美团点评万亿级 KV 存储架构与实践

齐泽斌

美团点评高级技术专家

自我介绍

美团点评高级技术专家，KV 存储团队负责人，有 8 年以上分布式存储研发经验。
2011 年天津大学毕业后加入百度，负责过分布式文件系统 MFS 和分布式 KV BDRP 系统研发及运营。
2014 年加入美团，负责过分布式 KV 存储 Cellar、分布式缓存 Squirrel、数据传输 Databus 等系统研发及运营，主要关注于分布式存储技术领域。

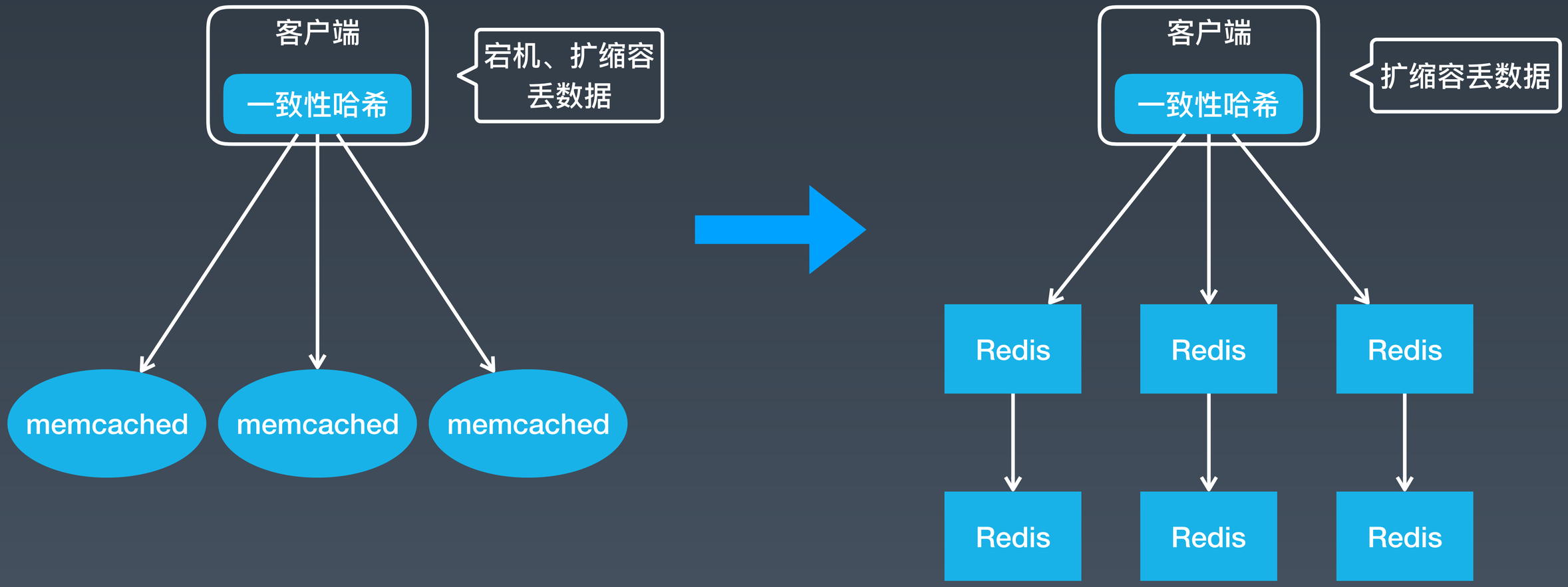
目录

- 美团点评 KV 存储发展历程
- 内存 KV Squirrel 架构和实践
- 持久化 KV Cellar 架构和实践
- 发展规划和业界趋势

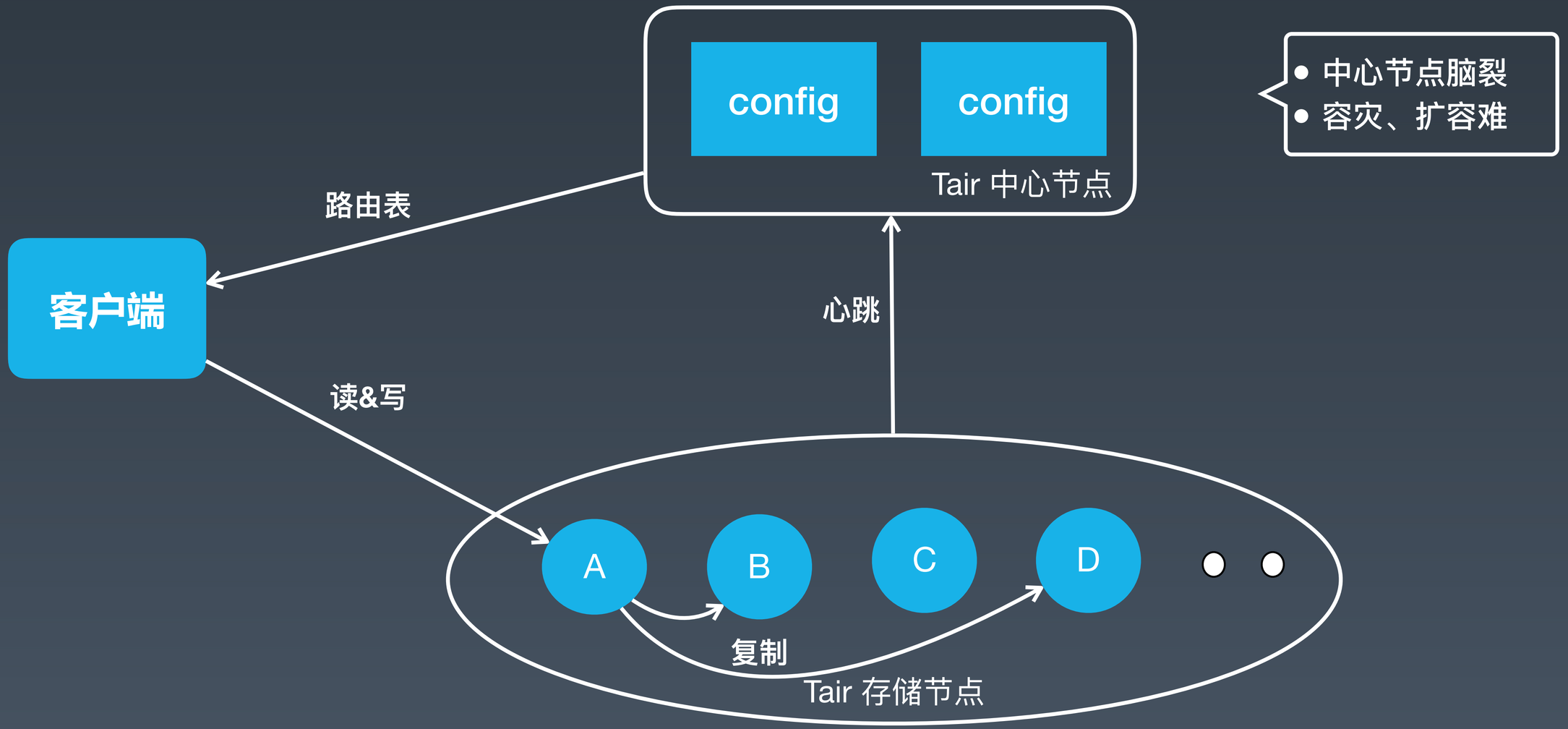
