

分布式KV存储Cellar演进之路

美团点评·基础架构 齐泽斌



个人简介

美团点评基础架构部，存储研发团队负责人

- Cellar : 分布式KV存储服务
- Databus : 数据库变更实时传输服务
- Venus : 图片服务

11年毕业于天津大学

11 年到 14 年任职于百度，负责分布式文件系统和 KV 存储系统研发
有多年分布式存储研发经验



目录

- Cellar起源
- 中心节点架构演进
- 节点高可用和异地容灾
- 服务可用性提升
- Cellar规划

目录

- Cellar起源
- 中心节点架构演进
- 节点高可用和异地容灾
- 服务可用性提升
- Cellar规划

Cellar起源

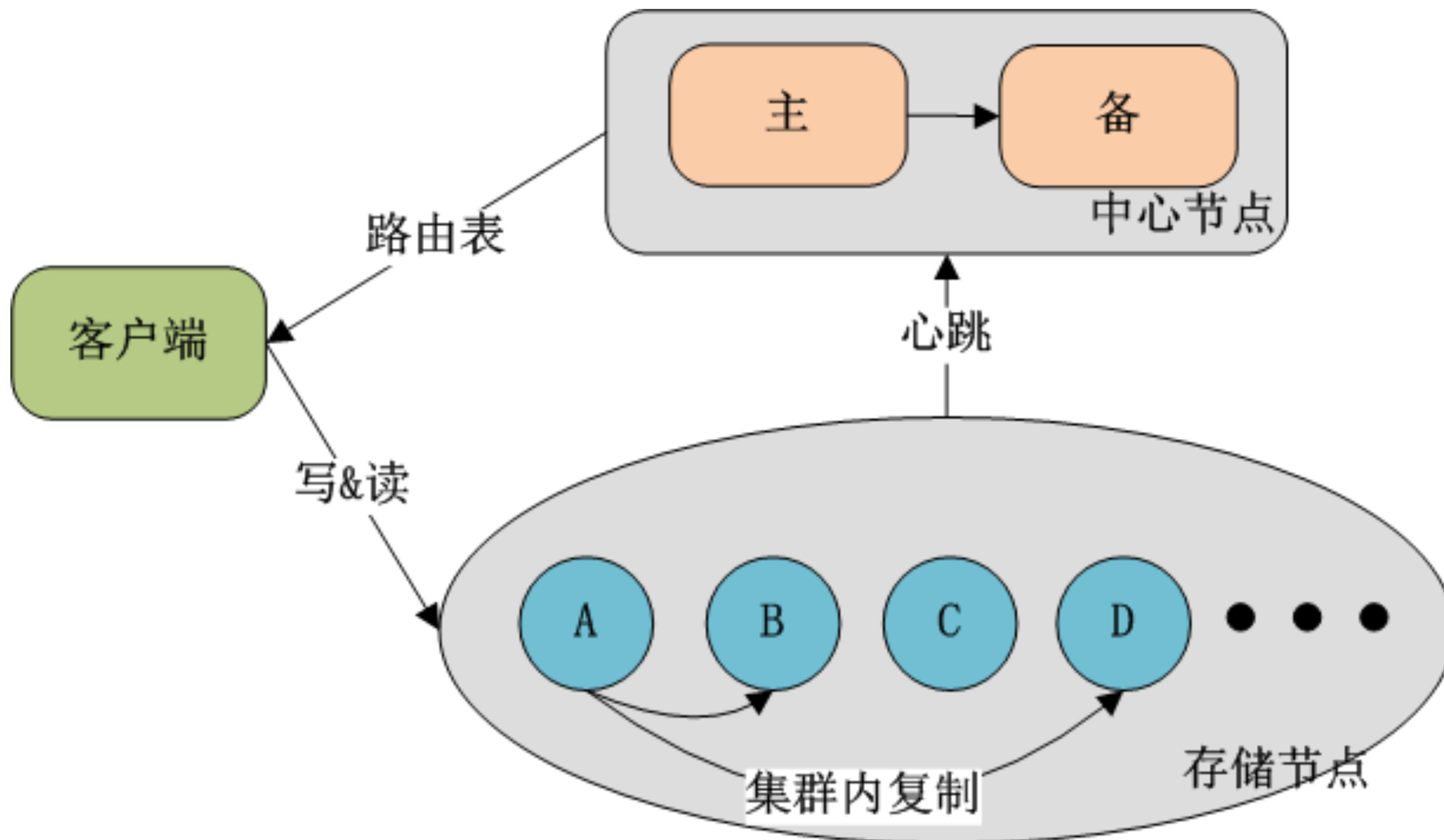
Cellar，英文原意是酒窖，项目取名Cellar，一方面借用其储藏之意，同时，也希望使用Cellar的用户，可以像用酒窖藏酒一样，越存越香。

Cellar起源

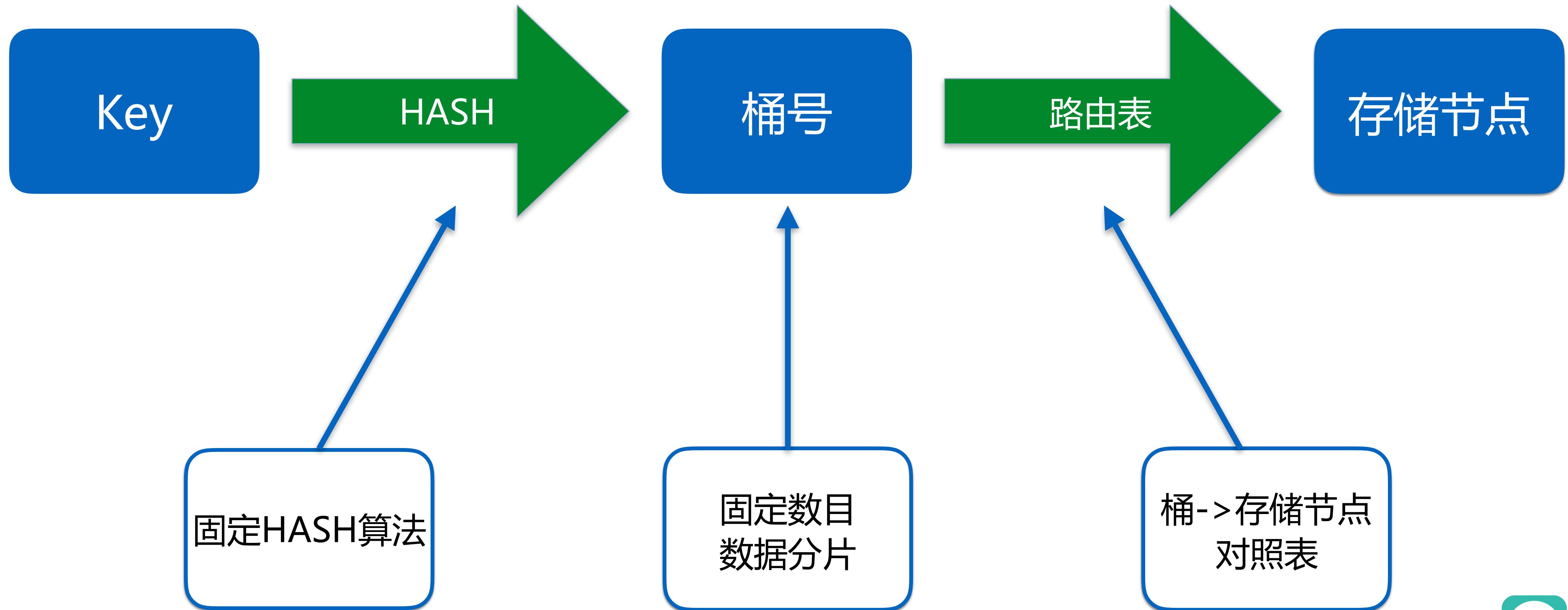
- 14年初 美团引入阿里Tair作为NoSQL存储
- 14年底 大范围应用，并对Tair修修补补，积累领域问题
- 16年初 基于开源版本研发新一代KV存储系统Cellar
- Now Cellar日请求量达万亿级，美团点评最大NoSQL存储



