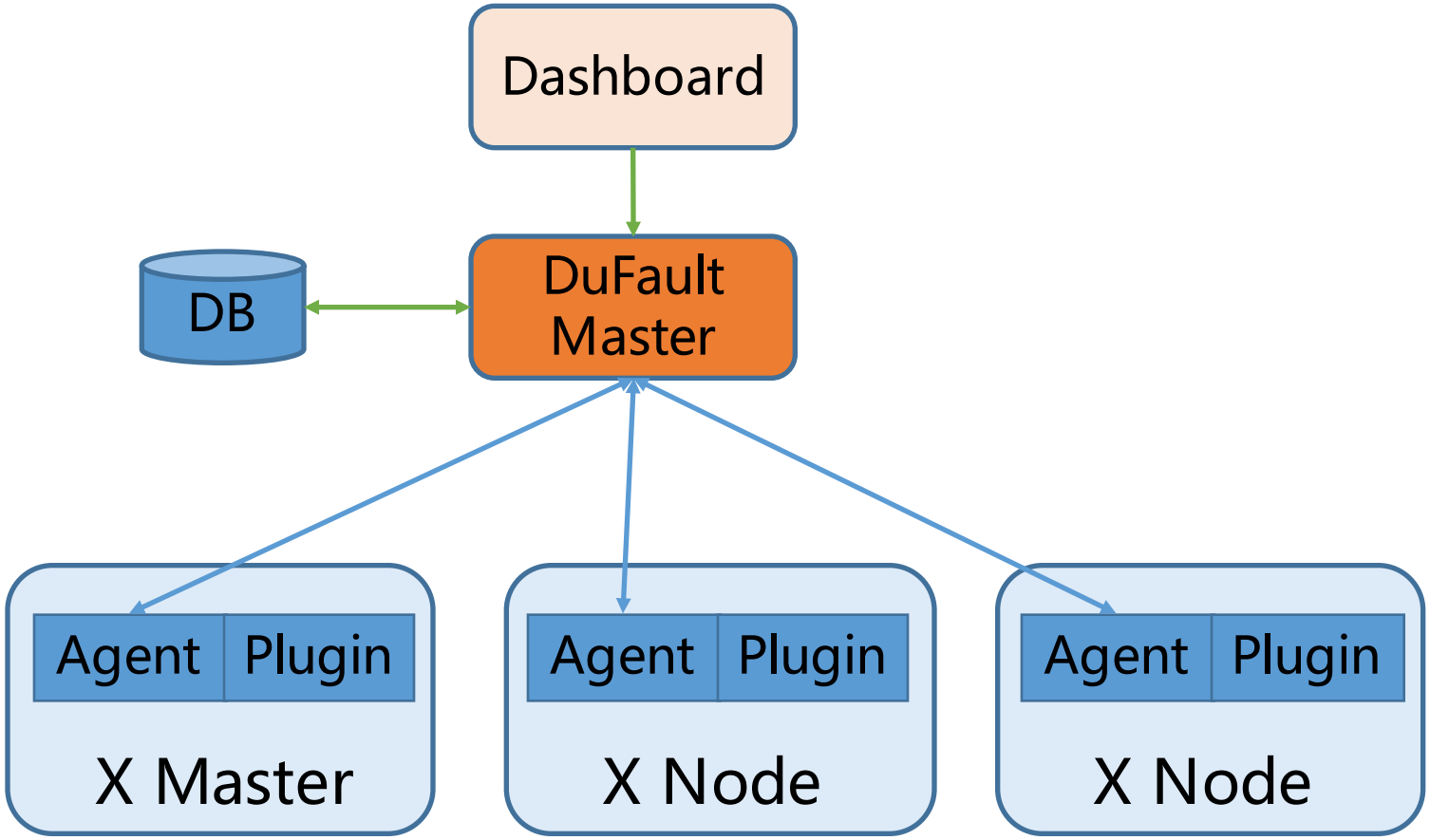

Block
Slice#1
Slice#2
Slice#3
.....
Slice#n
1. Order Snapshot
1. Write Trigger Snapshot
COW
Slice
3. Put and Merge
BLockSnapshotMeta
2. PutObject Backgroud
SnapshotMeta
BOS
SnapshotData

CDS回滚


Block
Slice#1
Slice#2
Slice#3
.....
Slice#n
2. Order Rollback
2. Write/Read Trigger Snapshot
1. Split and Get
BLockSnapshotMeta
SnapshotMeta
BOS
SnapshotData

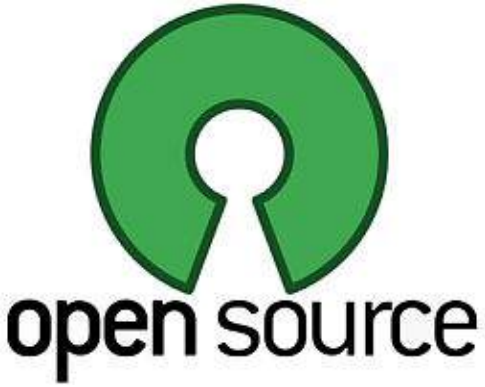
DuFault异常测试

- CPU/Mem异常
- Disk异常
- 进程异常
- 网络异常



广告

- 开源
  - https://github.com/brpc/brpc
  - https://github.com/brpc/braft



- 招聘
  - 分布式系统研发工程师
  - 虚拟网络研发工程师
  - 存储系统研发工程师

acut1-jobs@baidu.com

THANK YOU

cloud.baidu.com

主办方 Geekbang. 极客邦科技 InfoQ

GMTC 2018

全球大前端技术大会

大前端的下一站

<<扫码了解更多详情>>

极客邦科技
企业培训与咨询

Geekbang.

扫码关注

获取更多培训信息