目录

- **美团点评 KV 存储发展历程**
- 内存 KV Squirrel 架构和实践
- 持久化 KV Cellar 架构和实践
- 发展规划和业界趋势

美团点评 KV 存储发展历程



美团点评 KV 存储发展历程



美团点评 KV 存储发展历程

Redis Cluster

自研 + 社区

Squirrel

全内存、高吞吐、低延迟

Tair

自研

Cellar

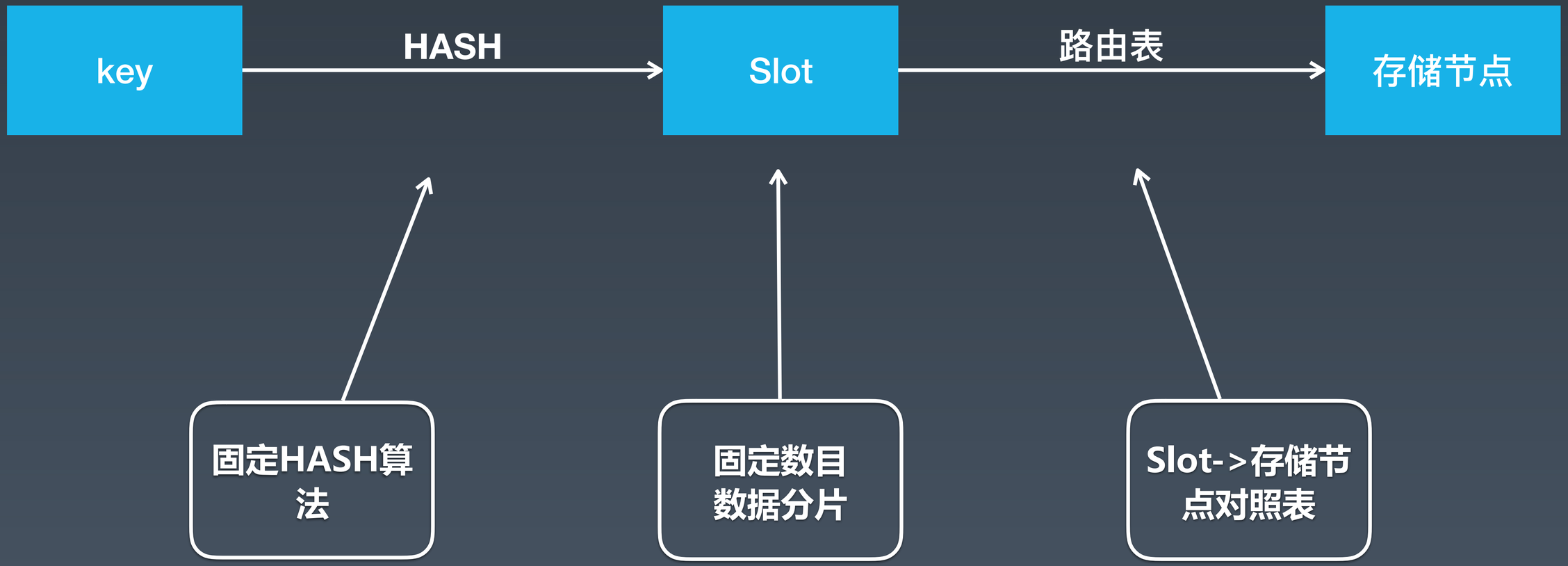
持久化、大容量、数据高可靠

- 日调用量均已破万亿
- 请求峰值均已破亿

目录

- 美团点评 KV 存储发展历程
- 内存 KV Squirrel 架构和实践
- 持久化 KV Cellar 架构和实践
- 发展规划和业界趋势

KV 数据分布介绍