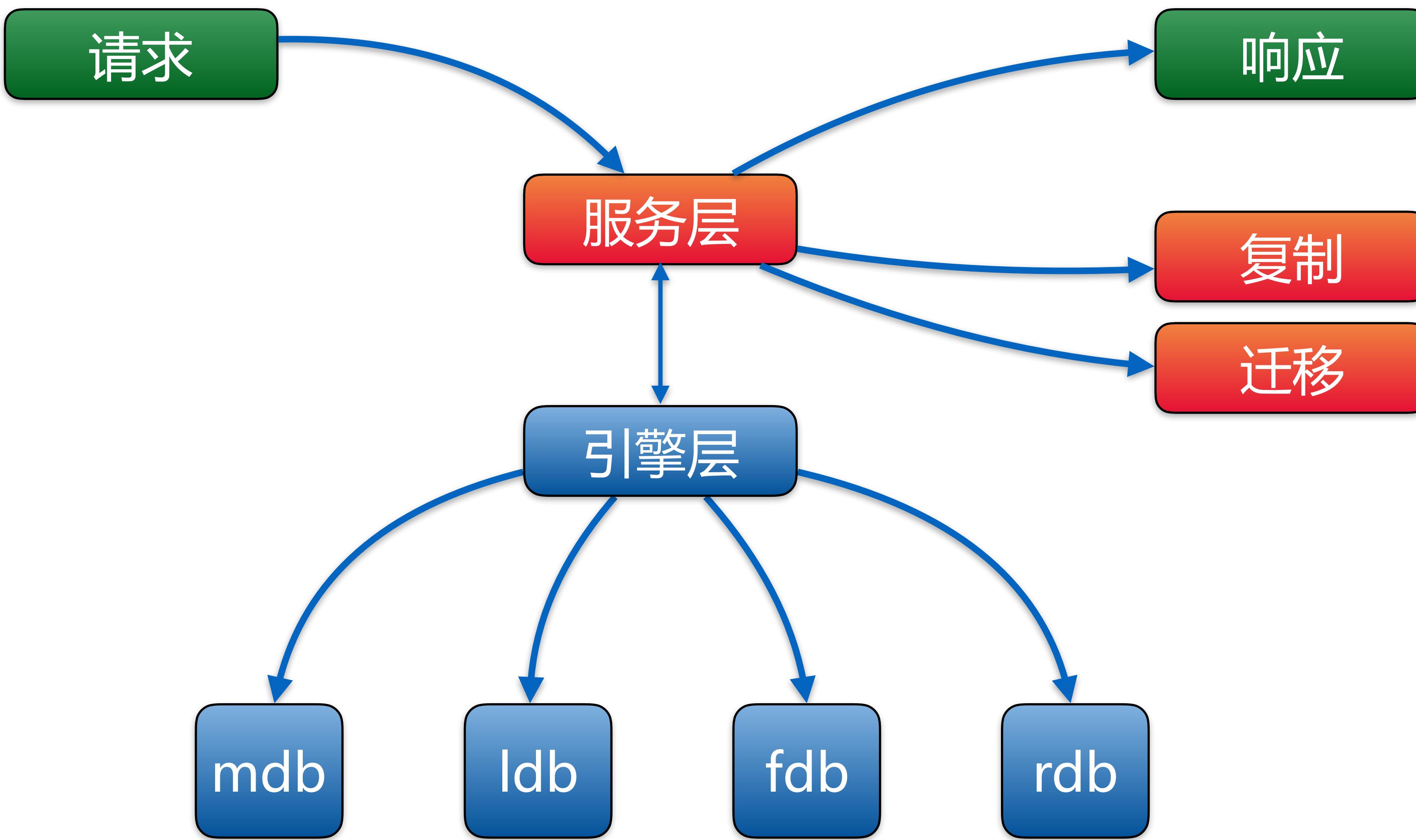
Cellar起源—Tair架构



Cellar起源—Tair架构



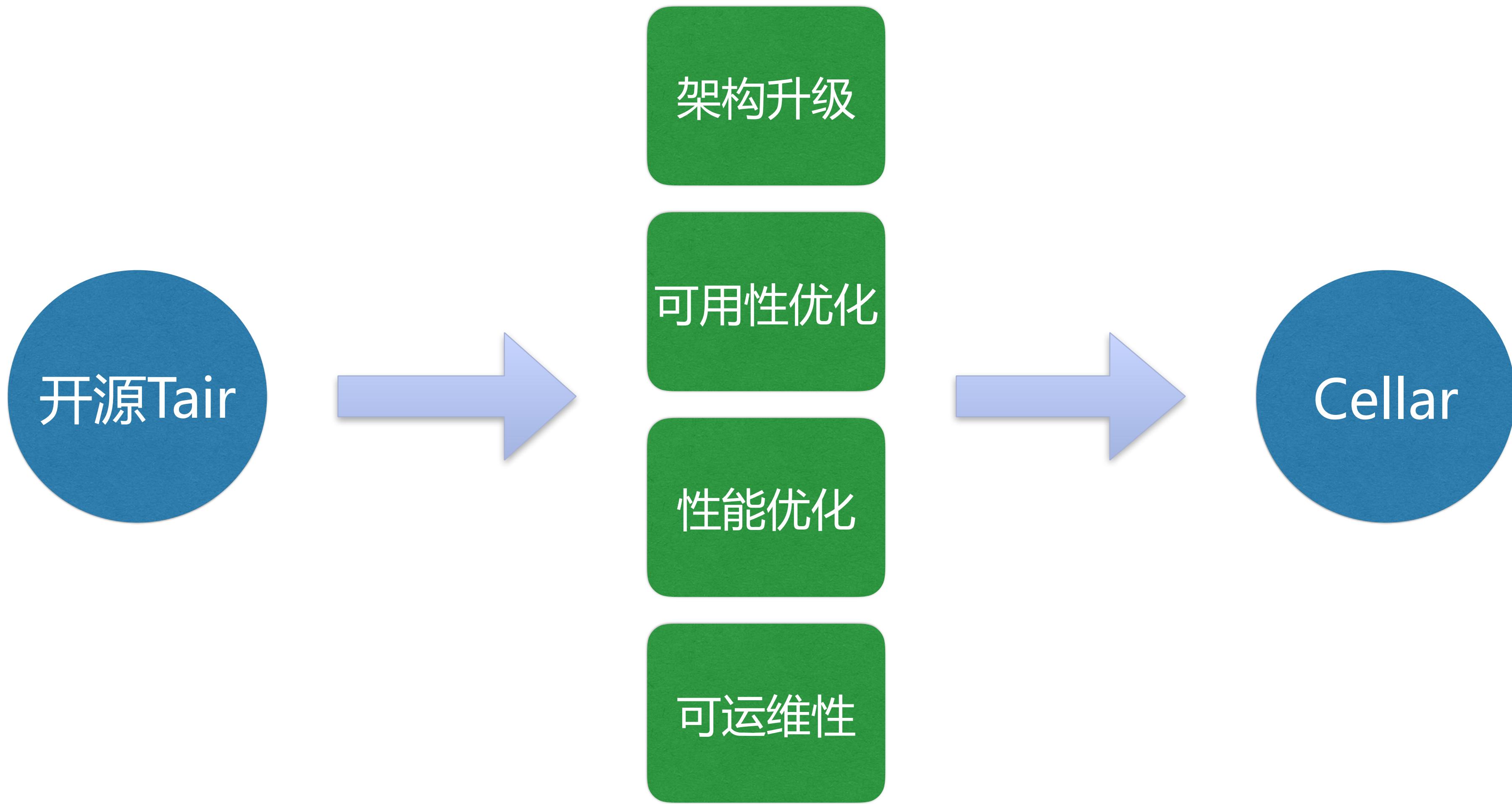
Cellar起源—Tair架构



Cellar起源—Tair问题

- 中心化集群问题
- 可用性问题
- 性能问题
- 运维问题

Cellar起源



目录

- Cellar起源
- 中心节点架构演进
- 节点高可用和异地容灾
- 服务可用性提升
- Cellar规划

Cellar—中心节点架构演进

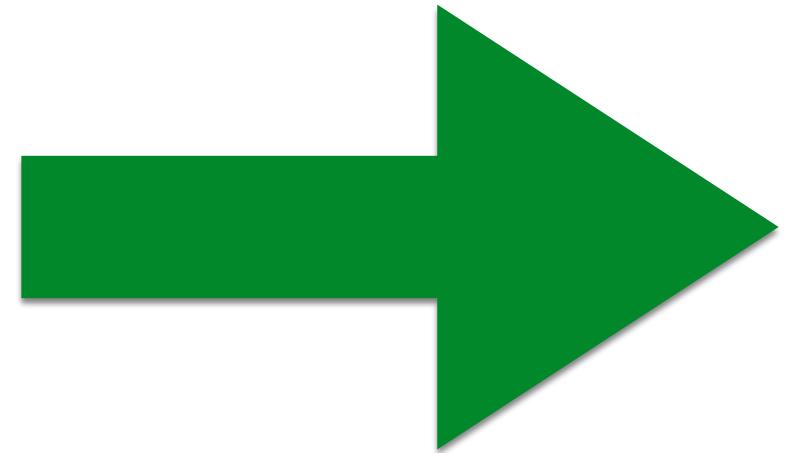
Cellar—中心节点架构演进

- 性能问题
客户端集中获取路由表
- 隔离性问题
中心节点暴露给客户端

Cellar—中心节点架构演进

- 性能问题

客户端集中获取路由表



- 隔离性问题

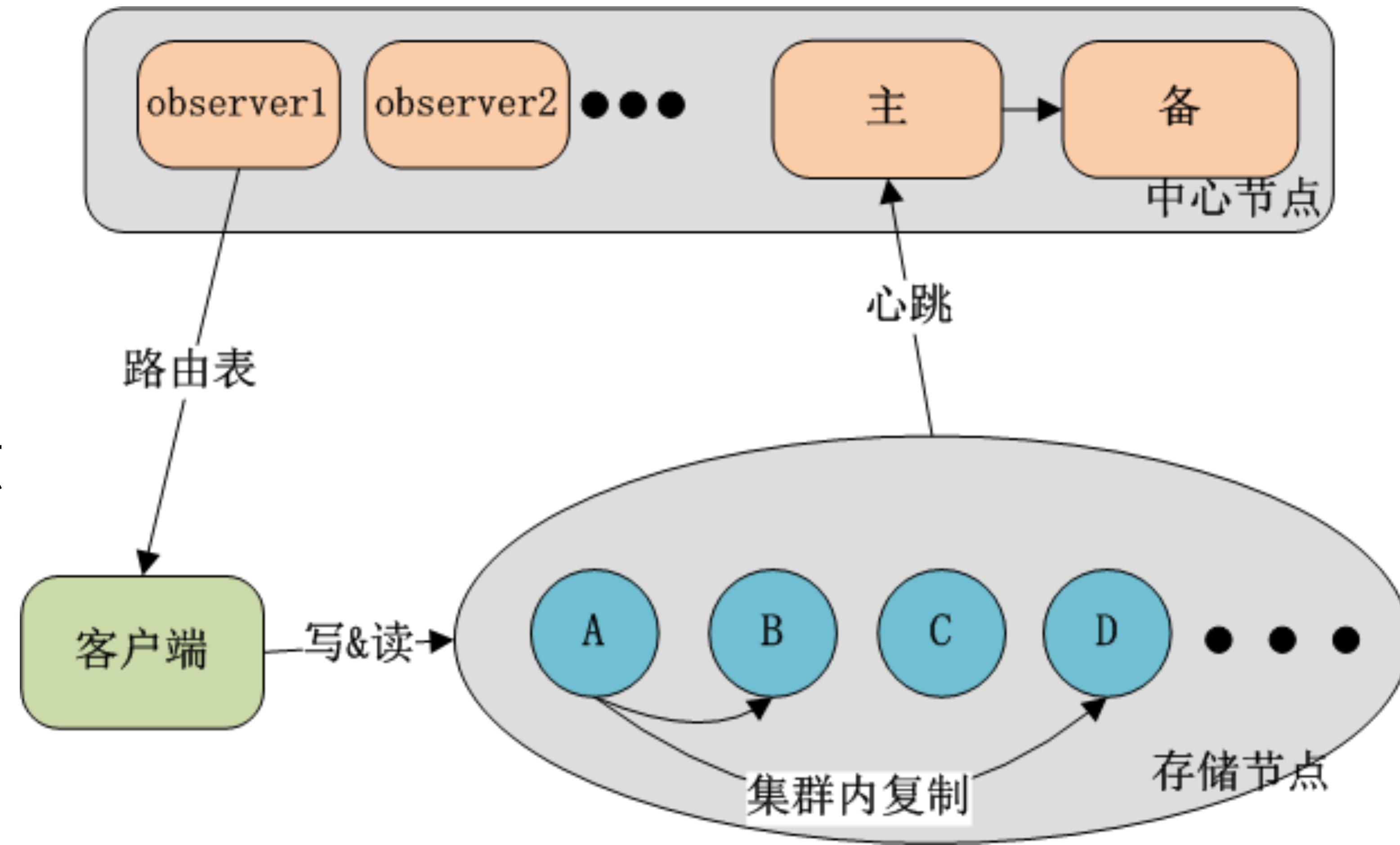
中心节点暴露给客户端

单独的路由表获取模块

Cellar—中心节点架构演进

- 可扩展性：
路由查询能力
