ECUG Con 十周年盛会

Raft在百度云存储实践

王耀

2017/01/15

About Me

- 王耀
- 百度云高级架构师
- 资深轮子党
- 分布式存储熟练工

ABC时代的分布式存储

容量

EB级存储需求

每日新增百PB数据

数据长期备份

性能

高吞吐

低延时

性能横向扩展

多样性

内容：网页、广告、日志、UGC

类型：文本、图片、视频

形式：结构化、半结构化、非结构化

百度云存储

- 分类
  - 消息队列
  - 文件系统
  - 块存储
  - 对象存储
  - 表格存储

CCDB存储体系

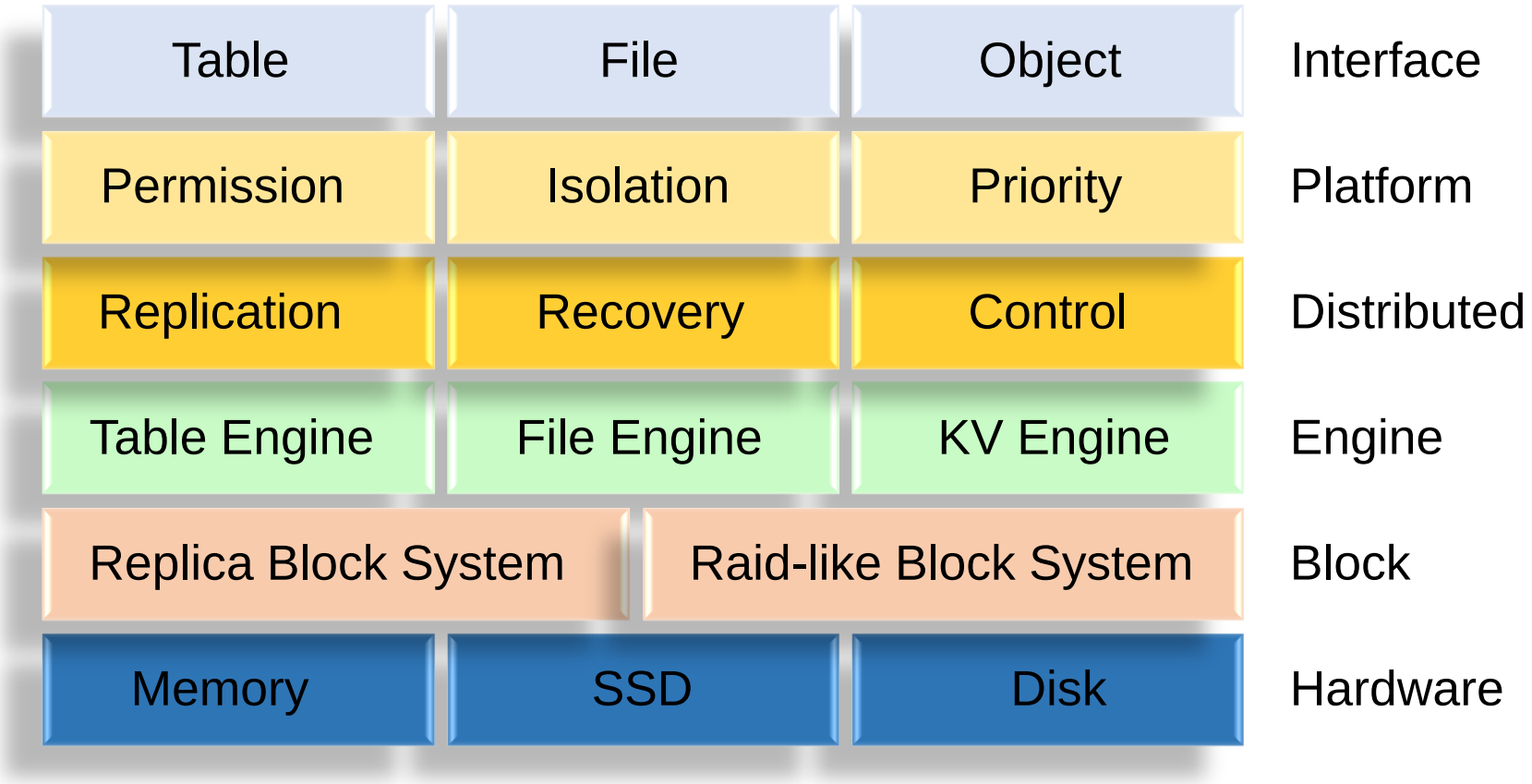


Table
File
Object
Interface
Permission
Isolation
Priority
Platform
Replication
Recovery
Control
Distributed
Table Engine
File Engine
KV Engine
Engine
Replica Block System
Raid-like Block System
Block
Memory
SSD
Disk
Hardware