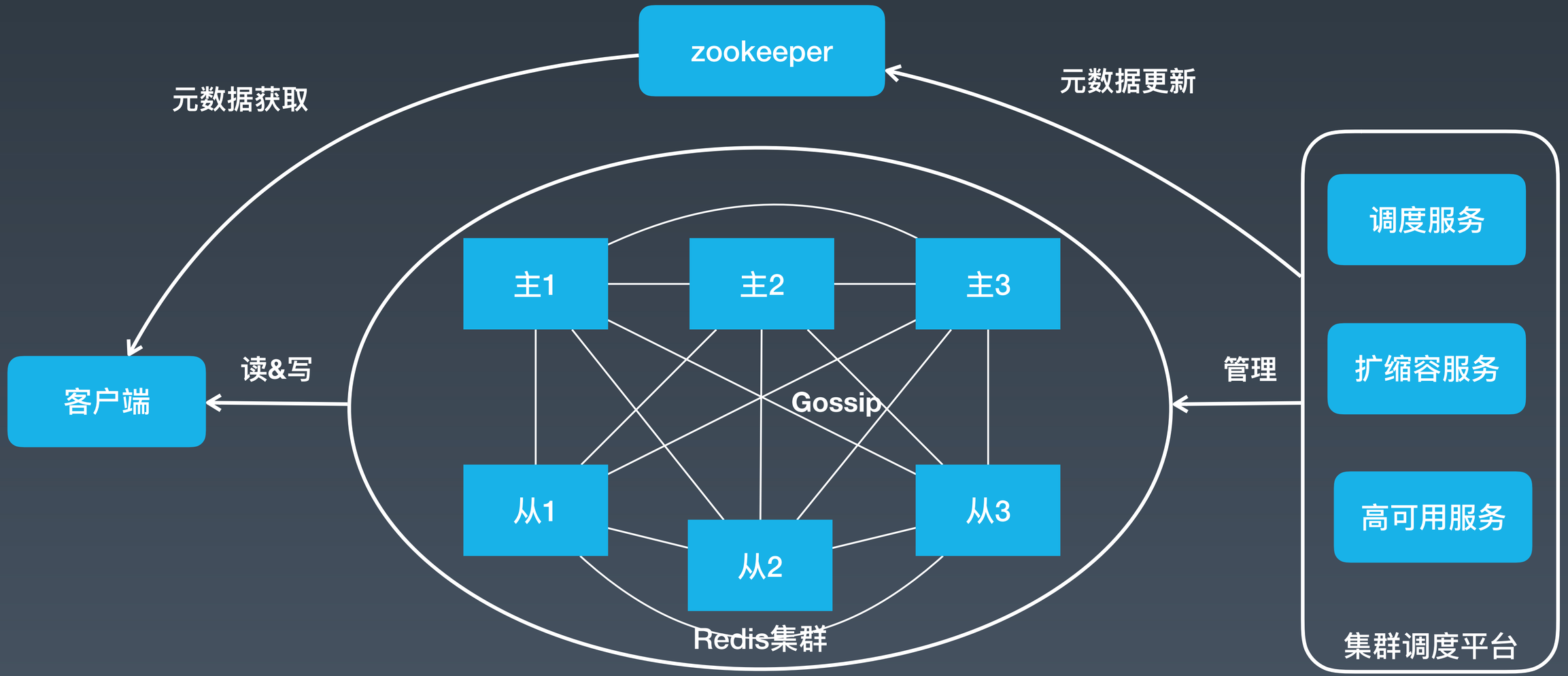


KV 架构和实践

高可用

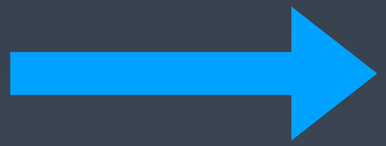
- **宏观：容灾**
- **微观：端到端成功率**

Squirrel 架构和实践



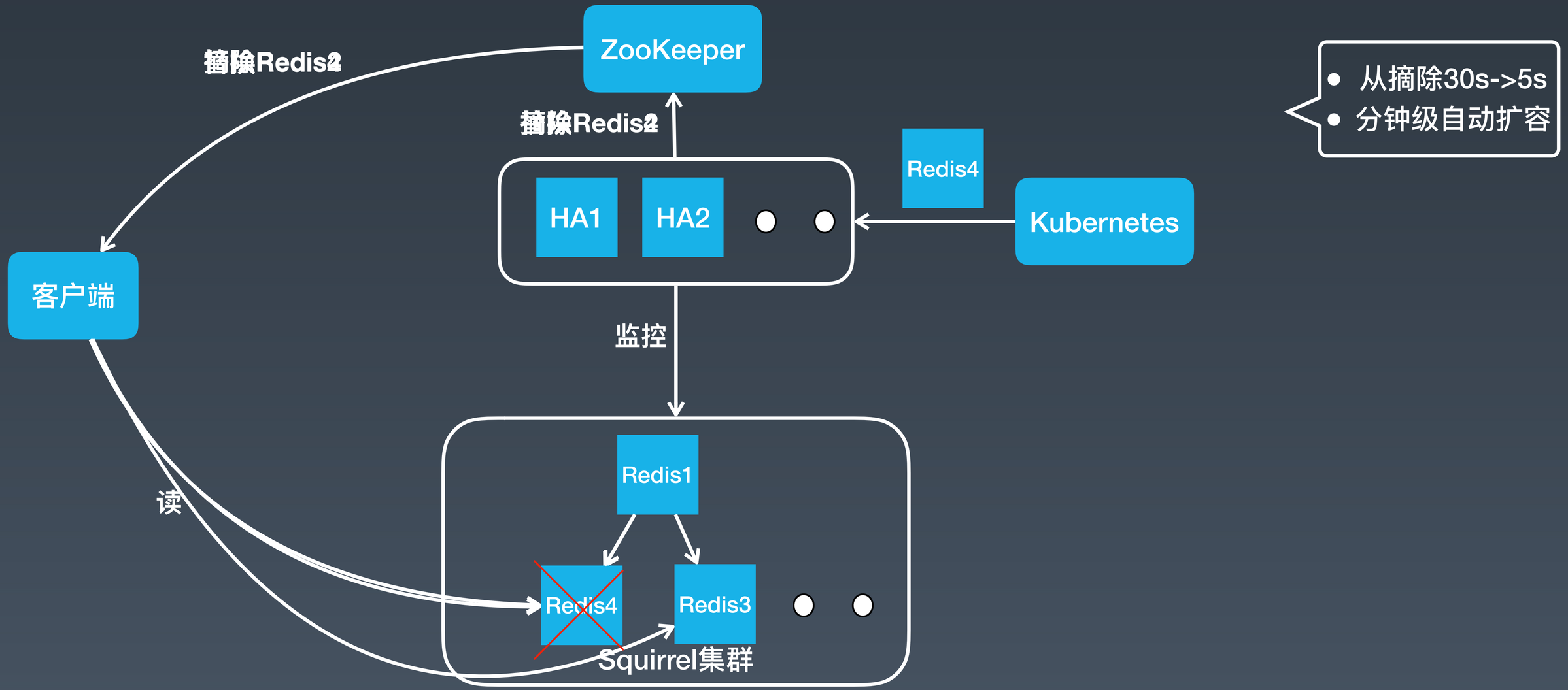
Squirrel—节点容灾

- 主库宕机恢复30s，从库有必要等这么久吗？
- 集群多，宕机后补副本累坏人？



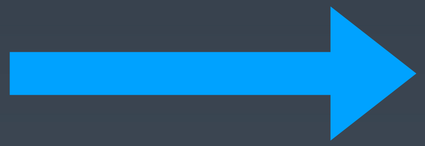
HA高可用服

Squirrel—节点容灾



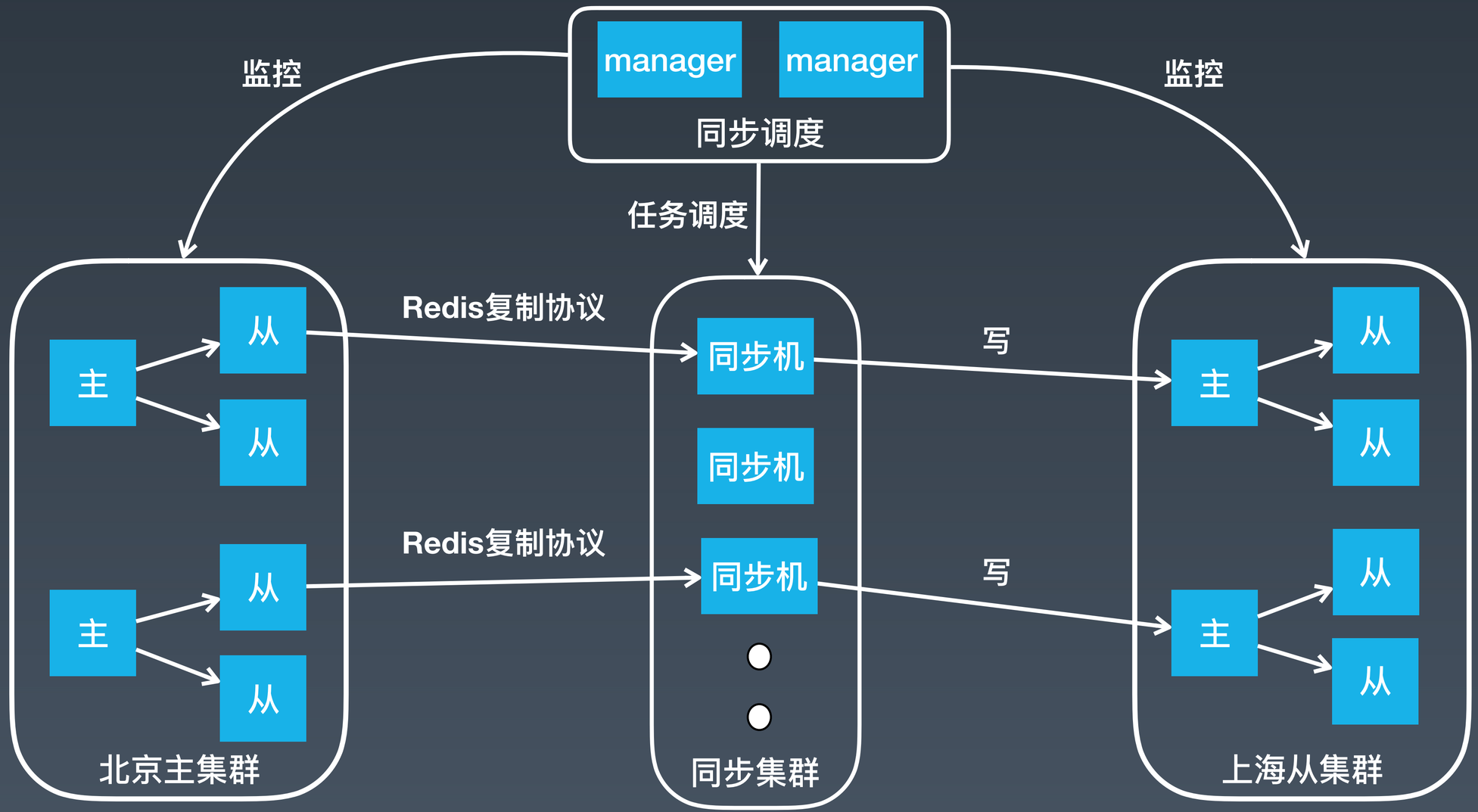
Squirrel—跨地域容灾

- 跨地域专线不稳定
- 跨地域专线有限的带宽