可线性扩展

- 隔离性：
客户端与中心节点
完全隔离



Cellar—中心节点架构演进

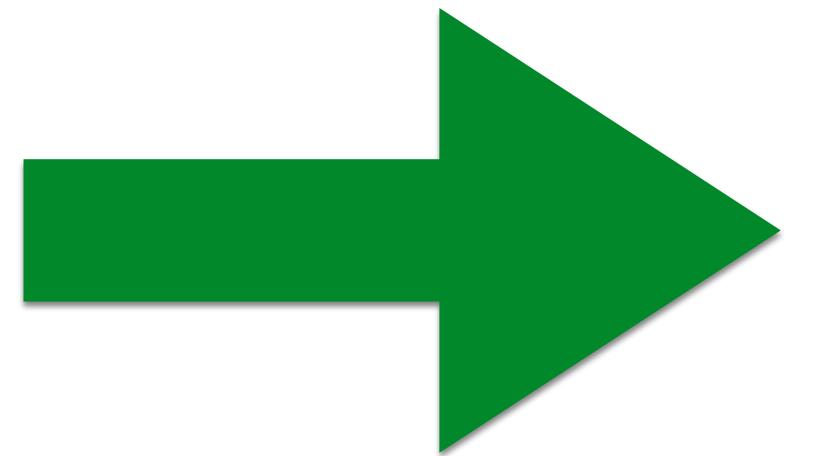
一致性

- 主备脑裂
- observer与config

Cellar—中心节点架构演进

一致性

- 主备脑裂
- observer与config

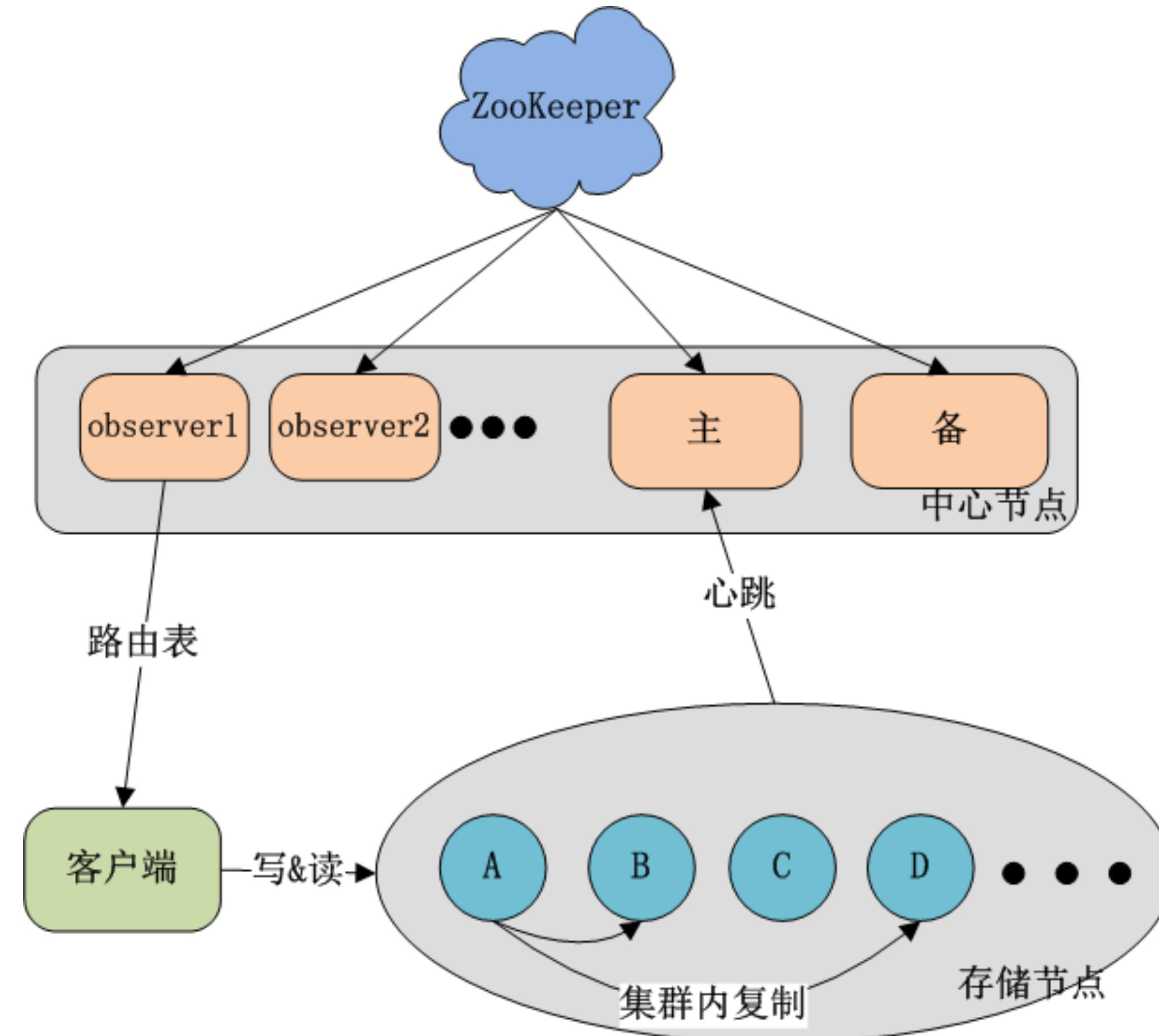


- Zookeeper选主
- 元数据Zookeeper存储

Cellar—中心节点架构演进

一致性：

- 主备强一致
- observer同步强一致



目录

- Cellar起源
- 中心节点架构演进
- 节点高可用和异地容灾
- 服务可用性提升
- Cellar规划

Cellar—节点高可用

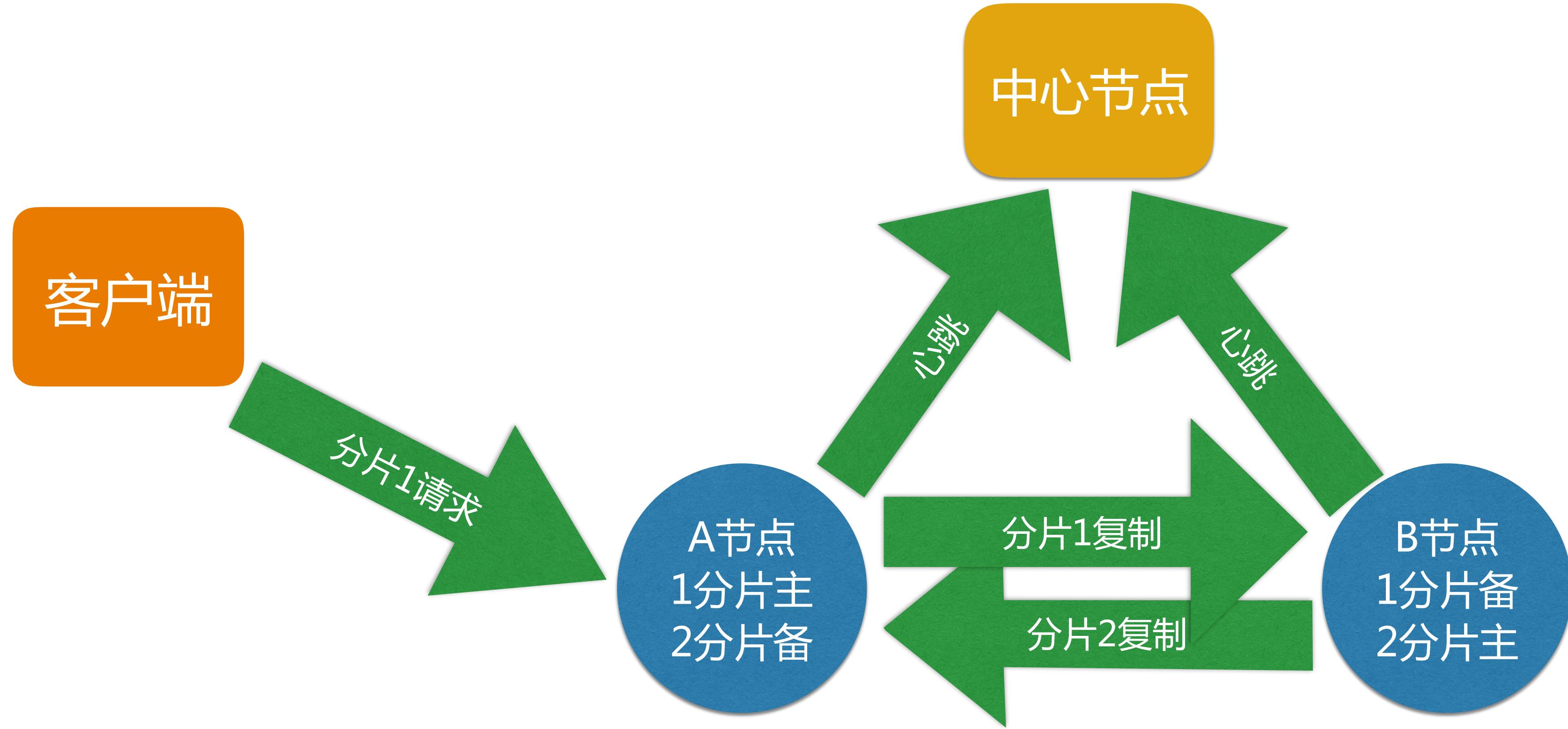
存储节点Failover，越快越好？

- 数据补全对业务影响
- 机器宕机五分钟，数据补全两小时

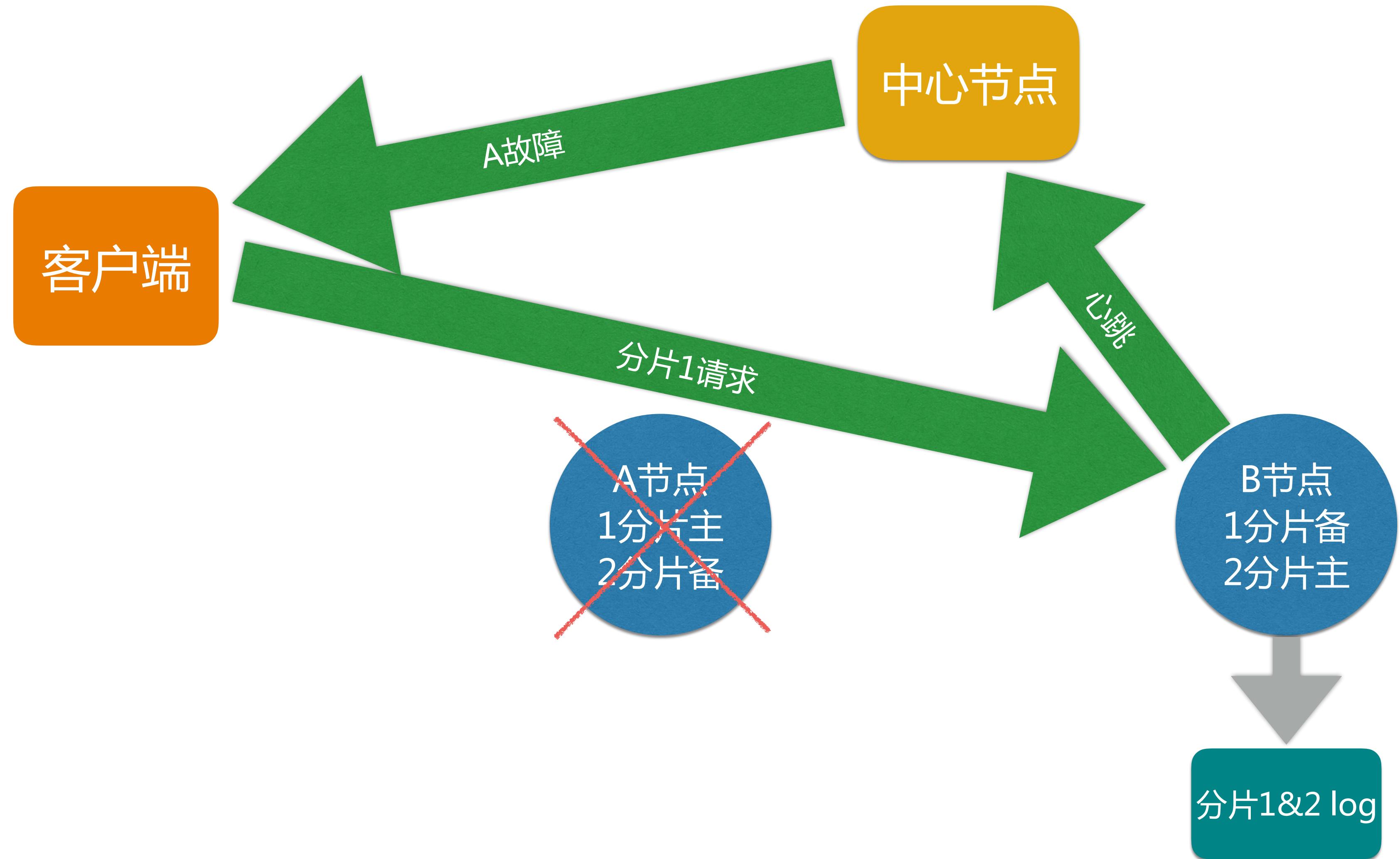
节点升级，先切走流量再操作？