AFS新存储体系

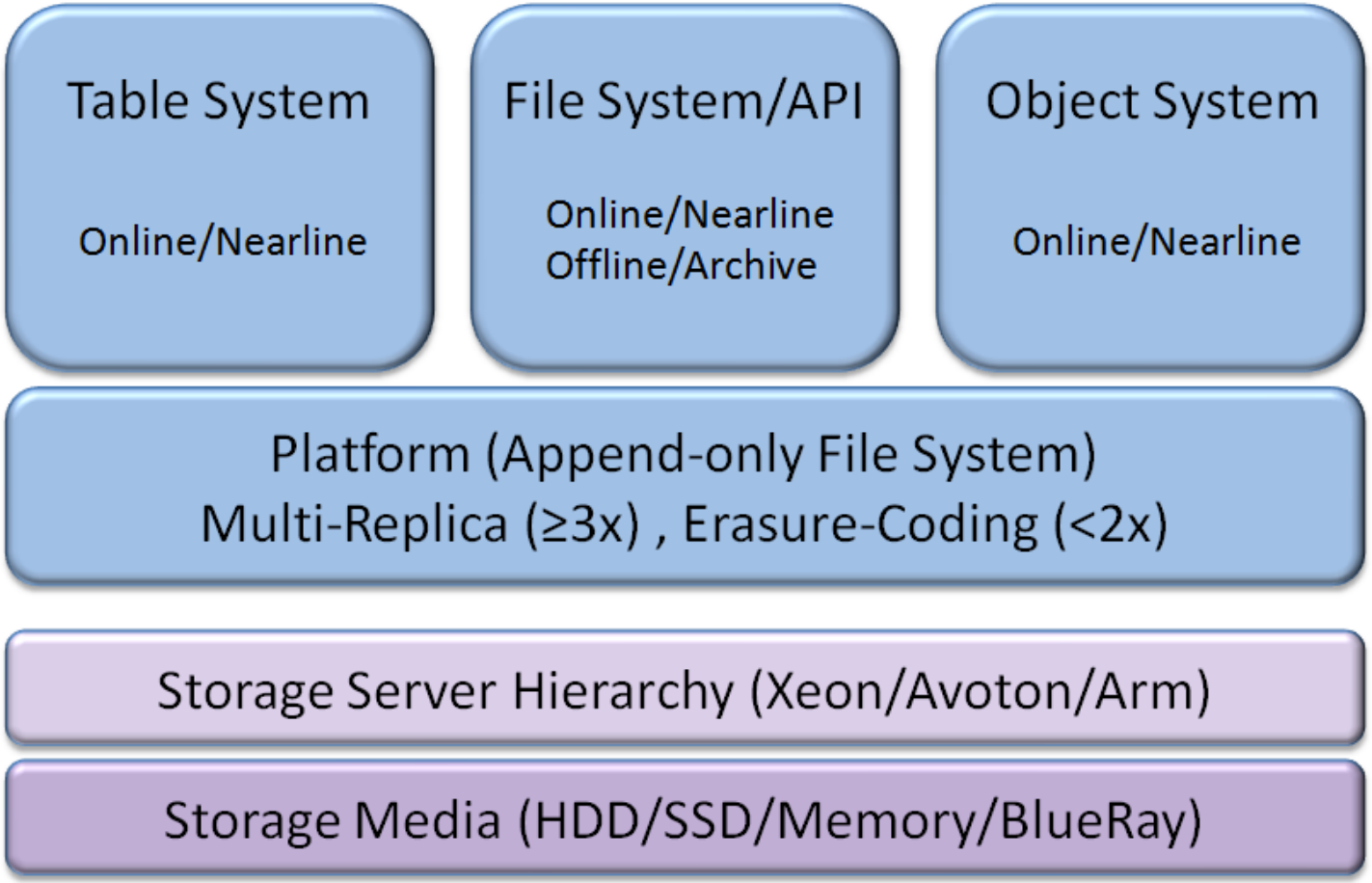
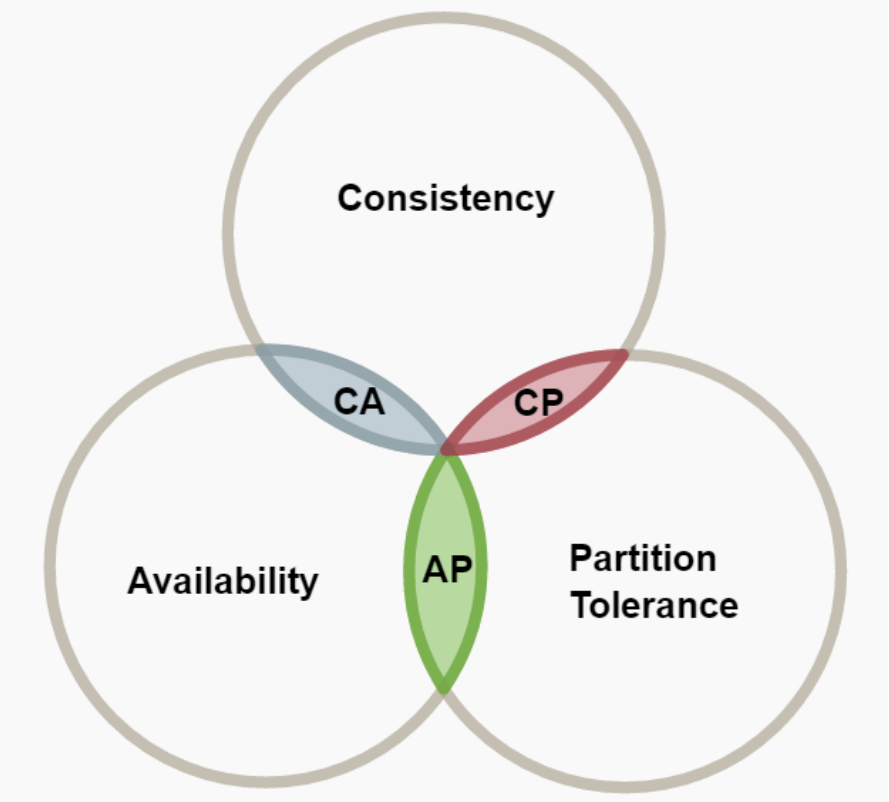
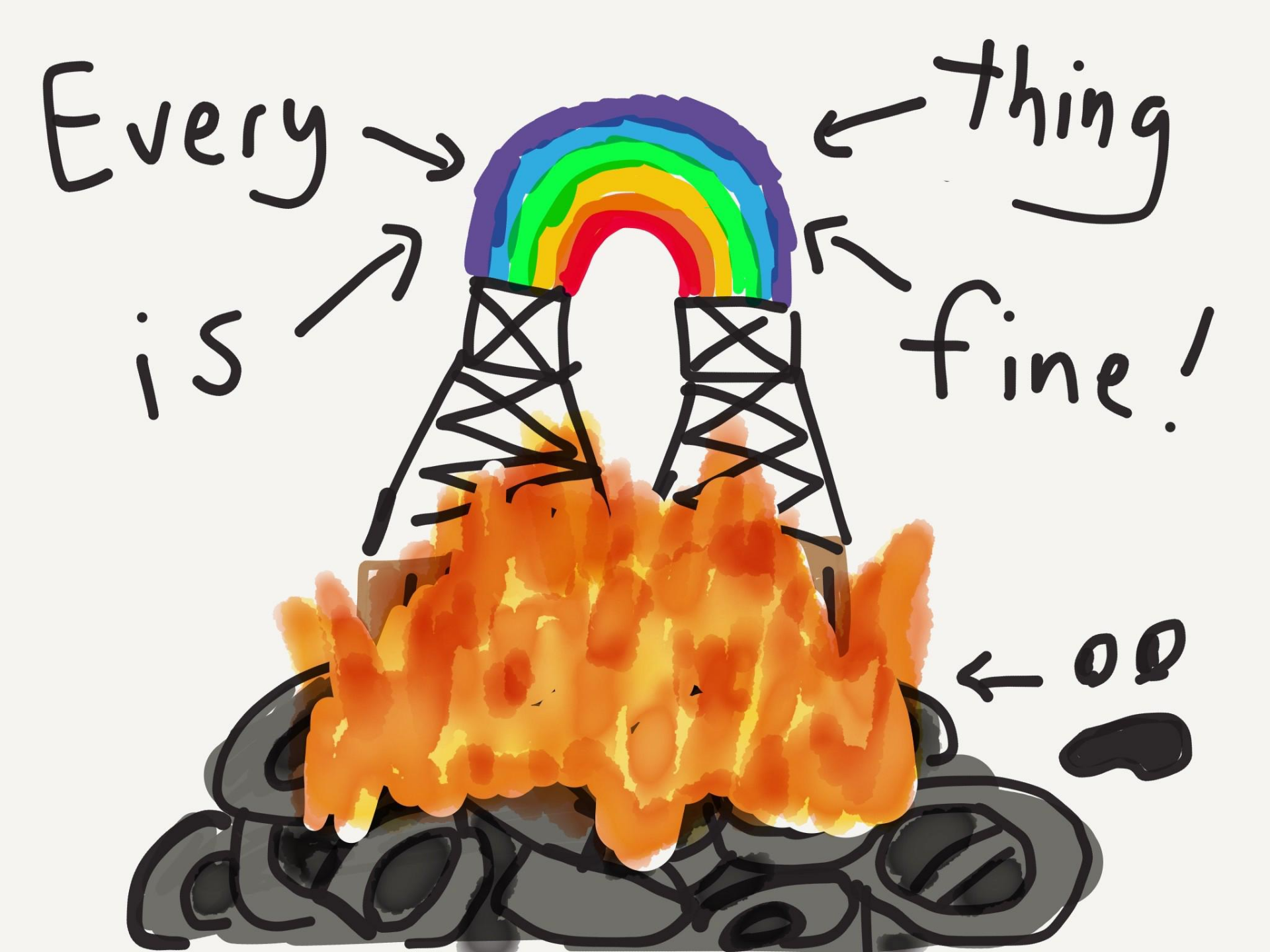