- 单元化部署，多活架构



集群间复制

Squirrel—跨地域容灾



Squirrel—端到端成功率

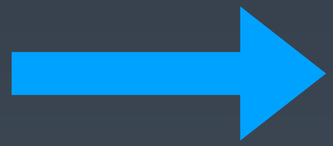
影响端到端成功率的因素：

- **数据迁移造成超时抖动**
- **持久化造成超时抖动**
- **热点key请求导致单节点过载**
- **□□□**

Squirrel—智能迁移

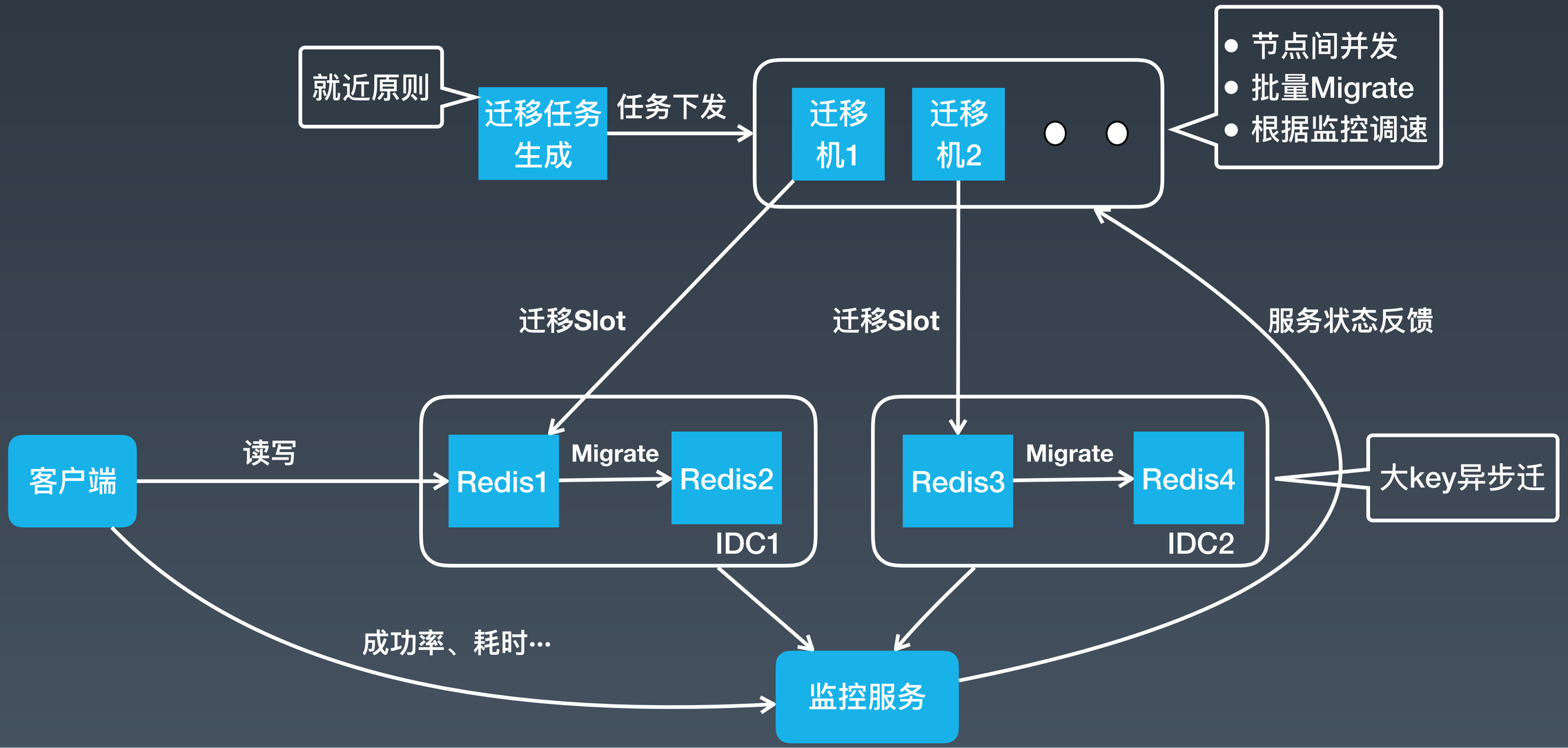
Redis提供数据迁移功能，但：

- Slot迁哪些、往哪迁、谁来迁？
- 想迁的快又怕太快影响业务？
- 迁移大key阻塞业务请求？



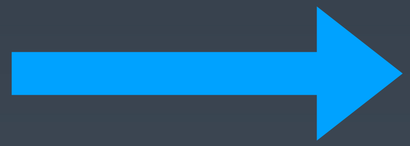
迁移服务

Squirrel—智能迁移