- 节点流量只能切到有其他副本的节点
- 升级后的节点缺少升级期间的写入

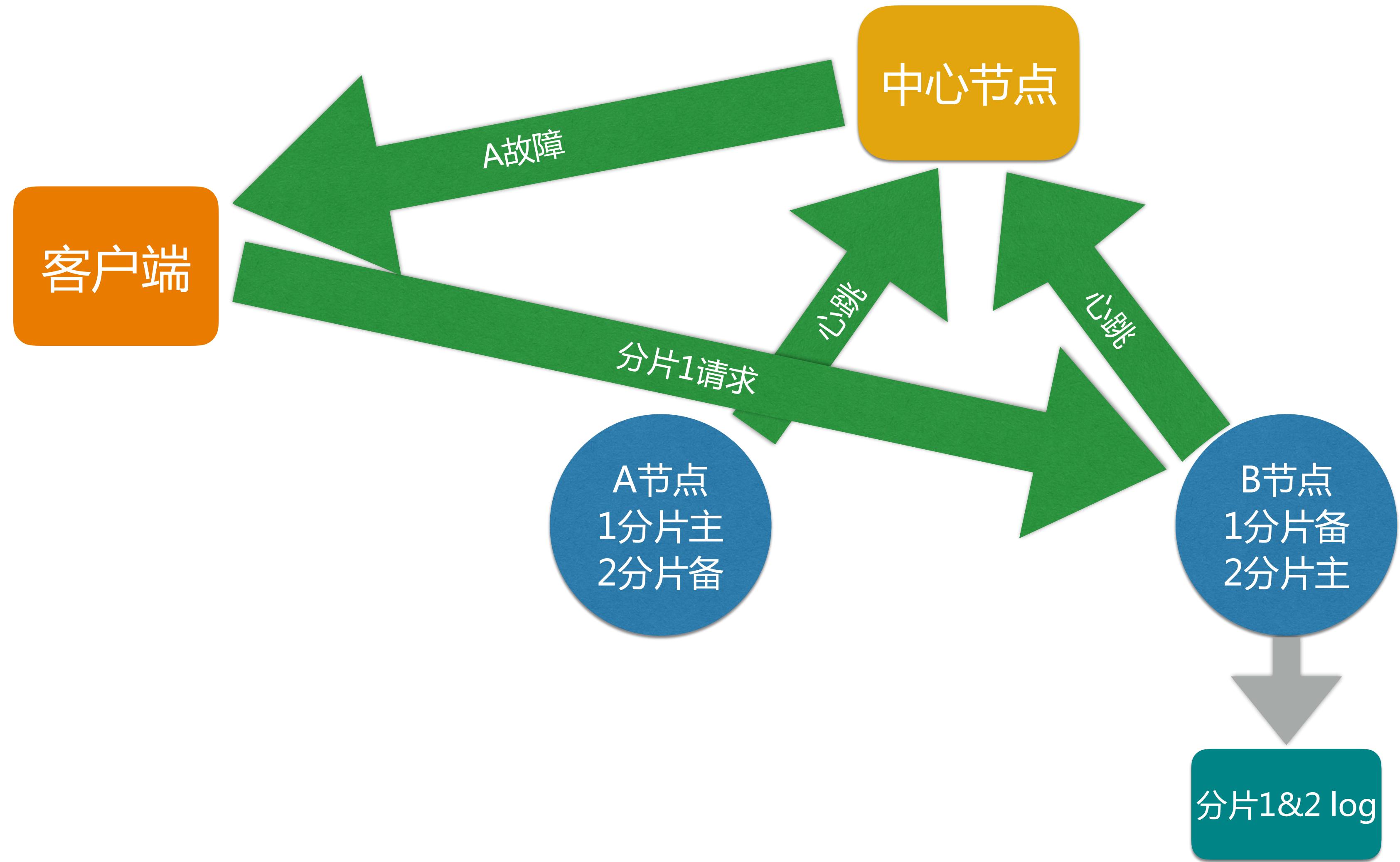
Cellar—节点高可用



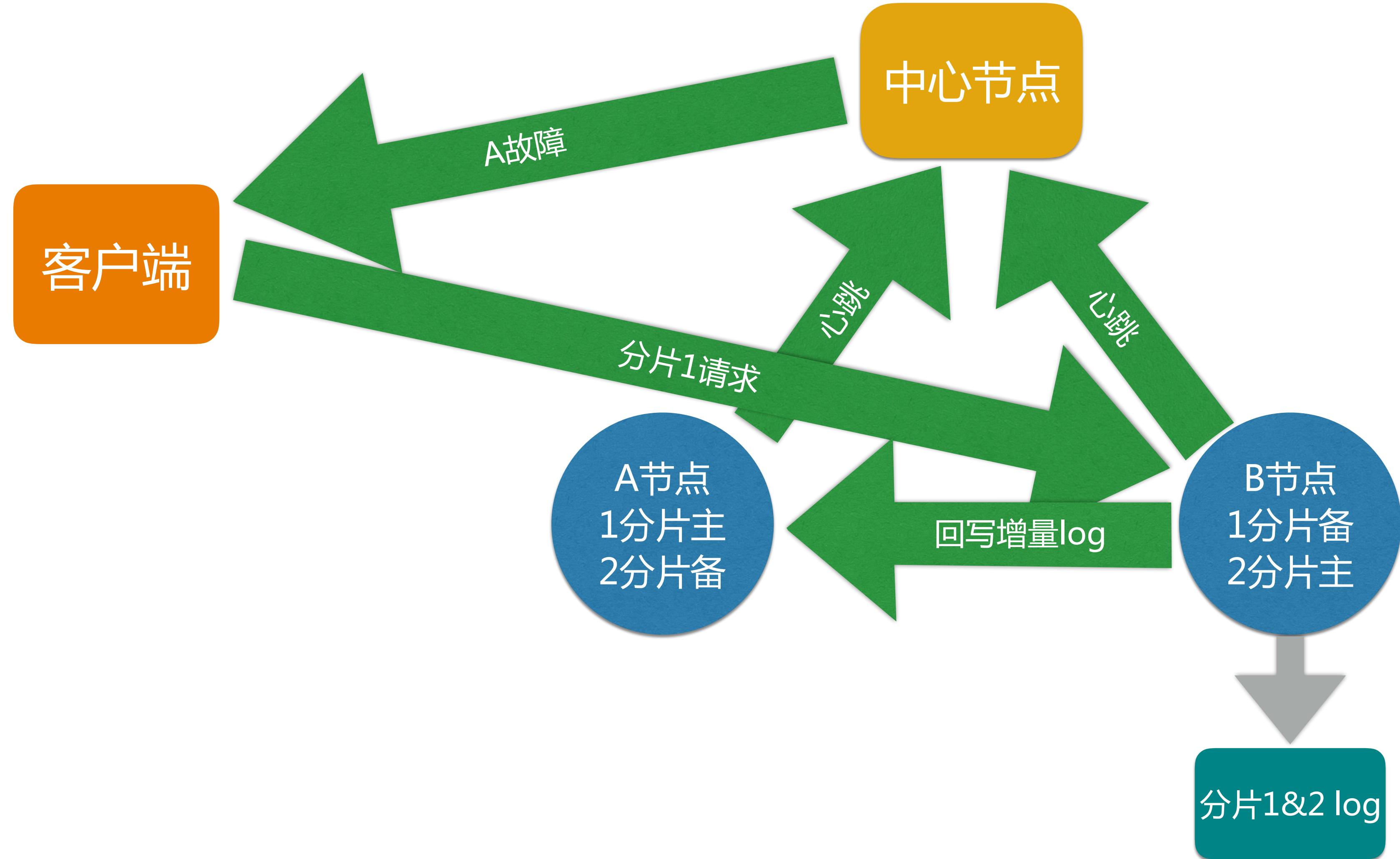
Cellar—节点高可用



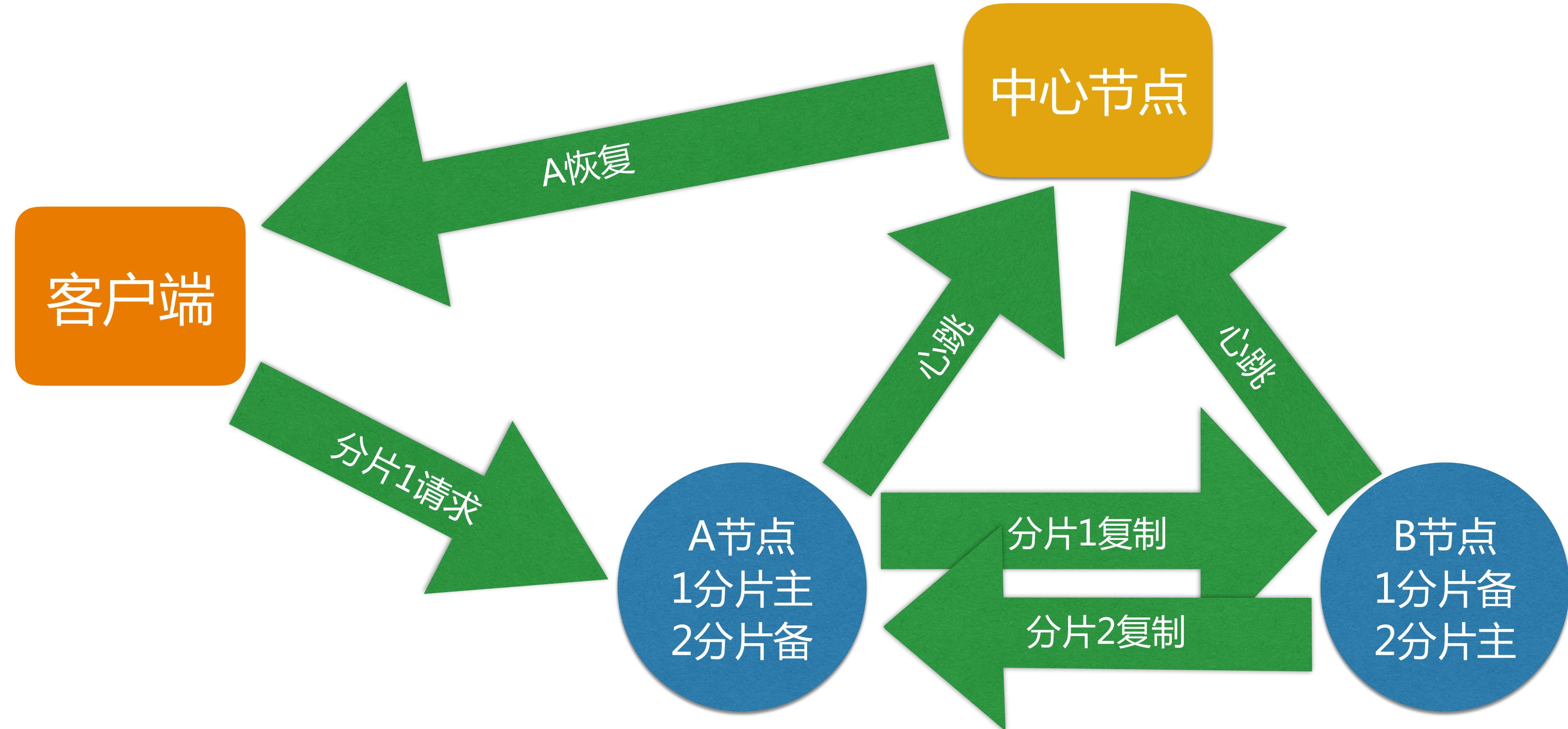
Cellar—节点高可用



Cellar—节点高可用

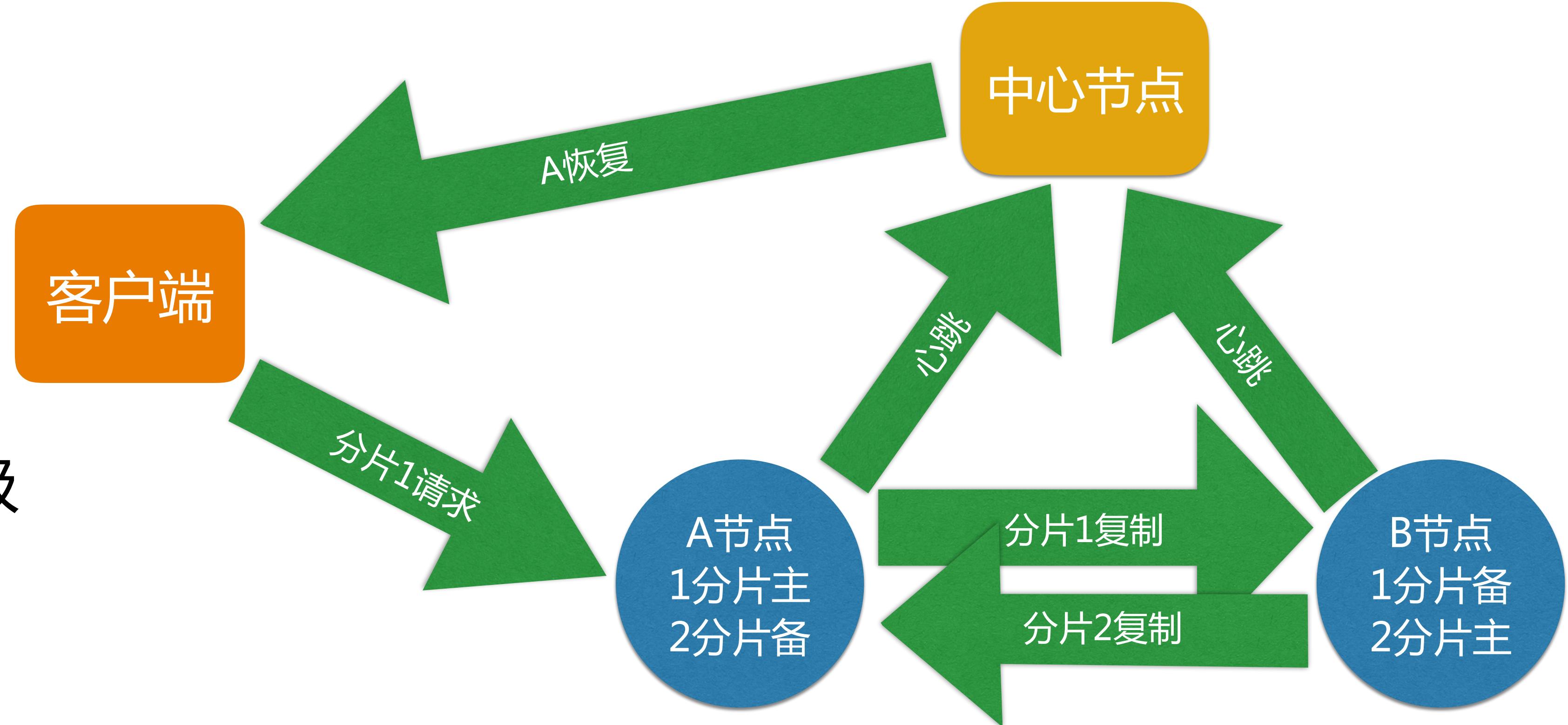


Cellar—节点高可用



Cellar—节点高可用

- 秒级容灾
无数据迁移
- 节点静默升级

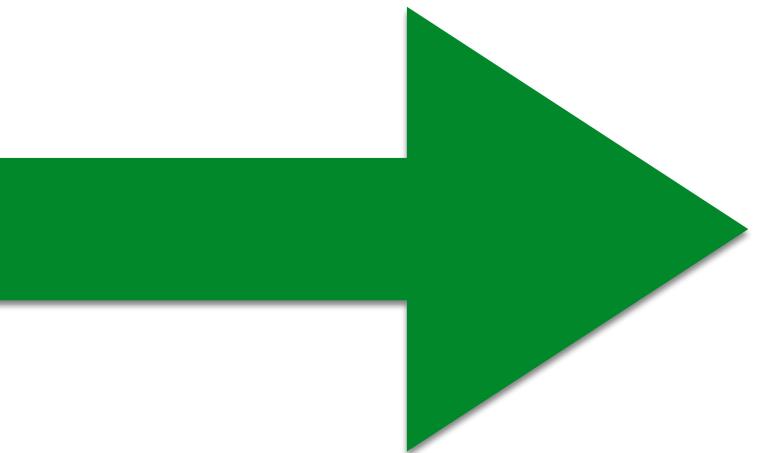


Cellar—异地容灾

- 多机房建设
 网络延迟大
 专线稳定性差
- 异地容灾需求

Cellar—异地容灾

- 多机房建设
网络延迟大
专线稳定性差
- 异地容灾需求



跨集群数据同步

Cellar—异地容灾

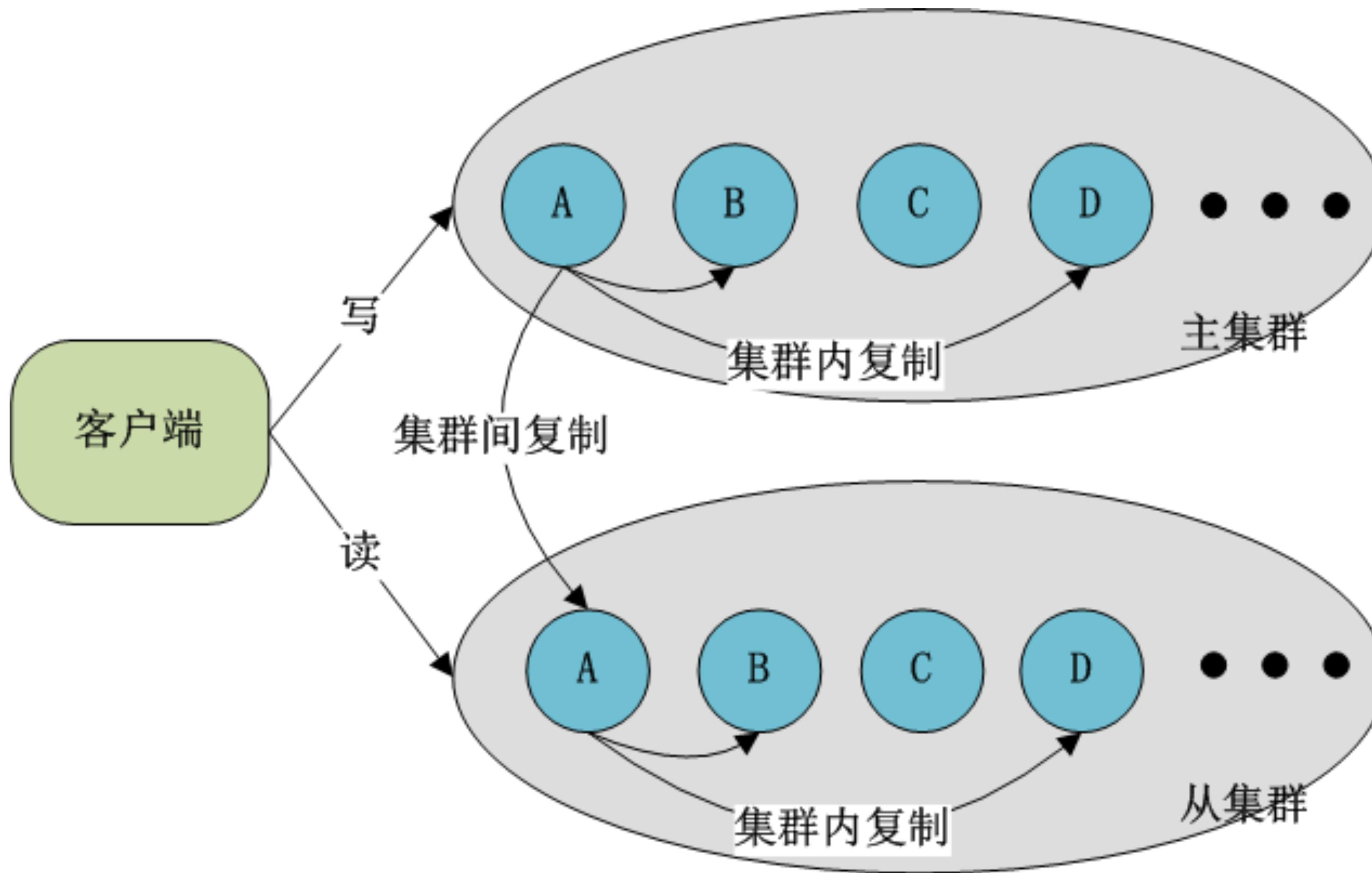