Table System
Online/Nearline
File System/API
Online/Nearline
Offline/Archive
Object System
Online/Nearline
Platform (Append-only File System)
Multi-Replica (≥3x) , Erasure-Coding (<2x)
Storage Server Hierarchy (Xeon/Avoton/Arm)
Storage Media (HDD/SSD/Memory/BlueRay)

分布式存储面临的问题

- 如何分片
- 如何复制
- 如何修复
  - 节点加入
  - 节点离开
- 如何负载均衡
- 如何规避IO慢节点



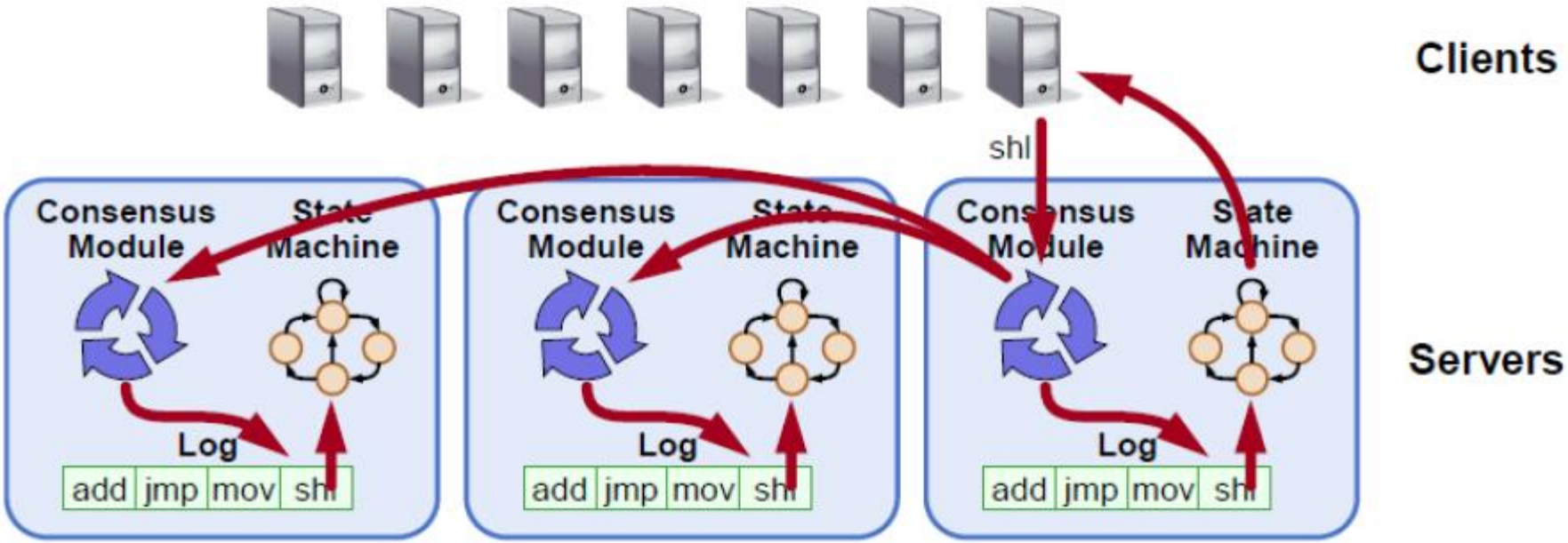

Every
thing
is
fine!
oil

一致性复制协议

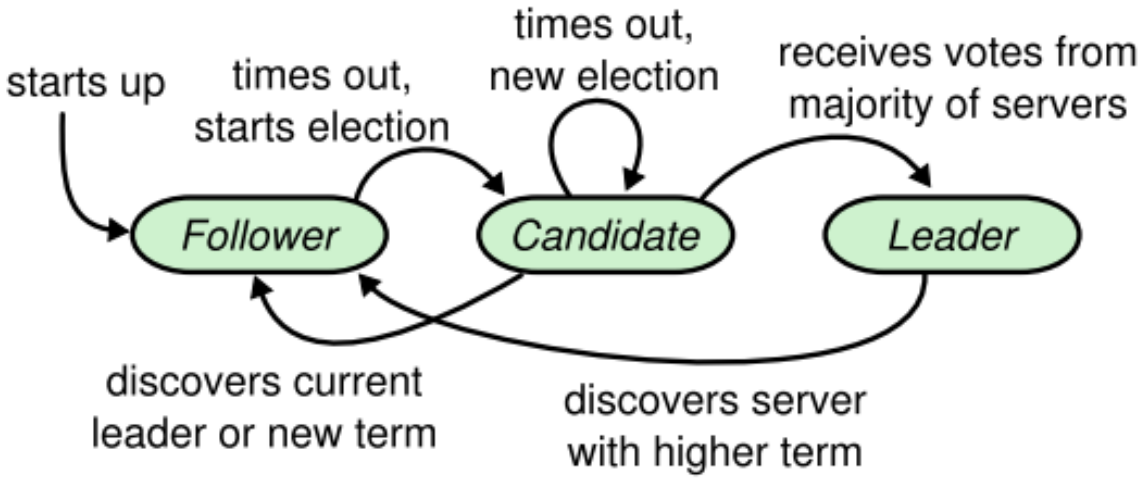
- Basic Paxos
- Multi Paxos
- Viewstamped Replication
- QJM
- ZAB

Raft简介

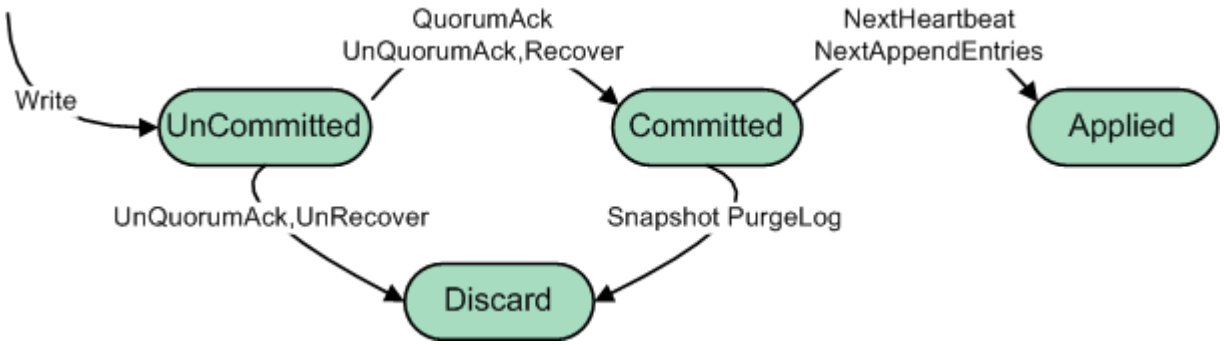
- Leader Election
- Log Replication
- Membership Change
- Log Compaction