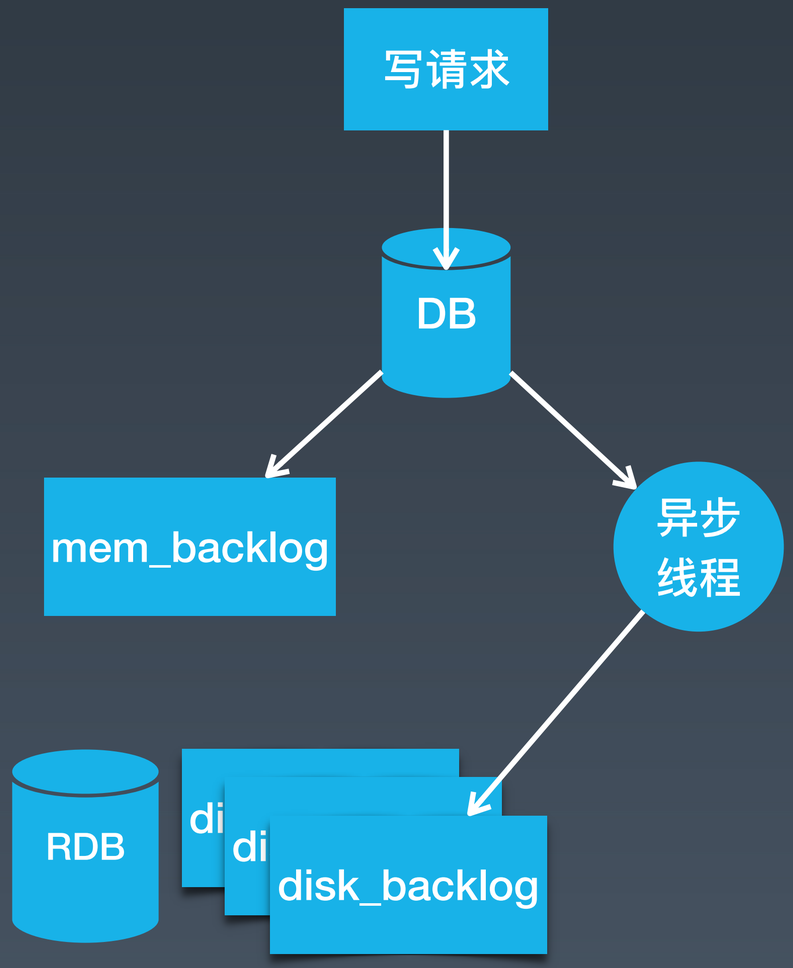
Squirrel—持久化重构

- 做不起的RDB
- 无法避免抖动的AOF



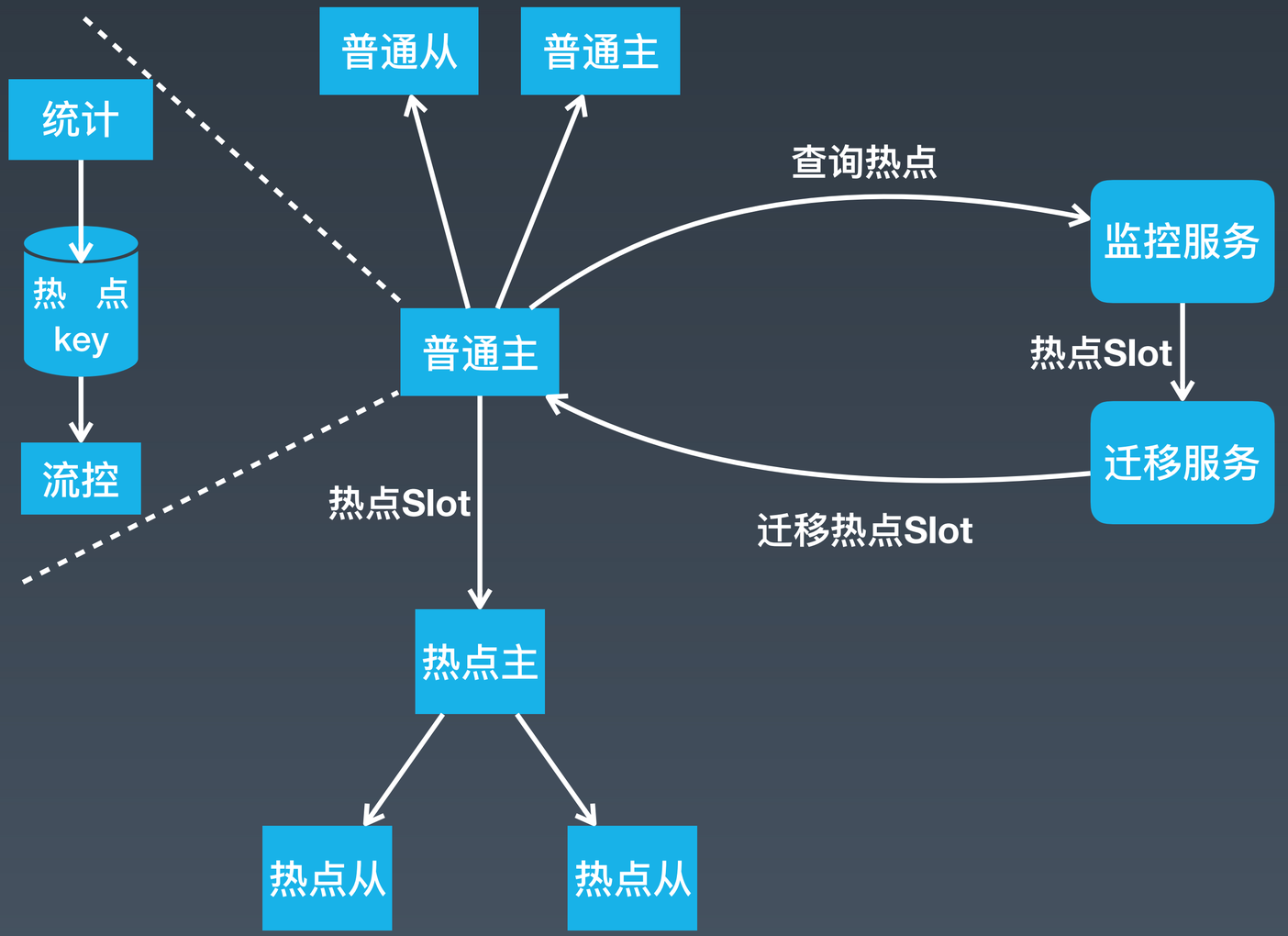
重构持久化机制

Squirrel—持久化重构



- 减少全量重传
- 减少并控制RDB
- 减少AOF写盘抖动
- 降低了数据可靠性

Squirrel—热点key

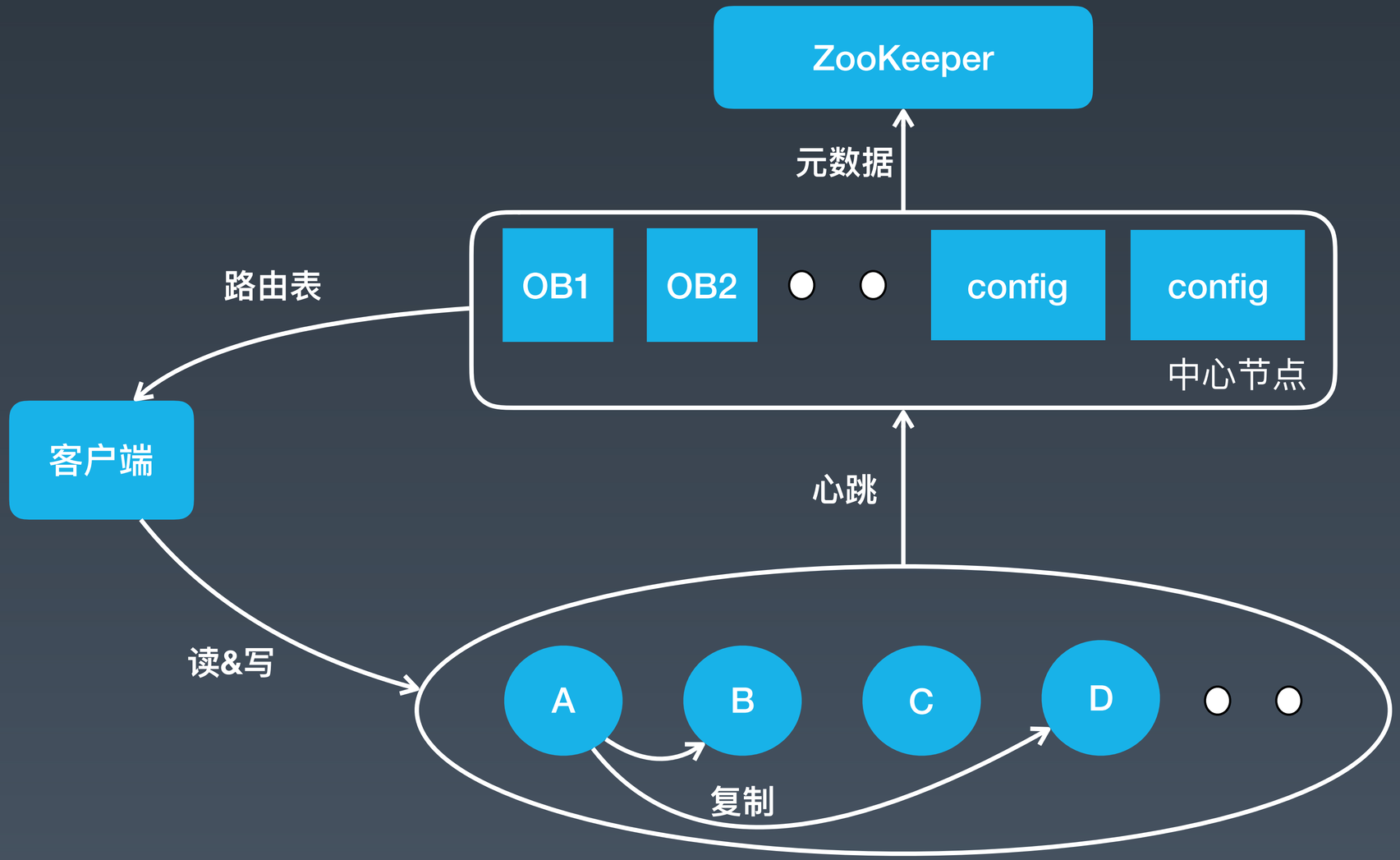


- 实时监控热点并止损
- 自动隔离热点并扩容

目录

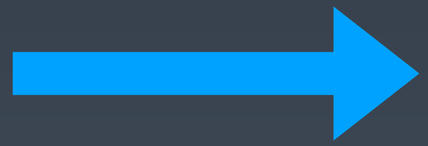
- 美团点评 KV 存储发展历程
- 内存 KV Squirrel 架构和实践
- 持久化 KV Cellar 架构和实践
- 发展规划和业界趋势