	集群节点同步	消息队列同步
复制延迟	低	高
系统复杂度	低	高
运维成本	低	高
实现难度	高	低
扩展性	低	高

Cellar—异地容灾

	集群节点同步	消息队列同步
复制延迟	低	高
系统复杂度	低	高
运维成本	低	高
实现难度	高	低
扩展性	低	高

- 低延迟
- 低复杂度（运维成本）

Cellar—异地容灾



目录

- Cellar起源
- 中心节点架构演进
- 节点高可用和异地容灾
- 服务可用性提升
- Cellar规划

Cellar—服务可用性提升

影响可用性的问题

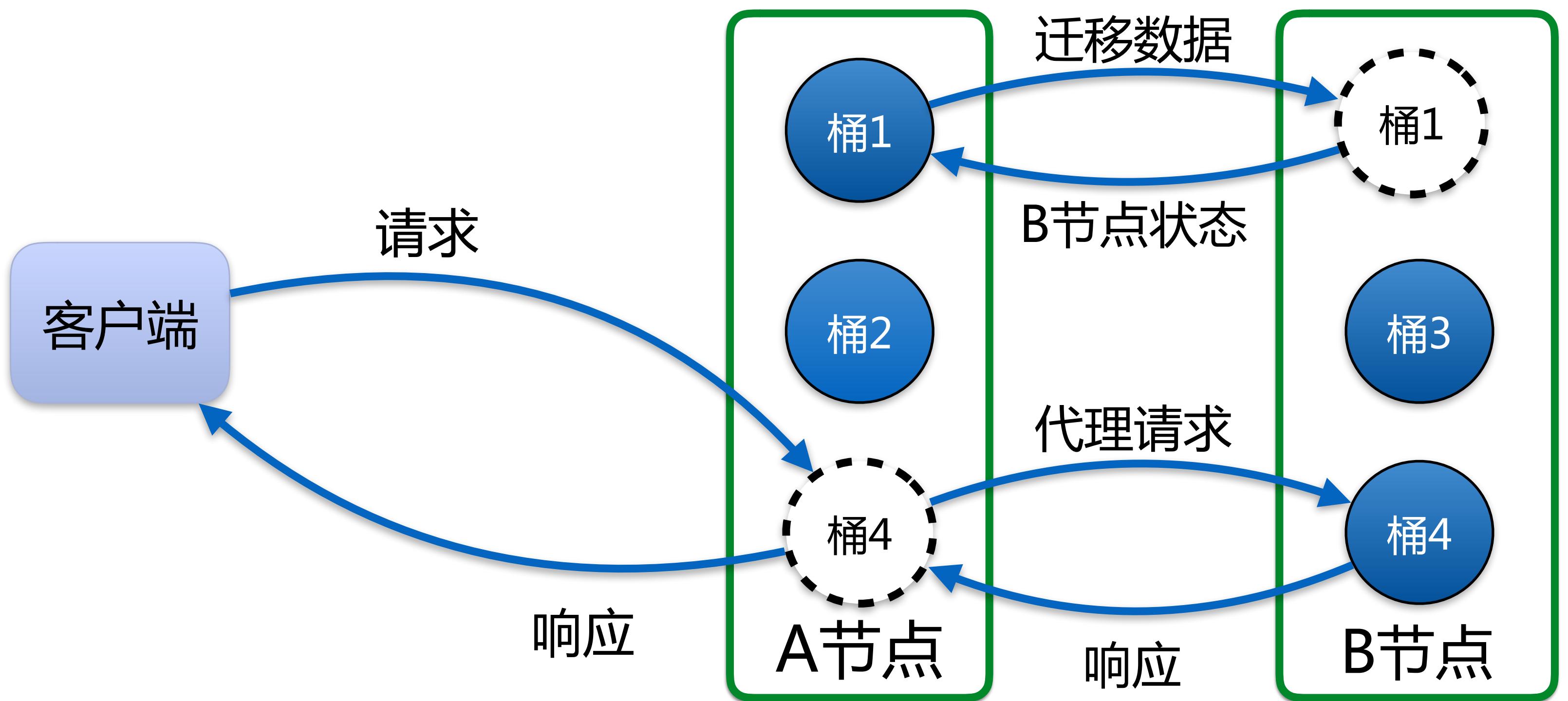
- 数据迁移
- 请求超时抖动

Cellar—无损数据迁移

数据迁移的问题

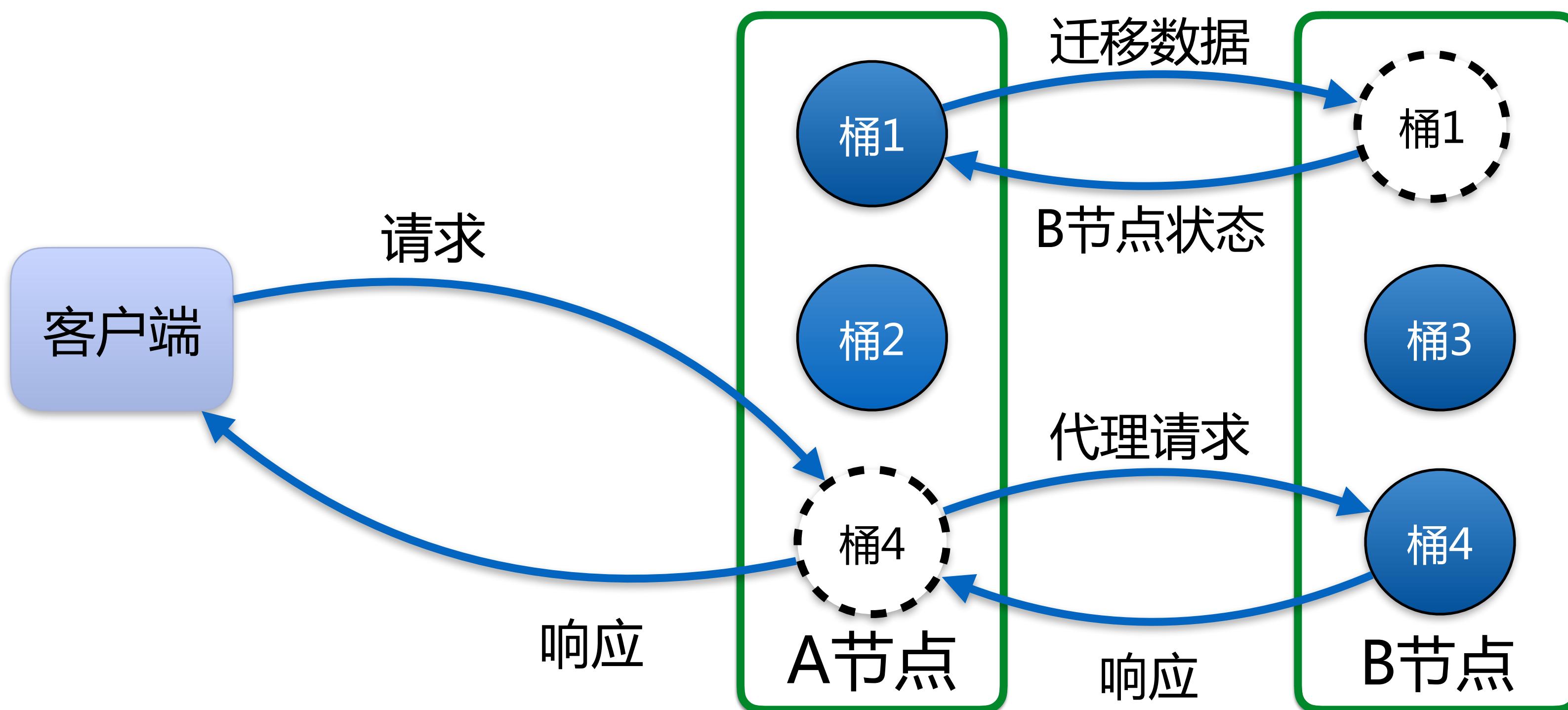
- 迁移速度不可控，易影响业务请求
- 路由表更新瞬间请求失败
- key级别迁移写入，引擎压力大

Cellar—无损数据迁移



智能调速 + 全程代理

Cellar—无损数据迁移



- 节点状态指标
 - 引擎压力
 - 网卡
 - 队列
 - QPS
 - ...