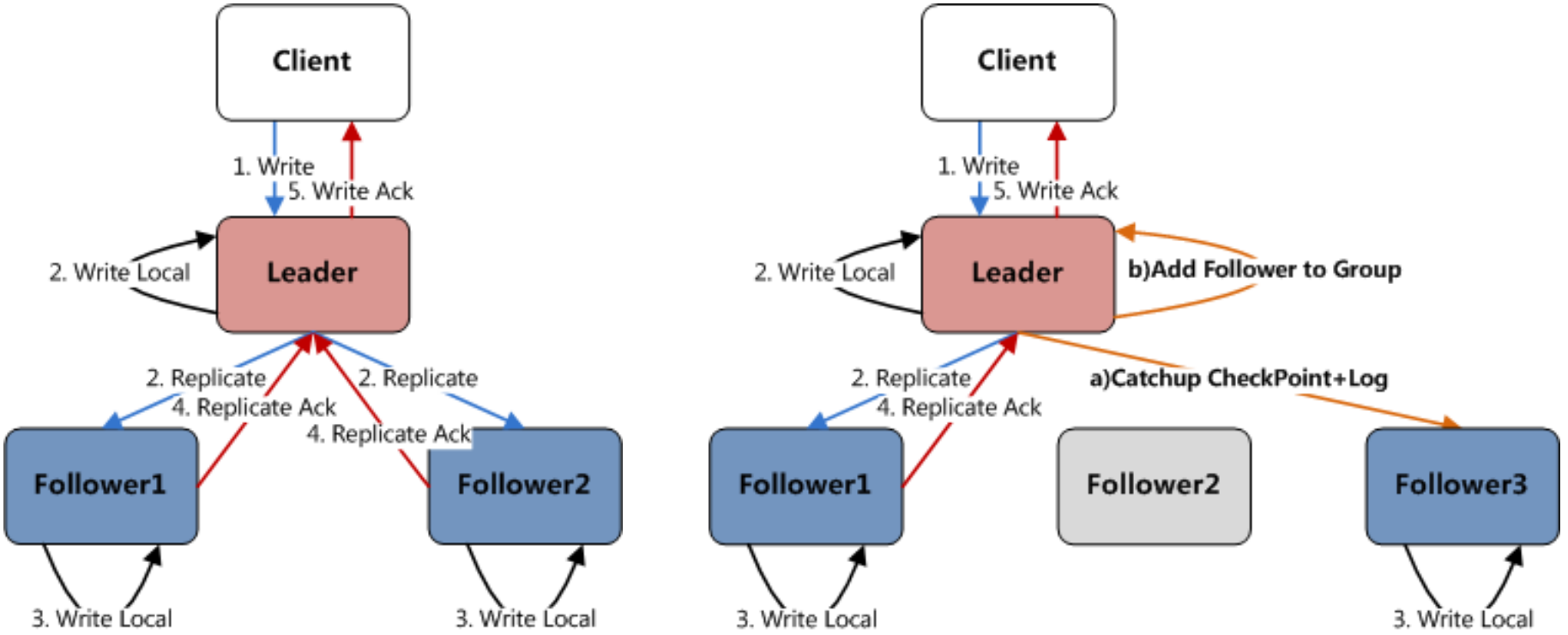
Raft之Node状态转移



Raft之Log状态转移

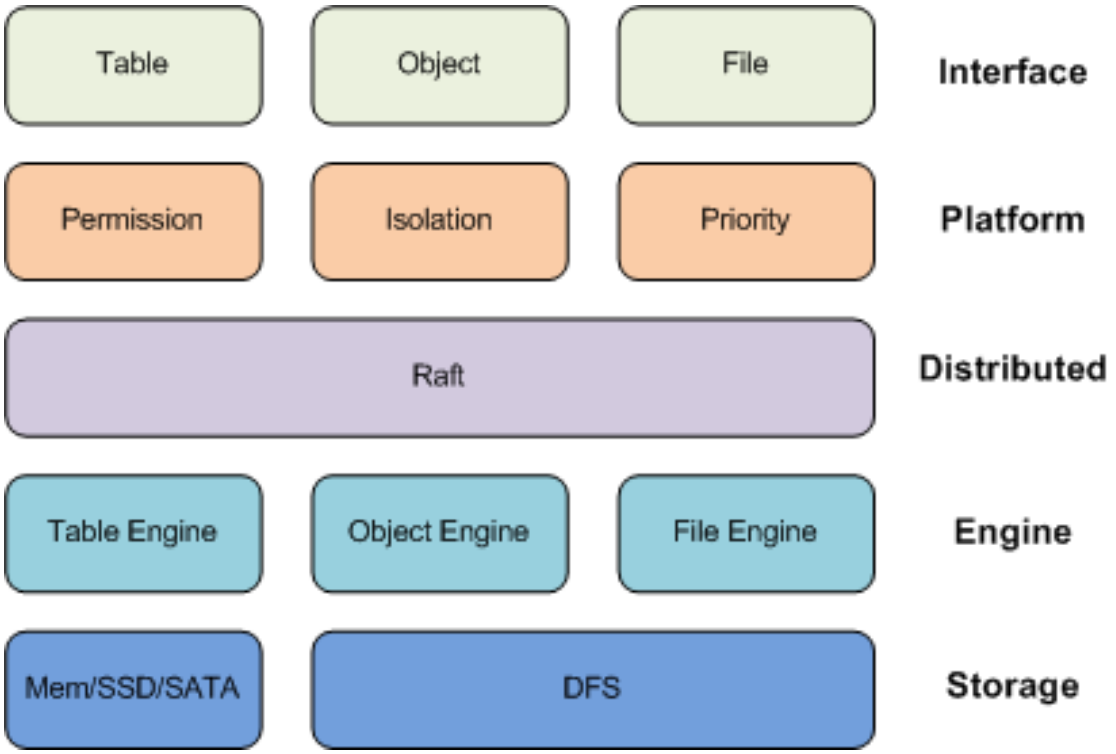


libraft之复制修复



raft在分布式存储

- Core Building
  - Lock
  - Block
  - Queue
  - Table
  - File
  - NewSQL

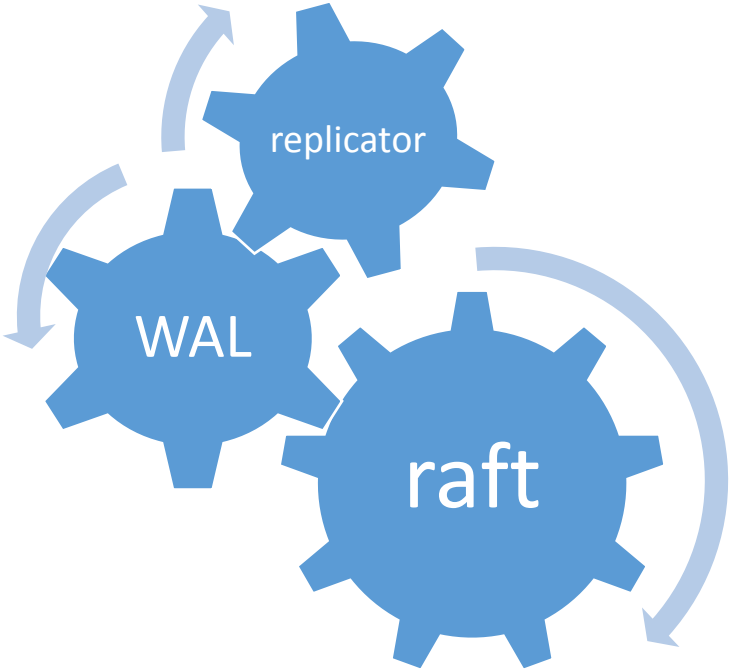


轮子libraft

- 业界现状
  - C++实现较少
  - 大部分类zk服务
  - 功能不完备
  - 性能不好
  - 测试不充分
- 需求目标
  - 高性能
  - 通用库
    - 自定义storage
  - 功能完备
    - prevote
    - leader transfer
  - 测试靠谱
    - jepsen test