Cellar 架构和实践



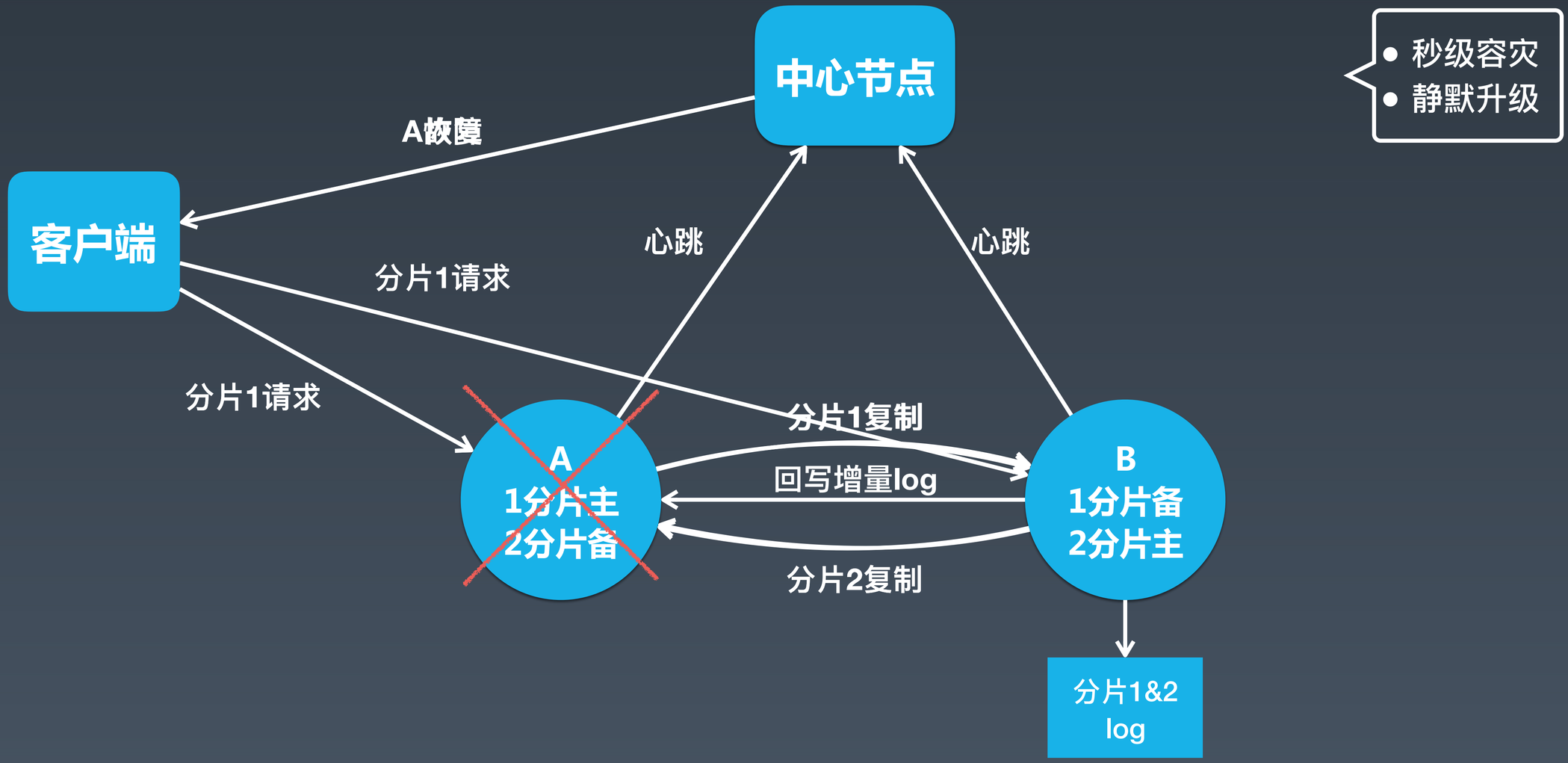
Cellar—节点容灾

- 想快速Failover却承担不起数据恢复的代价？
- 运维操作导致请求超时？



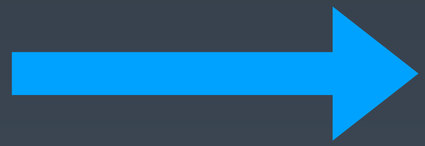
Handoff

Cellar—节点容灾


中心节点
客户端
A故障
心跳
心跳
分片1请求
分片1请求
分片1复制
回写增量log
分片2复制
A
1分片主
2分片备
B
1分片备
2分片主
分片1&2
log
秒级容灾
静默升级