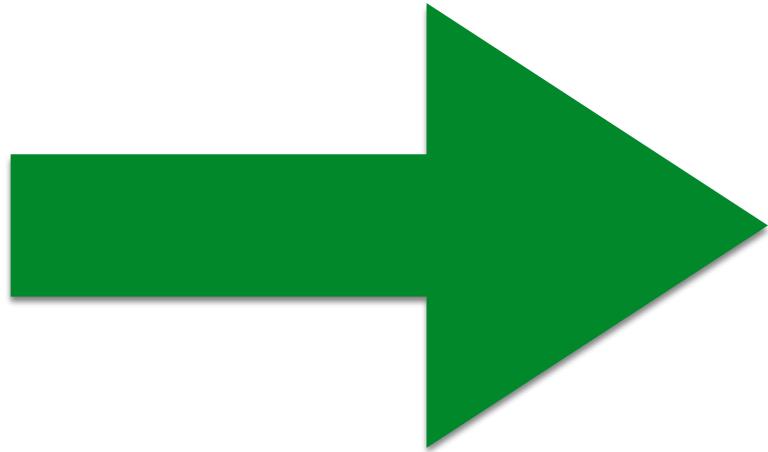
智能调速 + 全程代理

Cellar—请求超时原因

- 客户端问题
GC、CPU繁忙…
- 网络问题
- 服务器端问题
磁盘IO、慢请求…

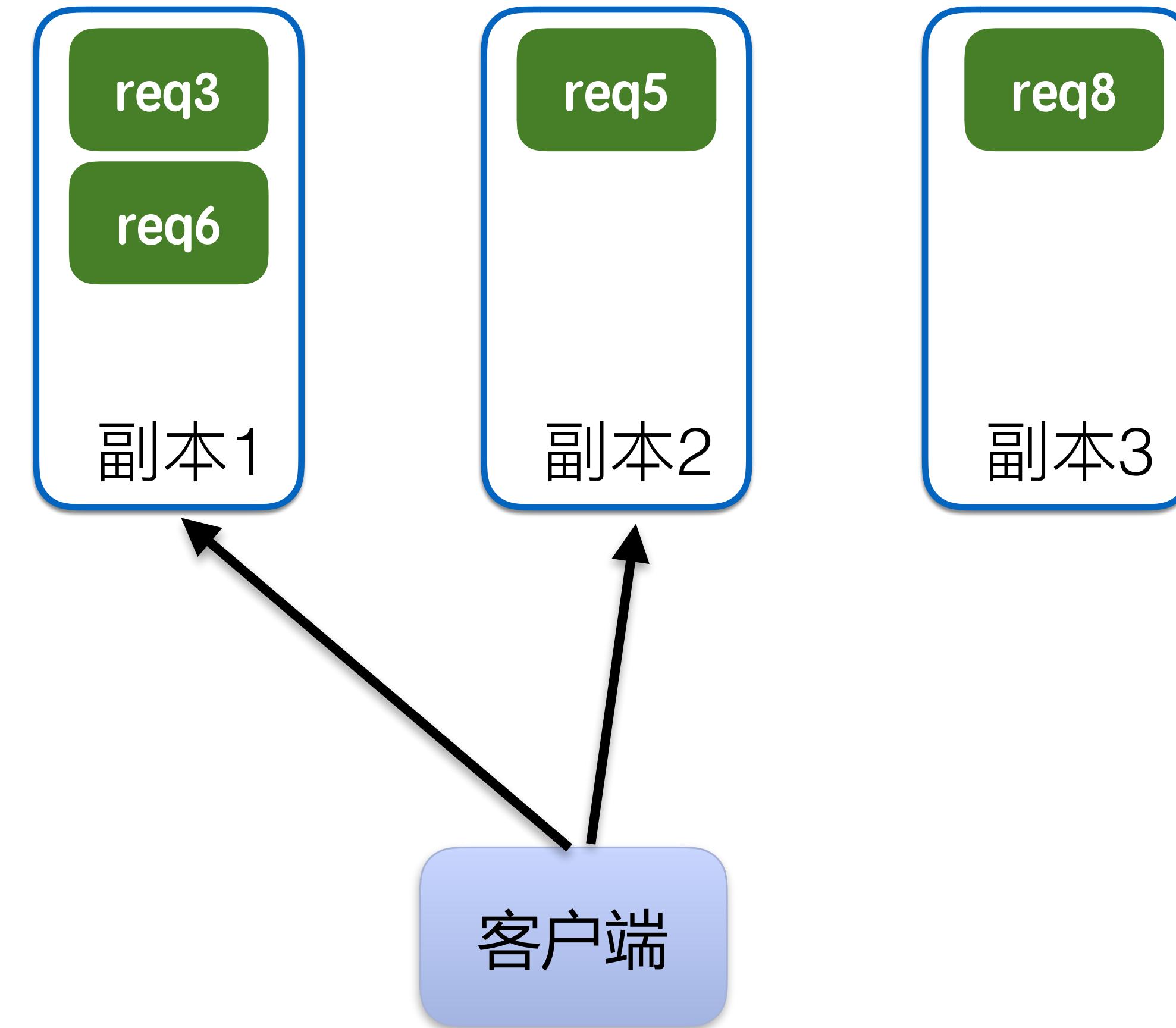
Cellar—请求超时原因

- 客户端问题
GC、CPU繁忙…
- 网络问题
- 服务器端问题
磁盘IO、慢请求…

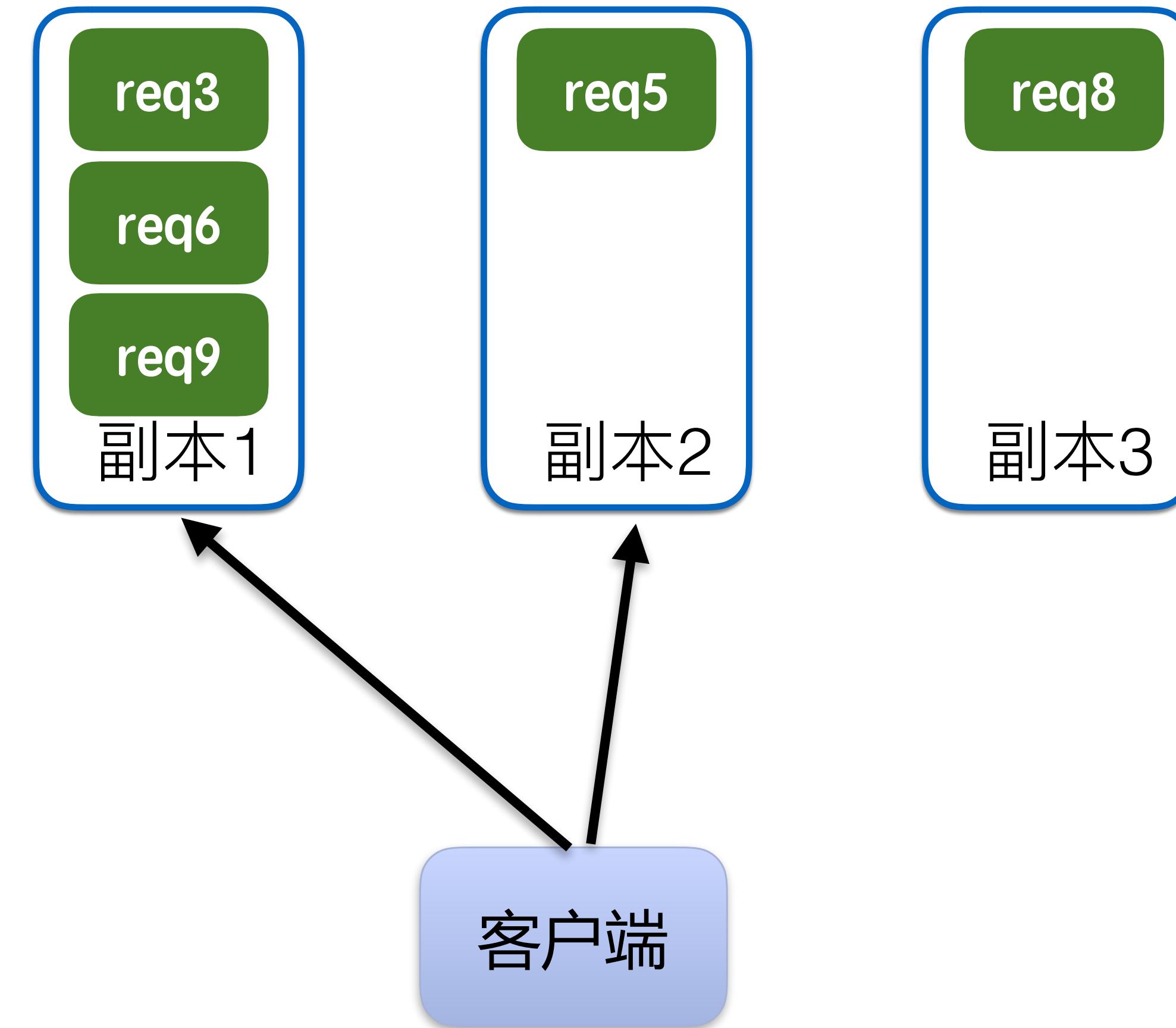


- 客户端
backup request
- 服务器端
快慢队列

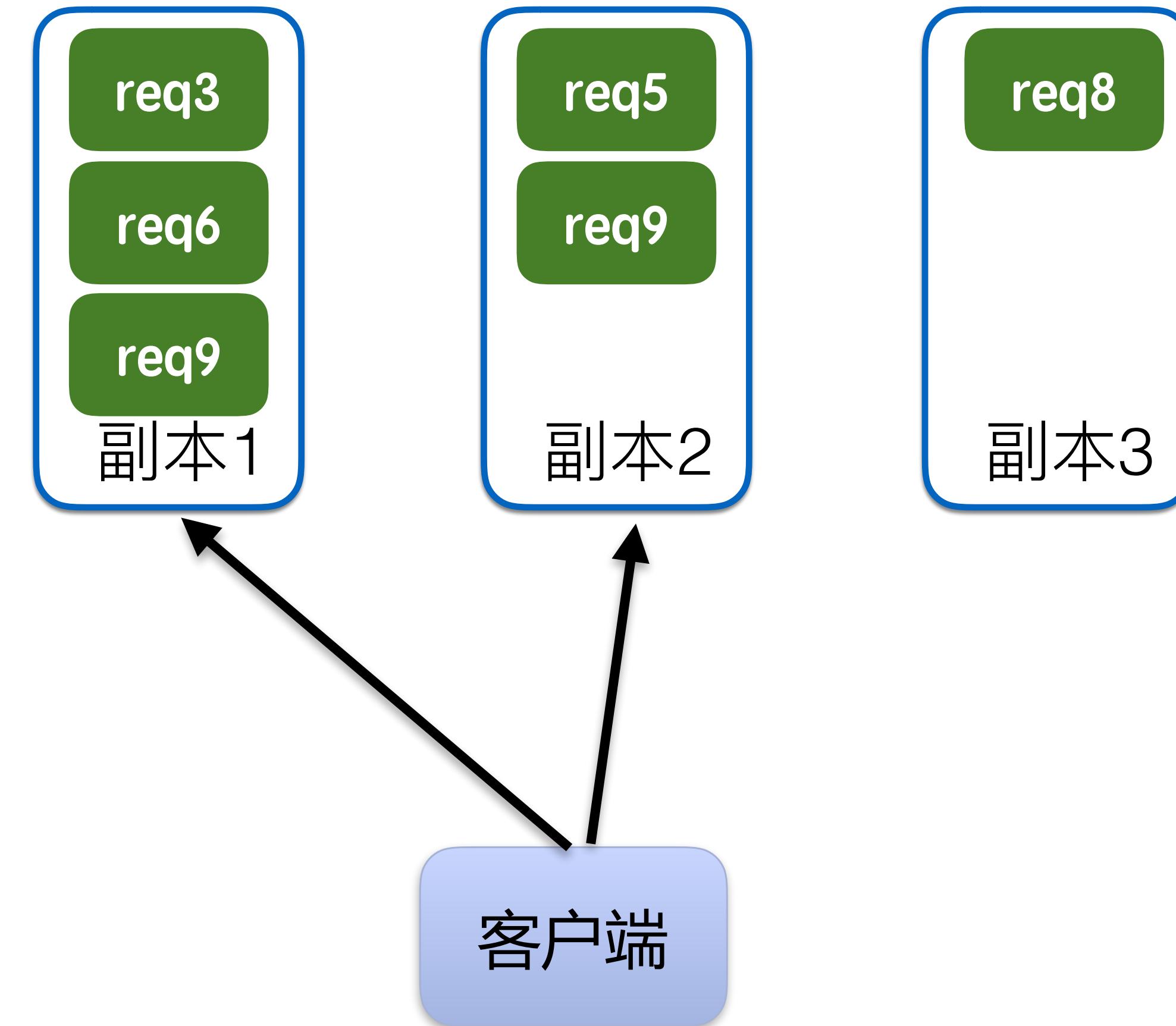
Cellar—backup request



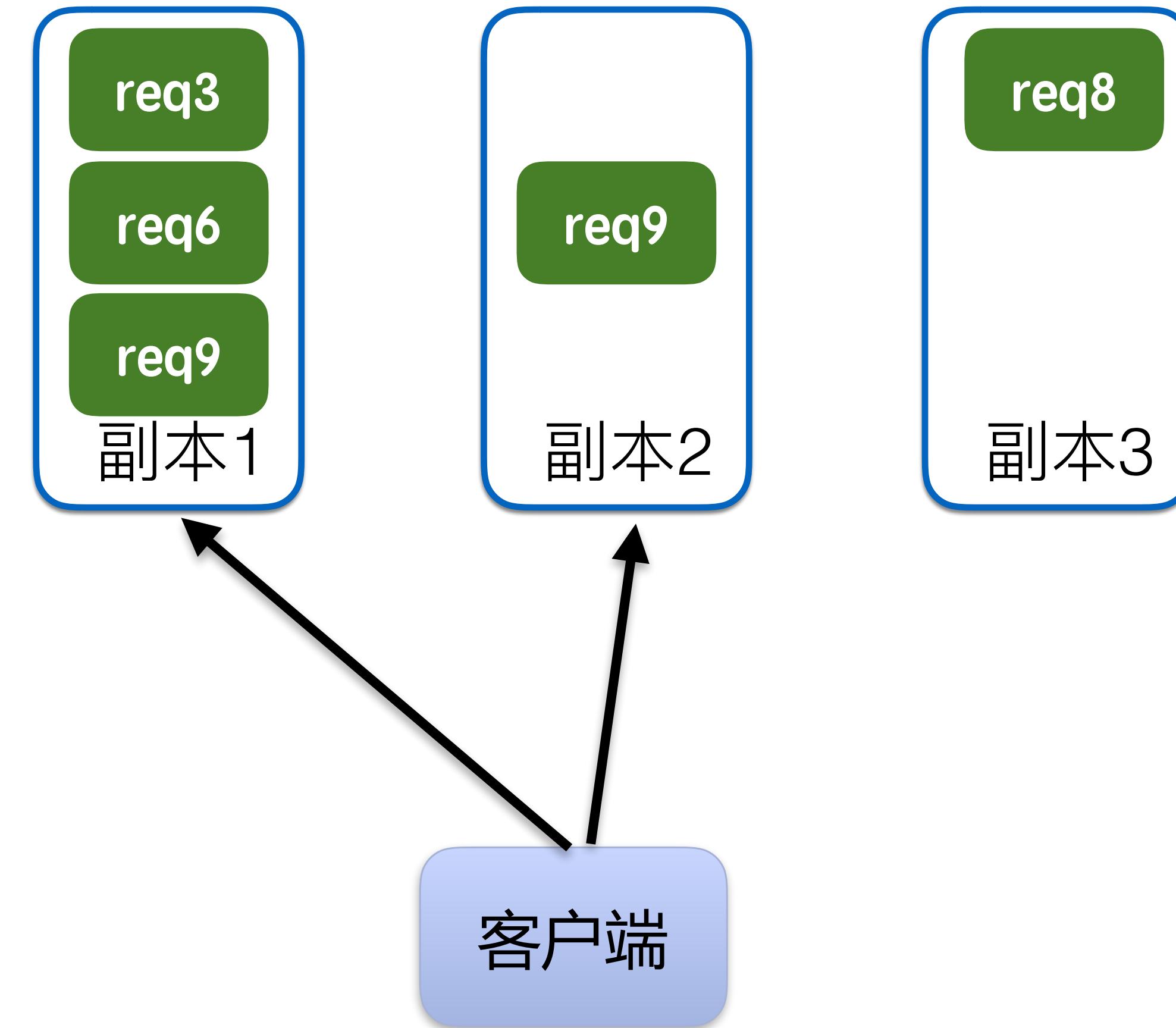
Cellar—backup request