libraft之WAL

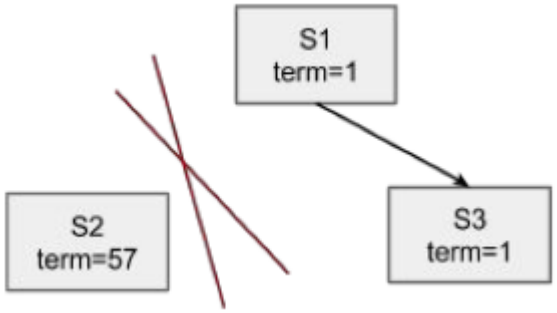
- 挑战
  - WAL的IO隔离
  - WAL阻塞Raft状态机
  - WAL双写影响吞吐



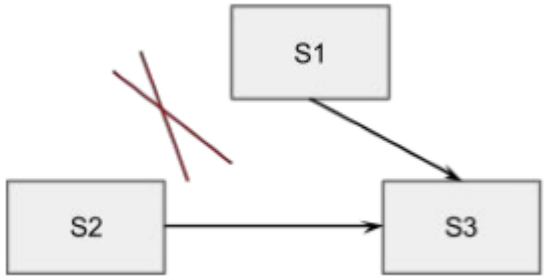
- 缓存
  - 内存缓存最近Entries
- 异步
  - WAL异步写
- 批量
  - Replicator批量发Entries
  - LogStorage批量写Entries

libraft之prevote

- 对称网络划分



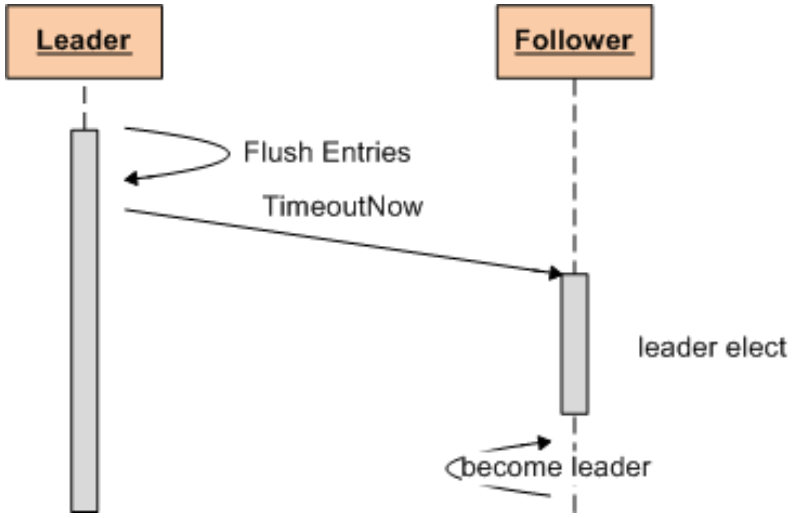
- 对称网络划分



- 增加term会导致leader stepdown
- prevote阻止数据不全节点选主
  - 不属于复制组中的节点
  - 属于复制组但网络划分的节点

libraft之leader transfer

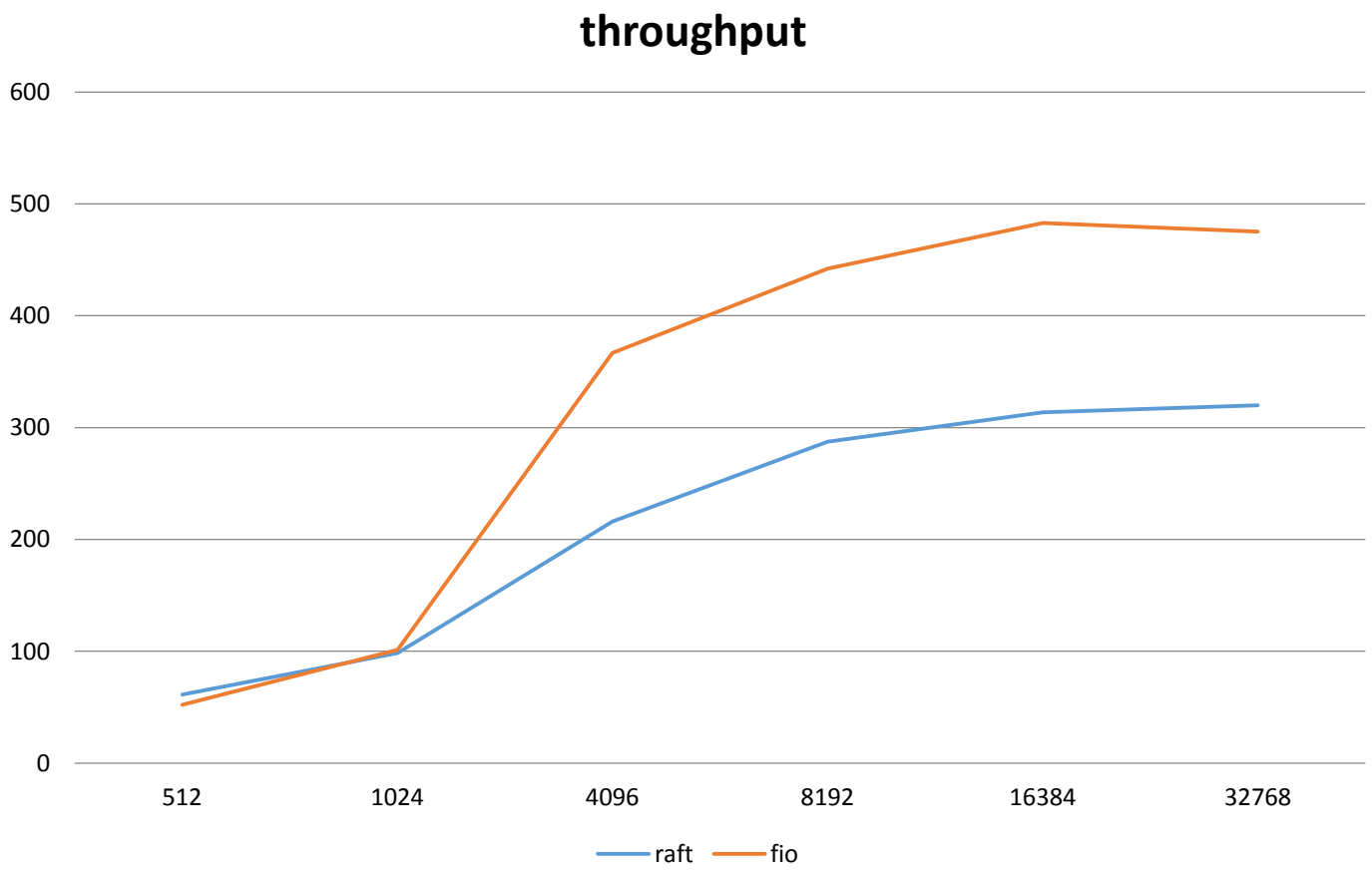
- Implement
  - TimeoutNow
- Case
  - rebalance
  - remove leader