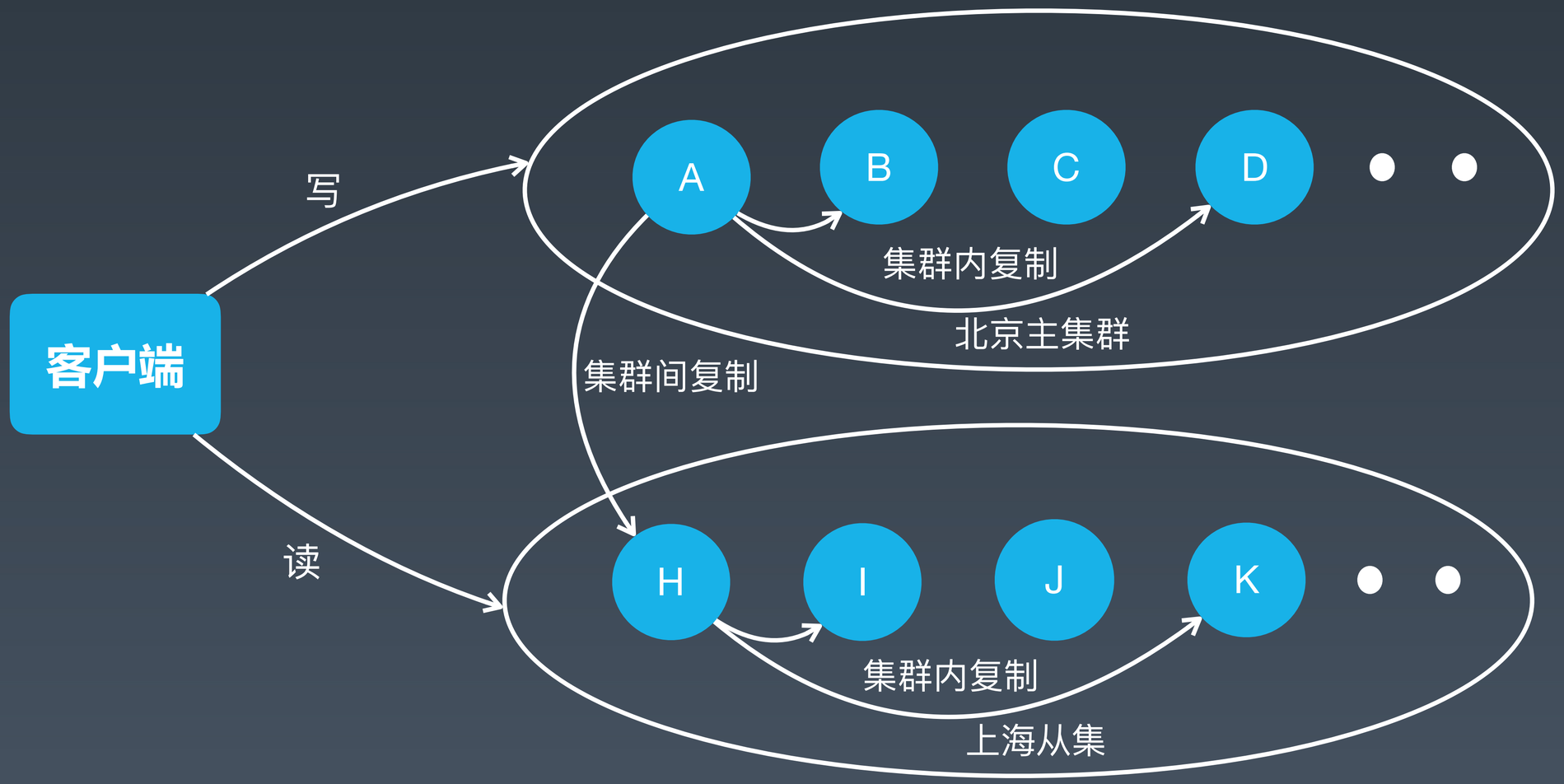
Cellar—跨地域容灾

- 跨地域专线不稳定
- 跨地域专线有限的带宽
- 单元化部署，多活架构



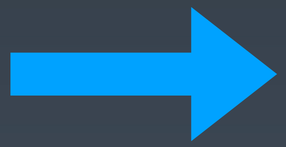
集群间复制

Cellar—跨地域容灾

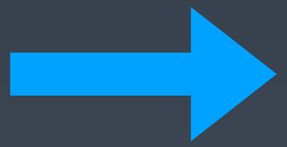


Cellar—强一致

支付等场景
数据不能丢

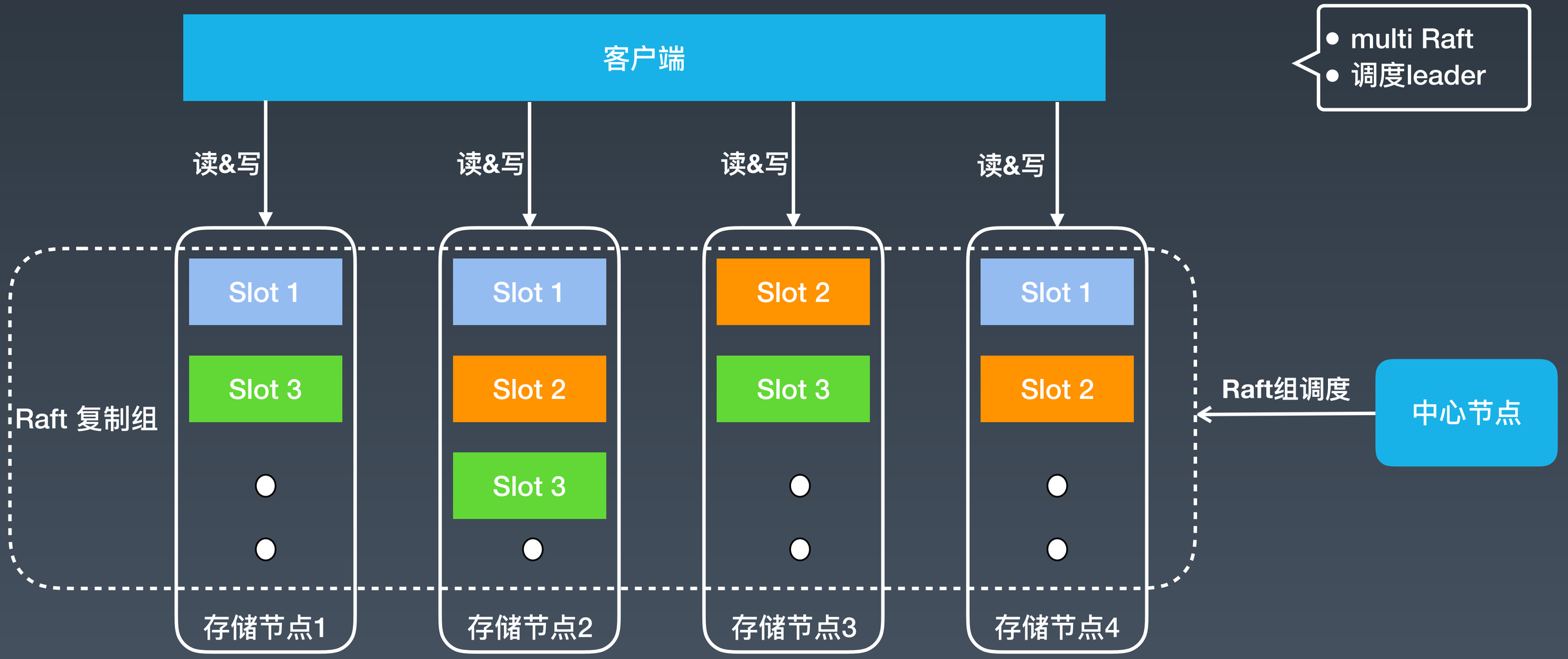


共识协议：
Paxos/Raft



Raft：
协议详实、工程实践

Cellar—强一致

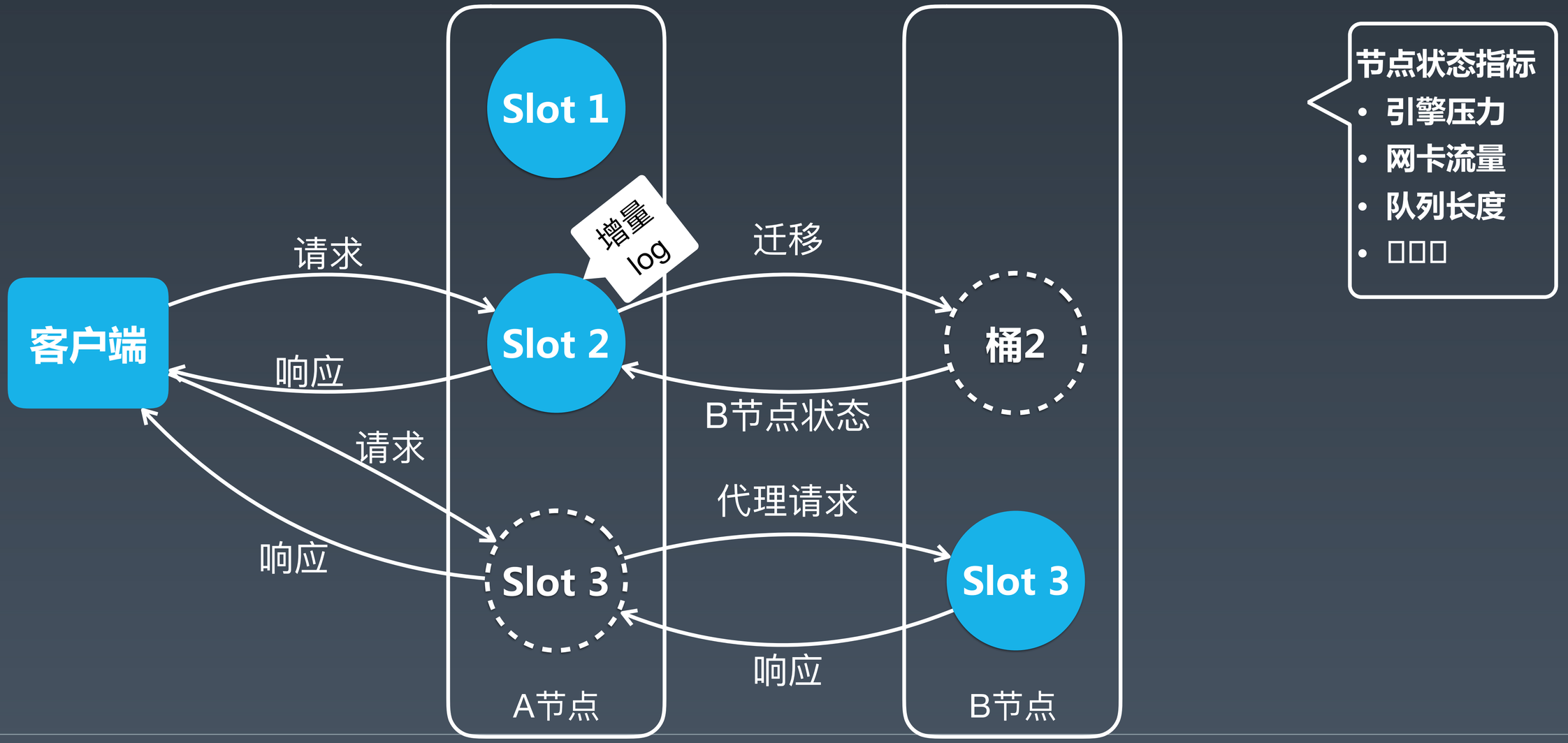

客户端
multi Raft
调度leader
读&写
读&写
读&写
读&写
Slot 1
Slot 3
Slot 1
Slot 2
Slot 3
Slot 2
Slot 3
Slot 1
Slot 2
Raft 复制组
Raft组调度
中心节点
存储节点1
存储节点2
存储节点3
存储节点4