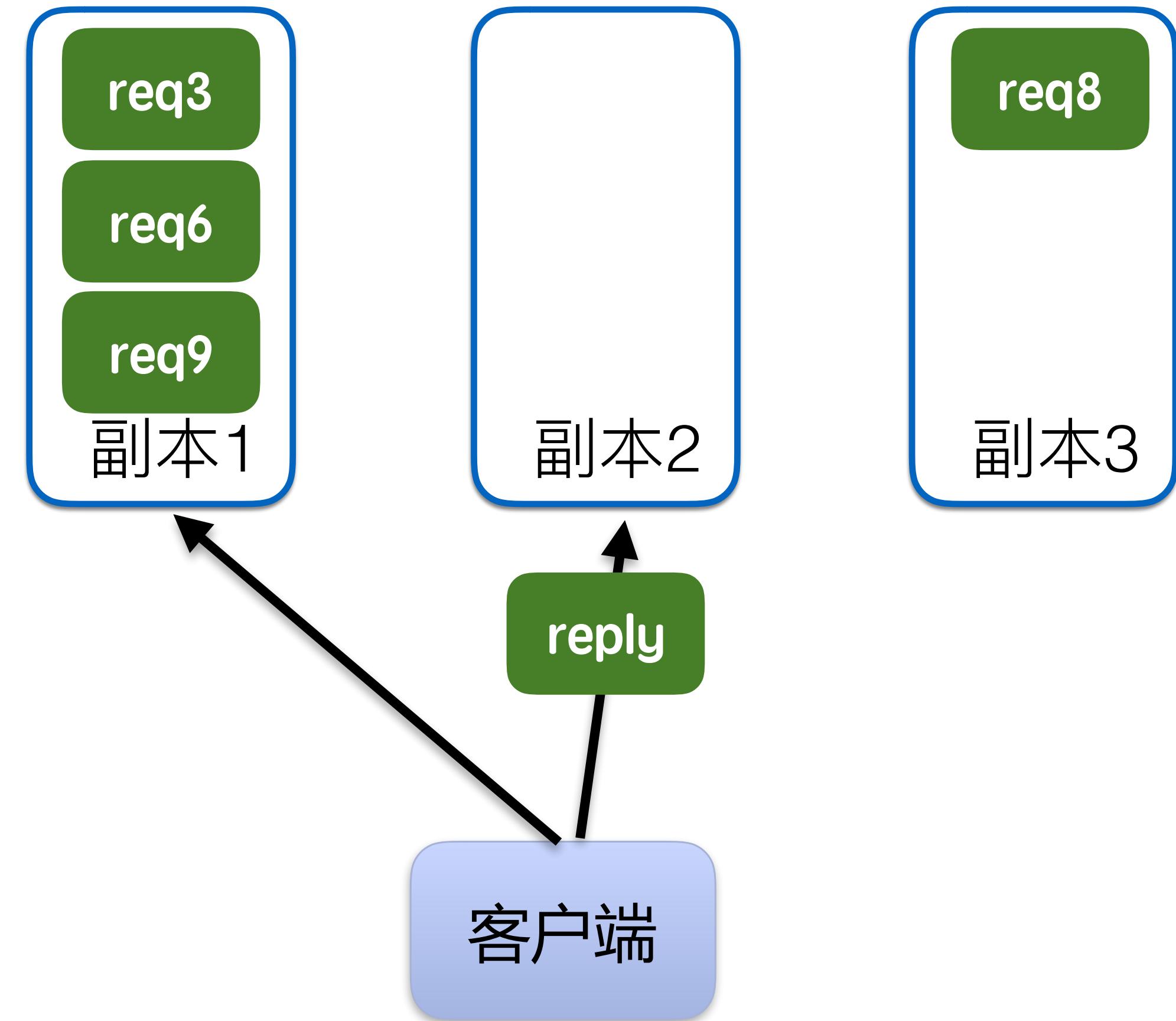
Cellar—backup request



Cellar—backup request

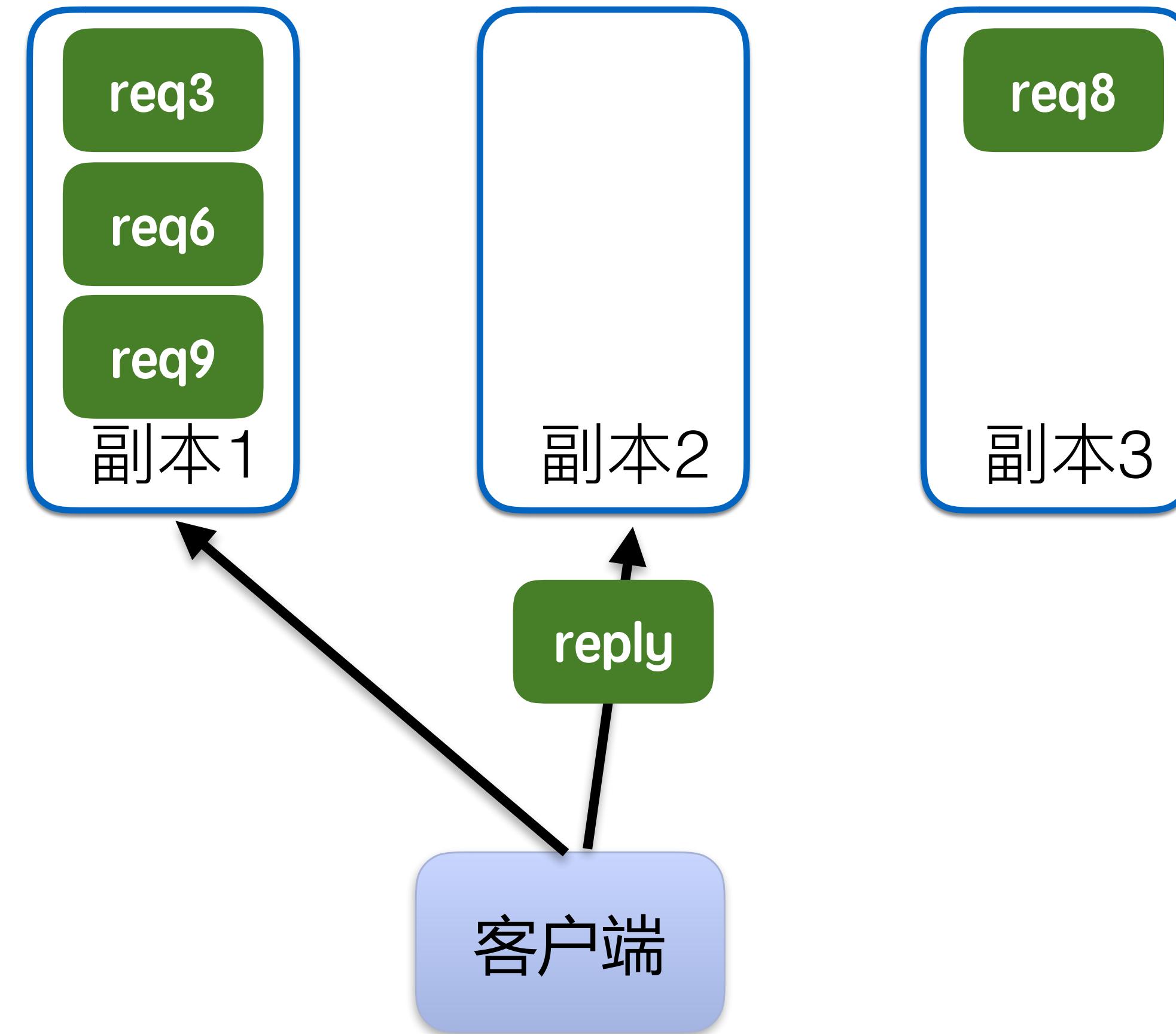


Cellar—backup request



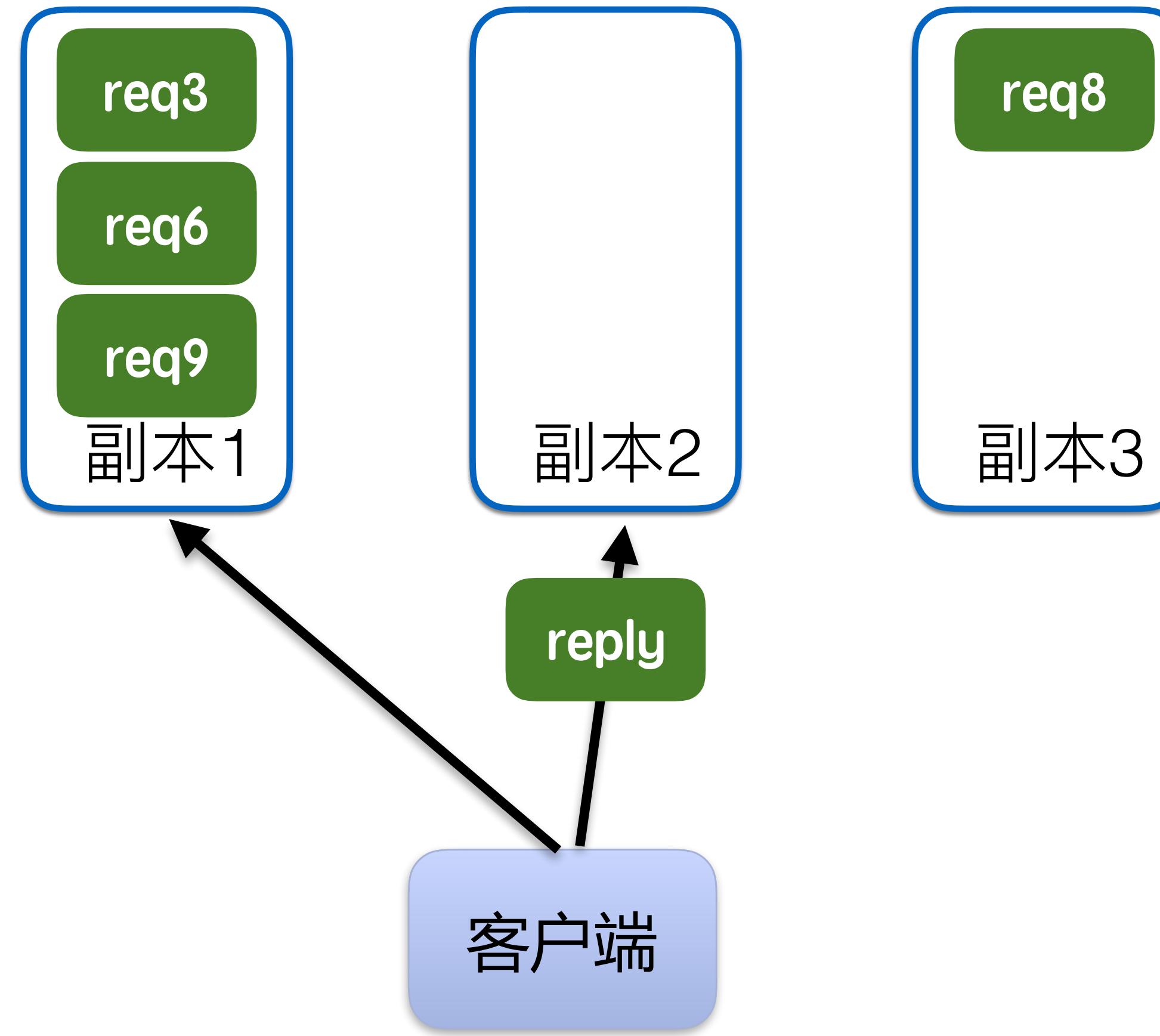
Cellar—backup request

- 什么时间
等待超过
超时时间一半
- 发几次
最多两次
- 重试比例
最大20%



Cellar—backup request

- 什么时间
等待超过
超时时间一半
- 发几次
最多两次
- 重试比例
最大20%



读请求超时降低90+%

Cellar—快慢队列

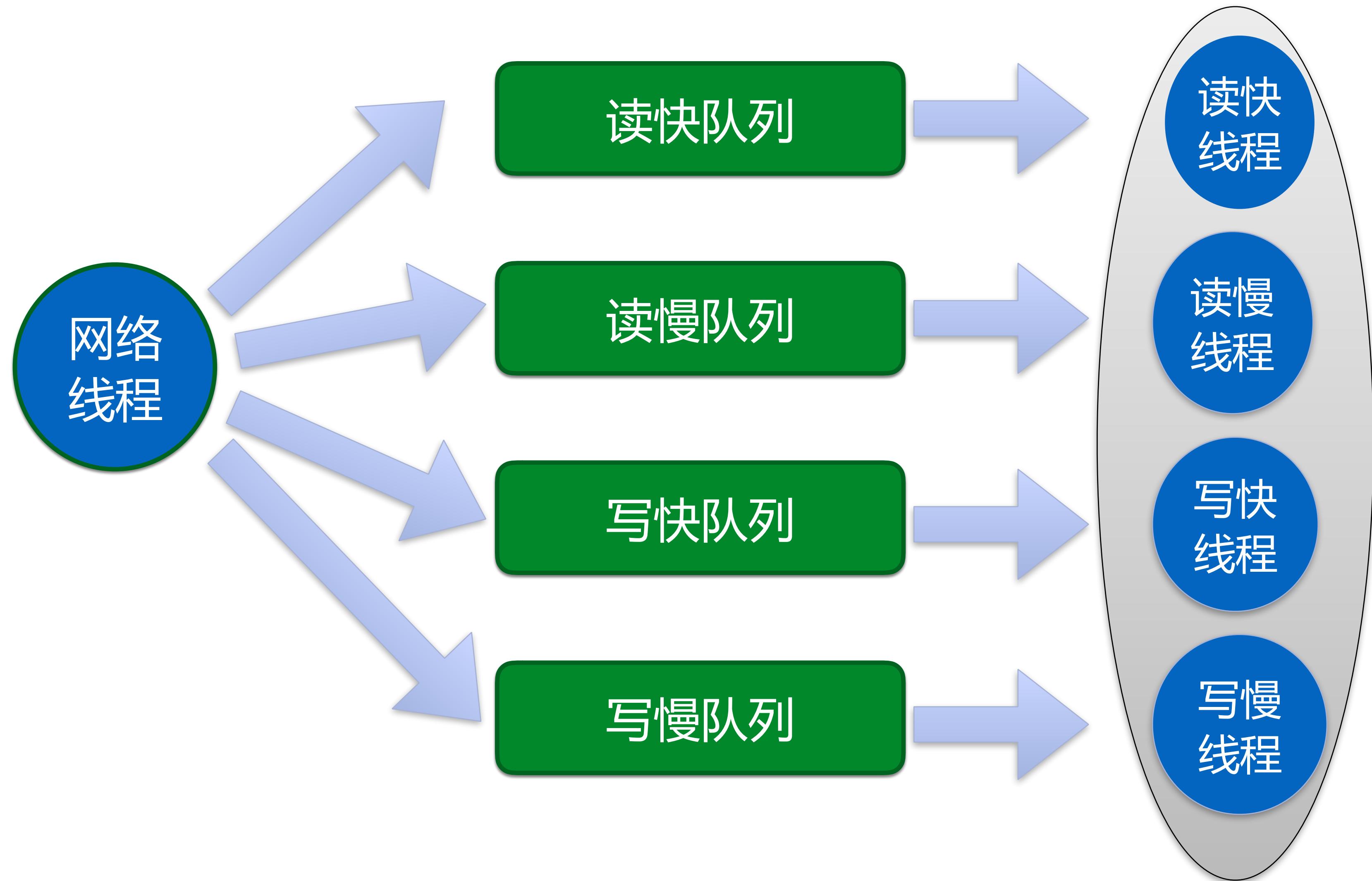


Cellar—快慢队列