libraft之tips

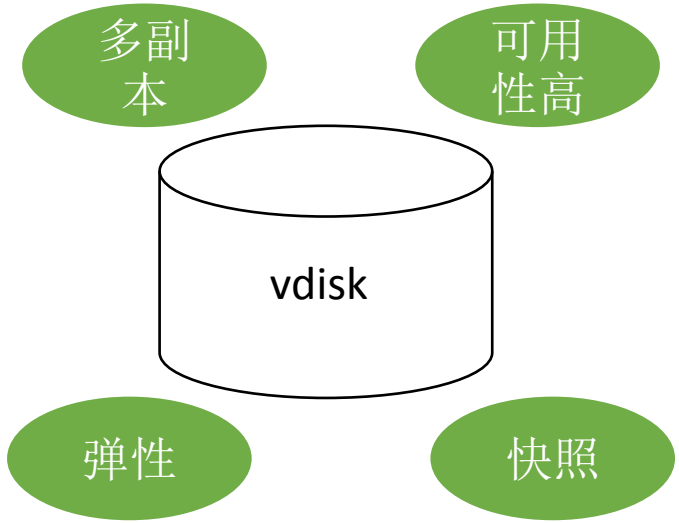
- on_snapshot_load开始先清空状态机
- on_apply保证主从执行结果一致
- on_leader_stop保证leader相关任务cancel
- proposal带上term保证非幂等操作的安全
- PeerId增加version机制