Cellar—端到端成功率

影响端到端成功率的因素：

- **数据迁移影响业务请求成功率**
- **慢请求阻塞服务队列**
- **热点key请求导致单节点过载**
- **□□□**

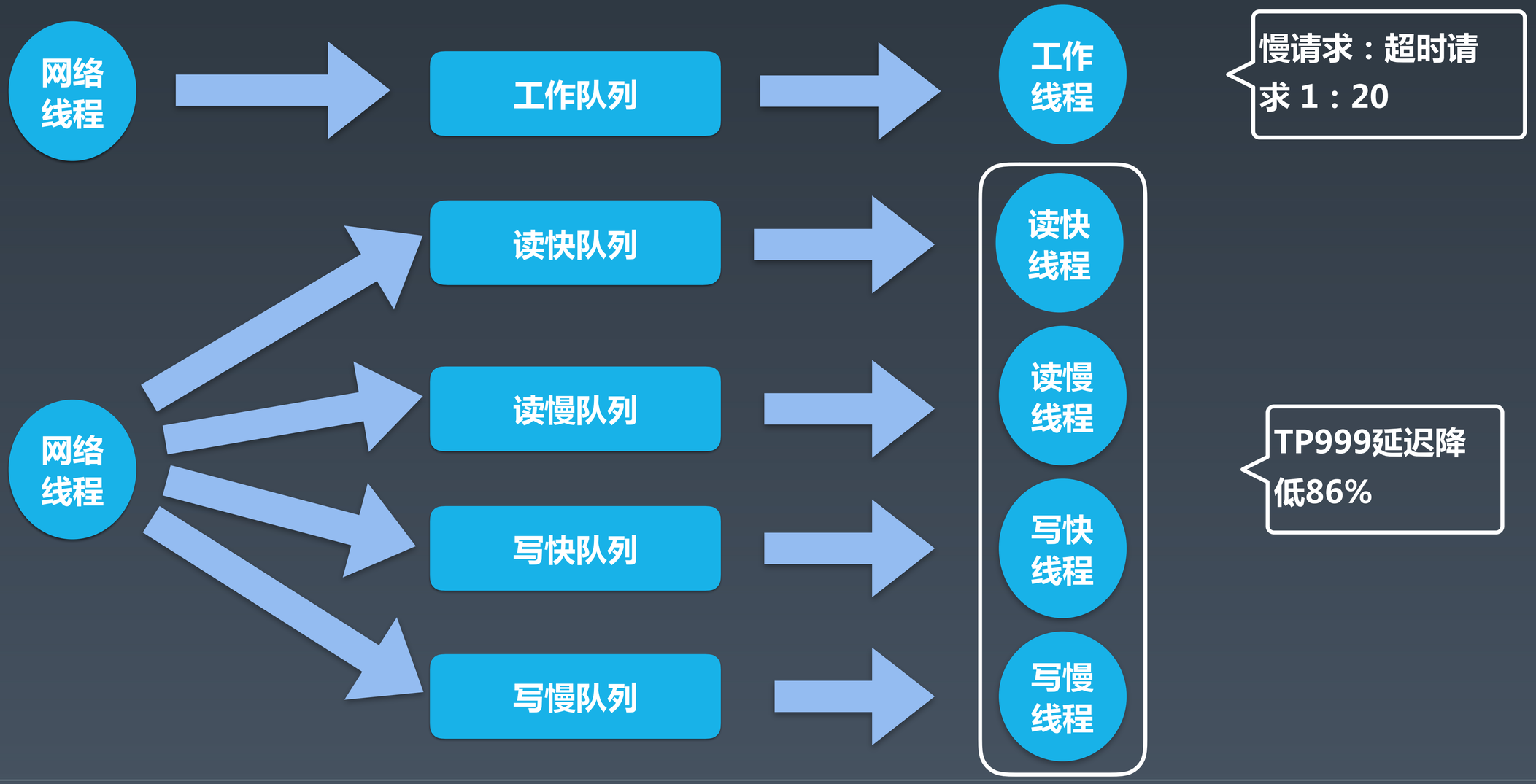
Cellar—智能迁移



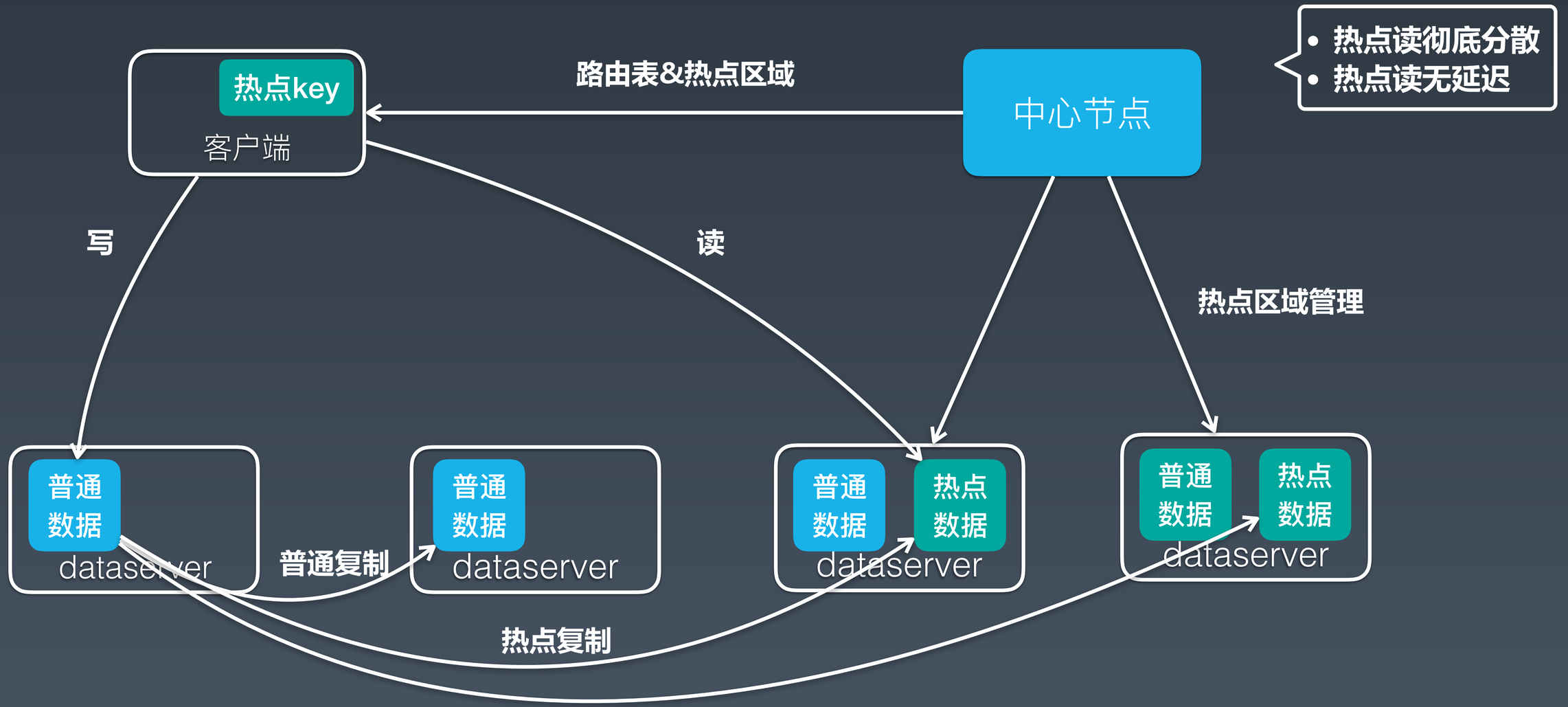
节点状态指标

- 引擎压力
- 网卡流量
- 队列长度
- □□□

Cellar—快慢队列


网络线程
网络线程
工作队列
读快队列
读慢队列
写快队列
写慢队列
工作线程
读快线程
读慢线程
写快线程
写慢线程
慢请求：超时请求 1：20
TP999延迟降低86%

Cellar—热点key



目录

- 美团点评 KV 存储发展历程
- 内存 KV Squirrel 架构和实践
- 持久化 KV Cellar 架构和实践
- 发展规划和业界趋势

发展规划和业界趋势

服务

- **Redis gossip 优化**
- **Cellar 中心节点 Raft**
- **Squirrel & Cellar API 统一**
- **□□□**

系统

- **Kernel bypass（DPDK、SPDK）**

硬件

- **RDMA**
- **3D XPoint（Optane、AEP）**
- **计算型硬件（SSD + FPGA）**

Geekbang 极客邦科技 | InfoQ 极客传媒

InfoQ官网 全新改版上线

促进软件开发领域知识与创新的传播

关注InfoQ网站
第一时间浏览原创IT新闻资讯

免费下载迷你书
阅读一线开发者的技术干货

THANKS!

QCon 10th

欢迎交流：qizebin@meituan.com