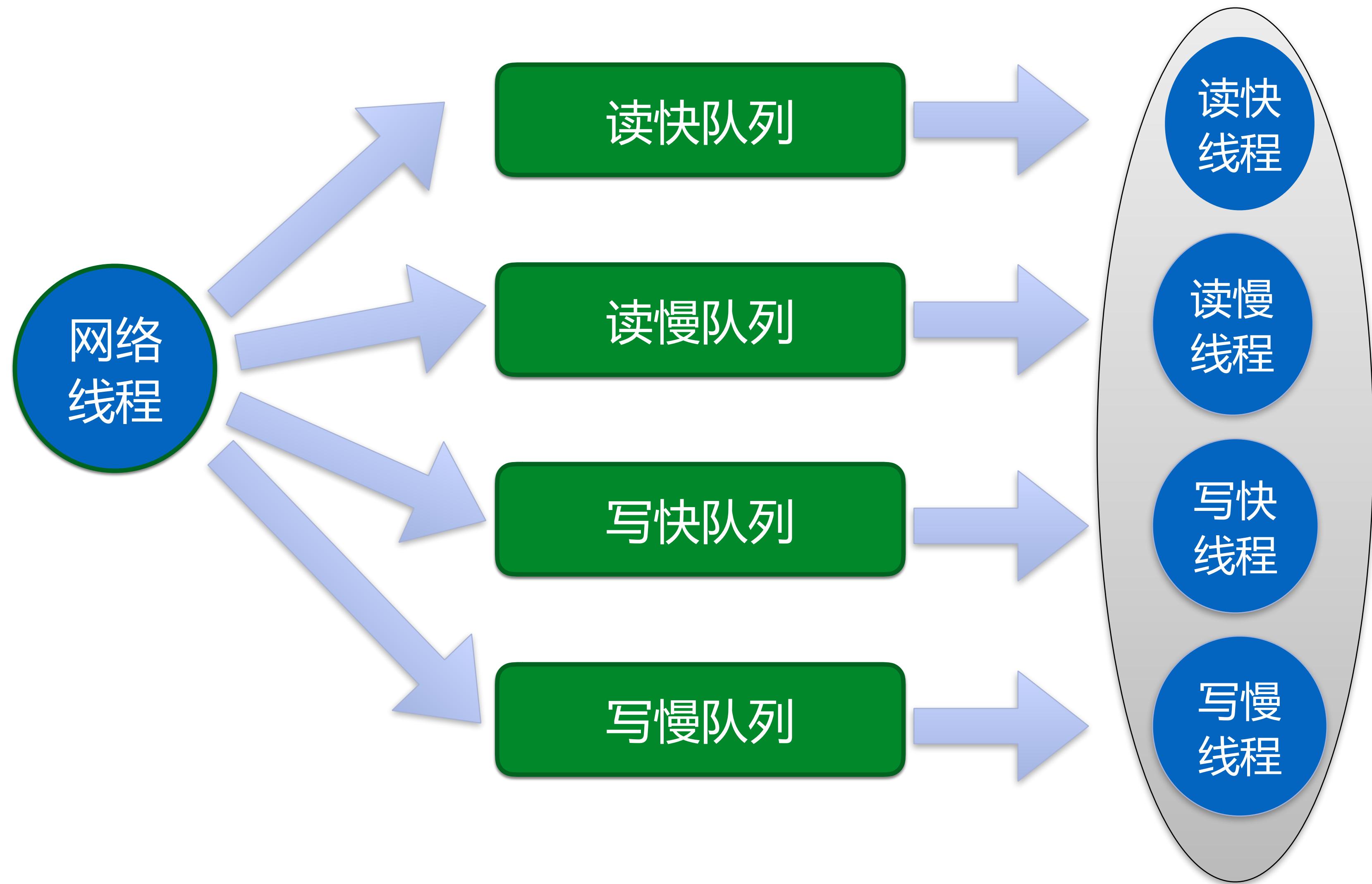


问题：
共用队列&线程
线上慢请求：超时请求 1 : 20

Cellar—快慢队列



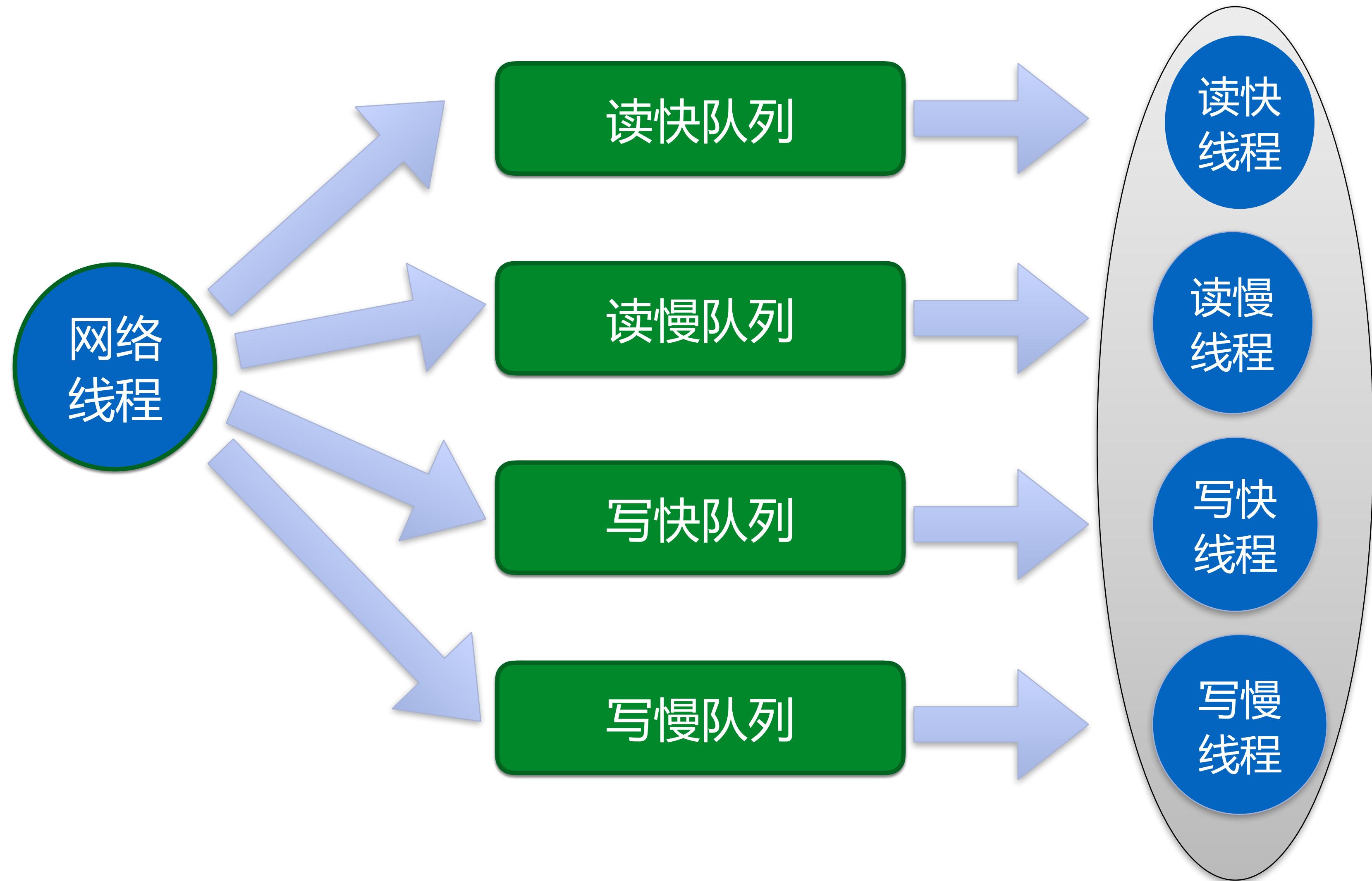
Cellar—快慢队列



慢请求判断：

- 耗时接口 (range...)
- value过大
- 单请求key过多
- ...

Cellar—快慢队列



慢请求判断：

- 耗时接口 (range...)
- value过大
- 单请求key过多
- ...

TP999延迟降低86%

目录

- Cellar起源
- 中心节点架构演进
- 节点高可用和异地容灾
- 服务可用性提升
- Cellar规划

Cellar规划

系统研发

- 异地多活
- 跨机房自动容灾
- 磁盘粒度容灾
- 数据迁移优化

可运维性

- 容器化
- 自动扩缩容

谢谢