libraft之benchmark



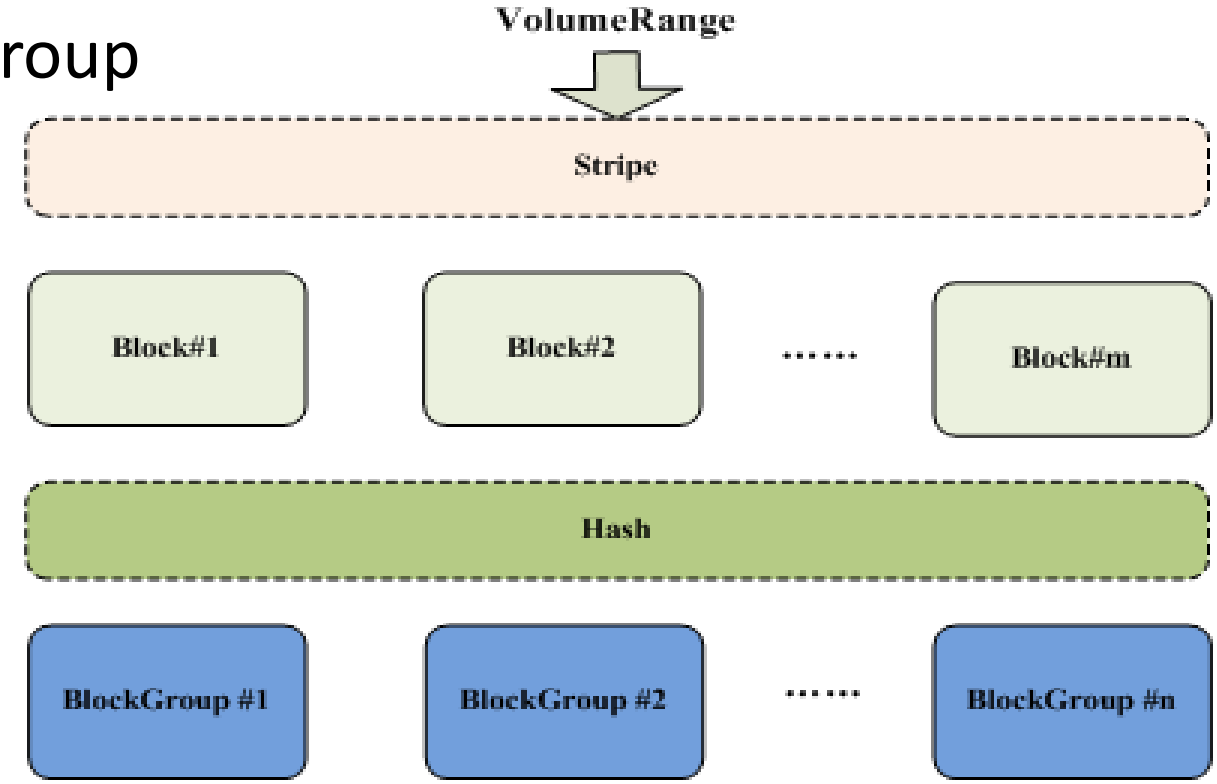
CDS简介

- 云磁盘服务
  - 为虚机提供可扩展的数据块级存储卷。
- 特性
  - 高可靠性
  - 高稳定性
  - 高性能



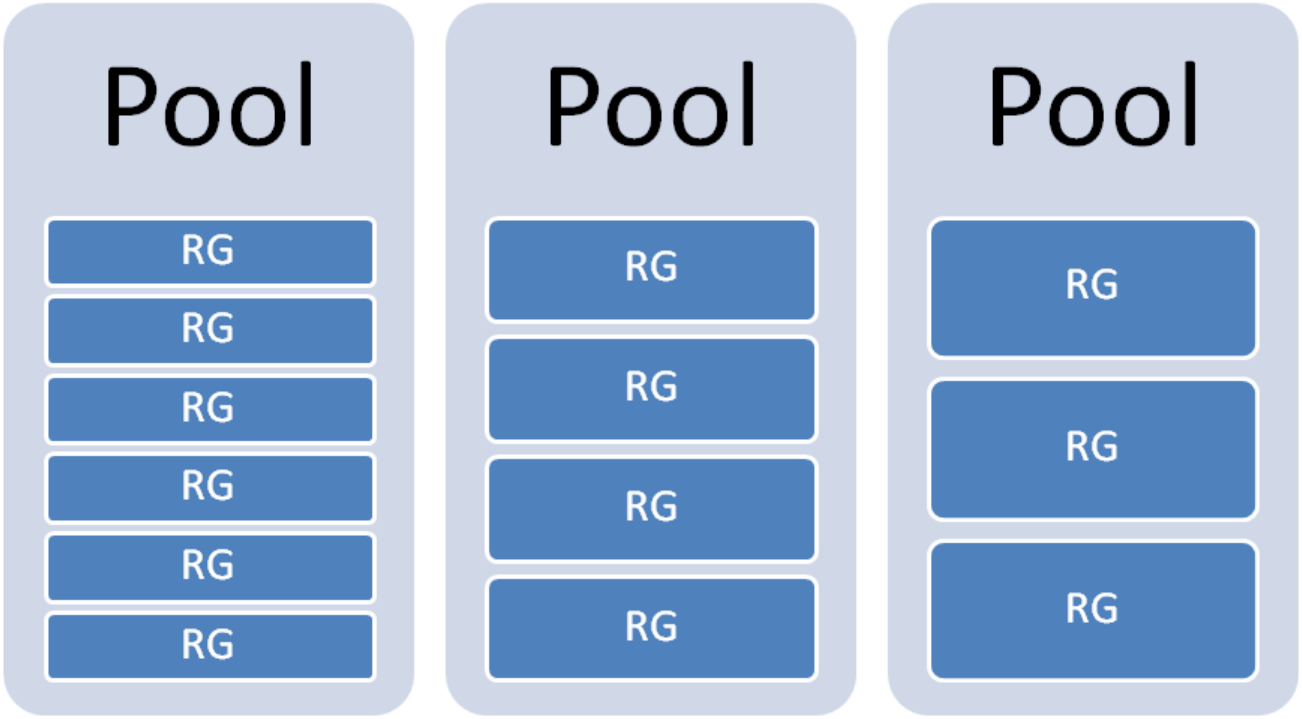
CDS数据模型

- Volume拆Block
- Block聚BlockGroup



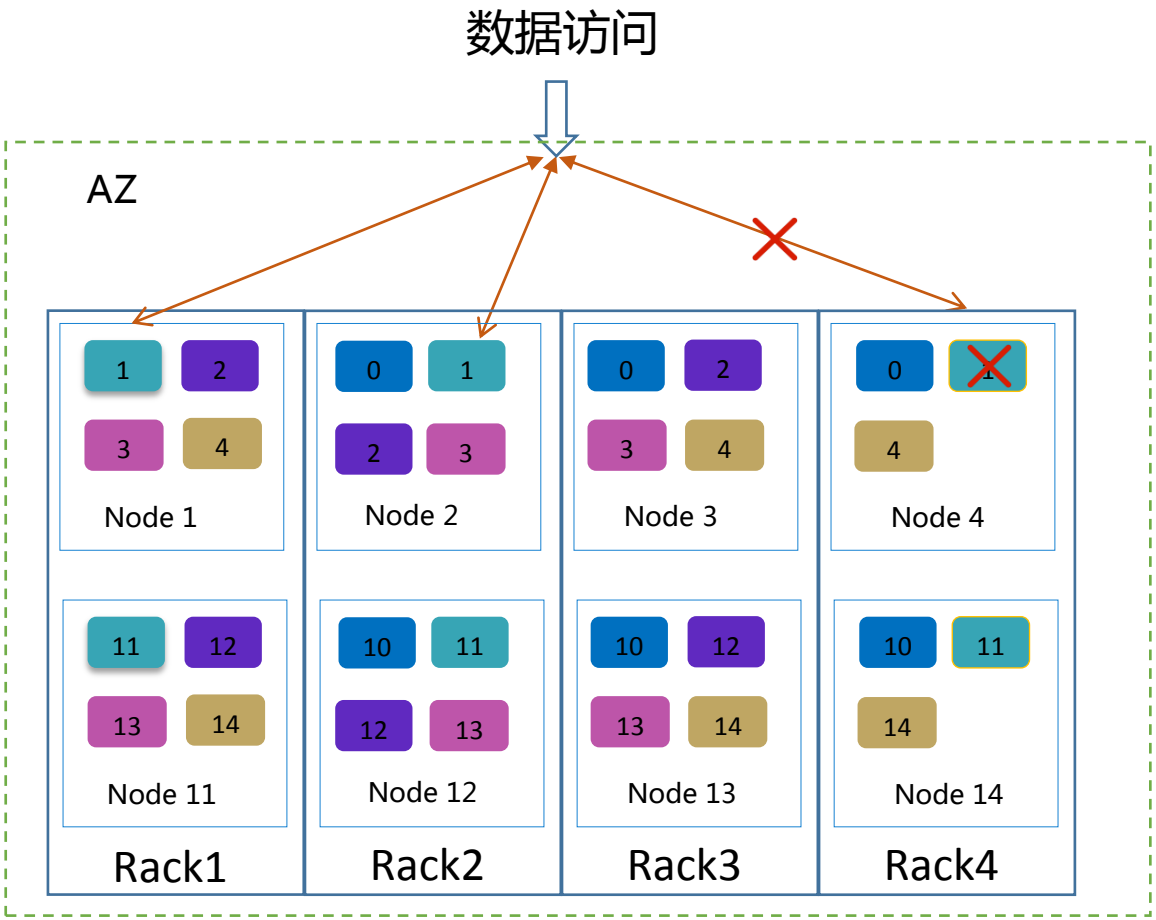
CDS逻辑数据分布

- 两级分布
  - Pool
  - ReplicaGroup

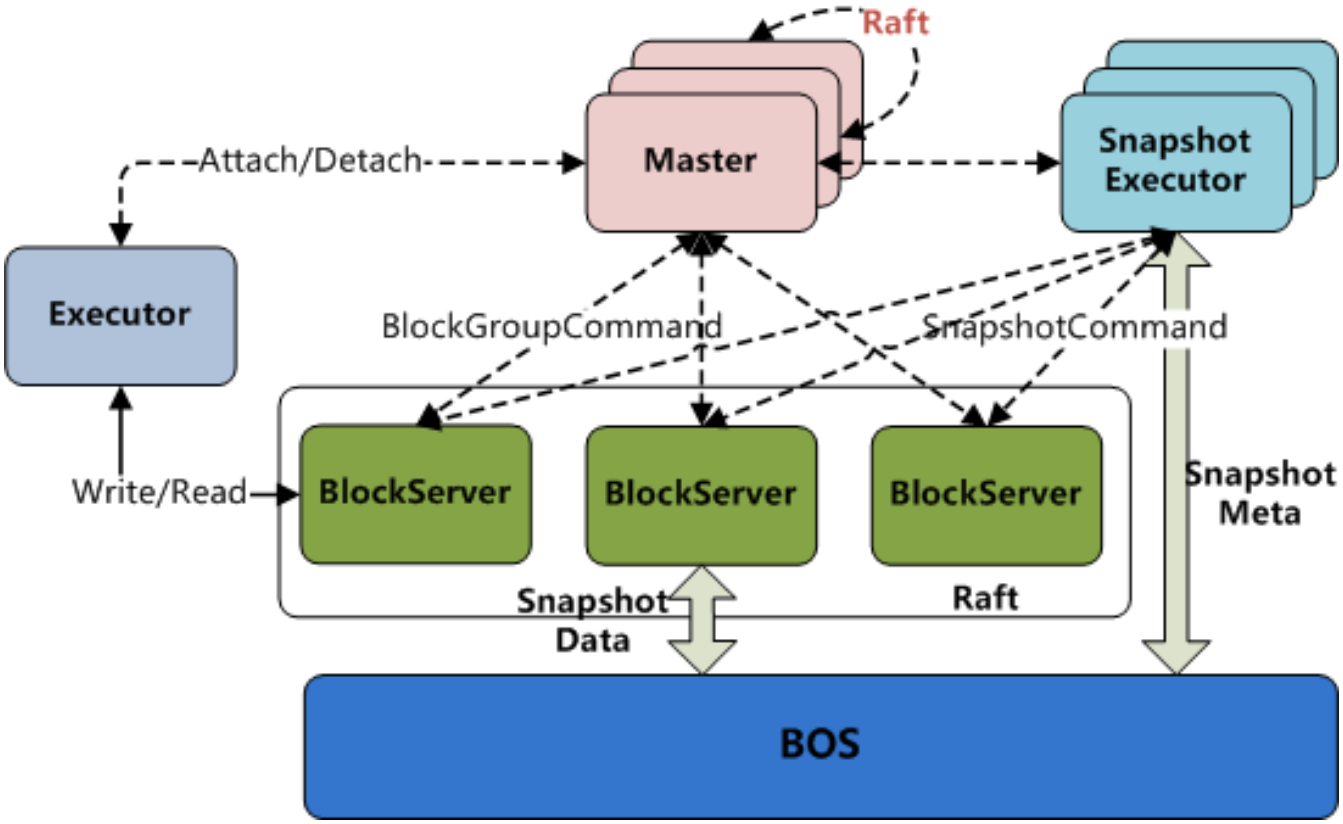


CDS物理数据分布

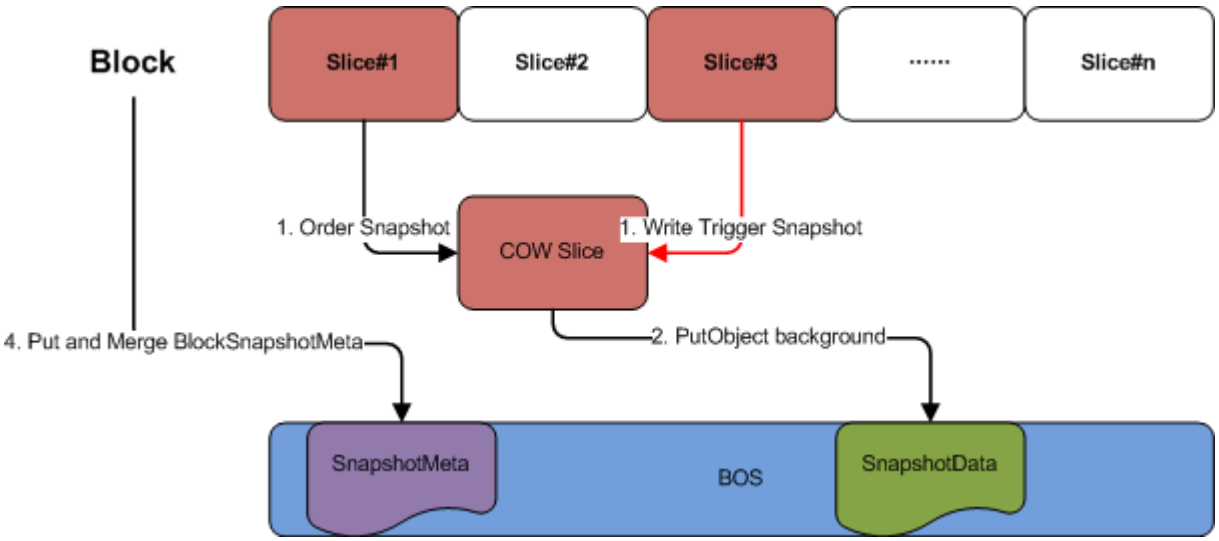
- 五级隔离
  - Region
  - Zone
  - Rack
  - Node
  - Disk



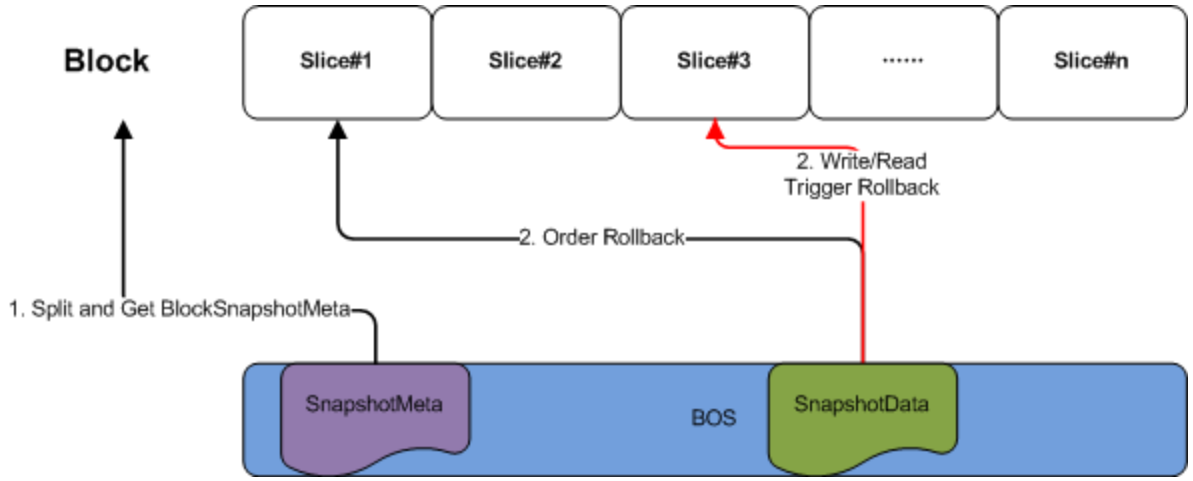
CDS架构


Raft
Master
Snapshot Executor
Attach/Detach
Executor
BlockGroupCommand
SnapshotCommand
Write/Read
BlockServer
BlockServer
BlockServer
Snapshot Meta
Snapshot Data
Raft
BOS

CDS快照


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

CDS回滚

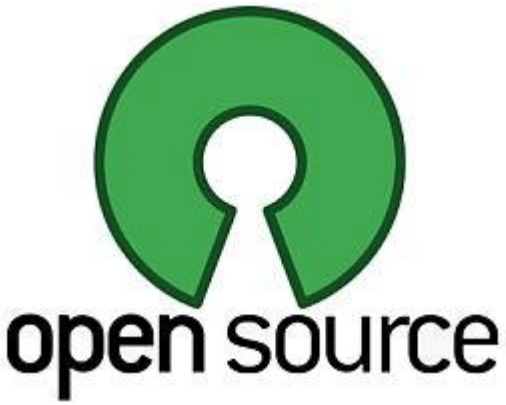

Block
Slice#1
Slice#2
Slice#3
.....
Slice#n
2. Write/Read
Trigger Rollback
2. Order Rollback
1. Split and Get BlockSnapshotMeta
SnapshotMeta
BOS
SnapshotData

CDS请求长尾优化

- quorum写入优化写请求
- backup request优化读请求
- 定期汇报进行主从均衡和数据迁移

即将开源

- bthread
- bvar
- baidu-rpc
- libraft



Q&A