

www.qconferences.com



微信后台存储架构

许家滔


sunnyxu@tencent.com

微信产品部 - 基础平台组




大纲

微信通用存储系统（Quorumkv）

 前言（需求背景）

 系统概况

 系统架构（强一致性协议，存储模型，分片）

 真实系统

前言－微信分布

 上海 天津 深圳 香港 加拿大 ...

 同城多园区分布

系统概况

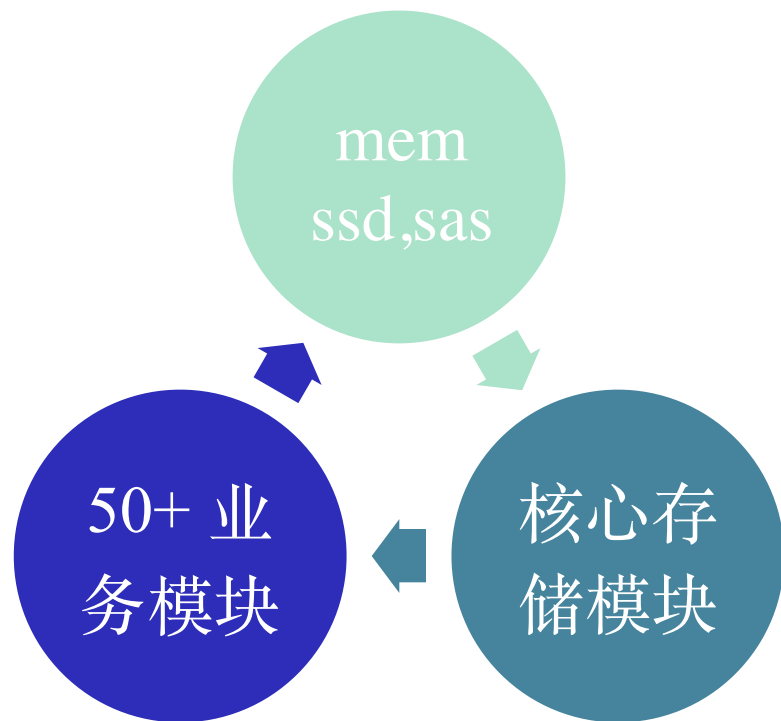
数据存储基础设施

分布式强一致

同城园区级容灾

支持类**SQL**查询

系统概况



系统架构 — 背景

A. 分布式问题收敛

后台逻辑模块专注逻辑，快速开发

可能读取到过时的数据是个痛点

需要看到一致的数据

系统架构 — 背景

B.内部定义

数据拥有两个以上的副本

如果成功提交了变更，那么不会再返回旧数据

系统架构 — 推演

1. 增加一个数据

系统架构 — 推演

2.序列号发生器，偏序

约束：只能有一个**client**操作

client有解决冲突的能力

问题转移：**client**如何分布？

系统架构

3.修改集群中一个指定key的value

1) 覆盖它

2) 根据value的内容做修改

if value = 1 then value : = 2

系统架构

1.通用解法:

1) paxos算法

工程难度

一切可控

系统架构 — 分布算法设计

2) Quorum算法 (2011)

在单个key上面运算

真实系统约束

类paxos方案，简化

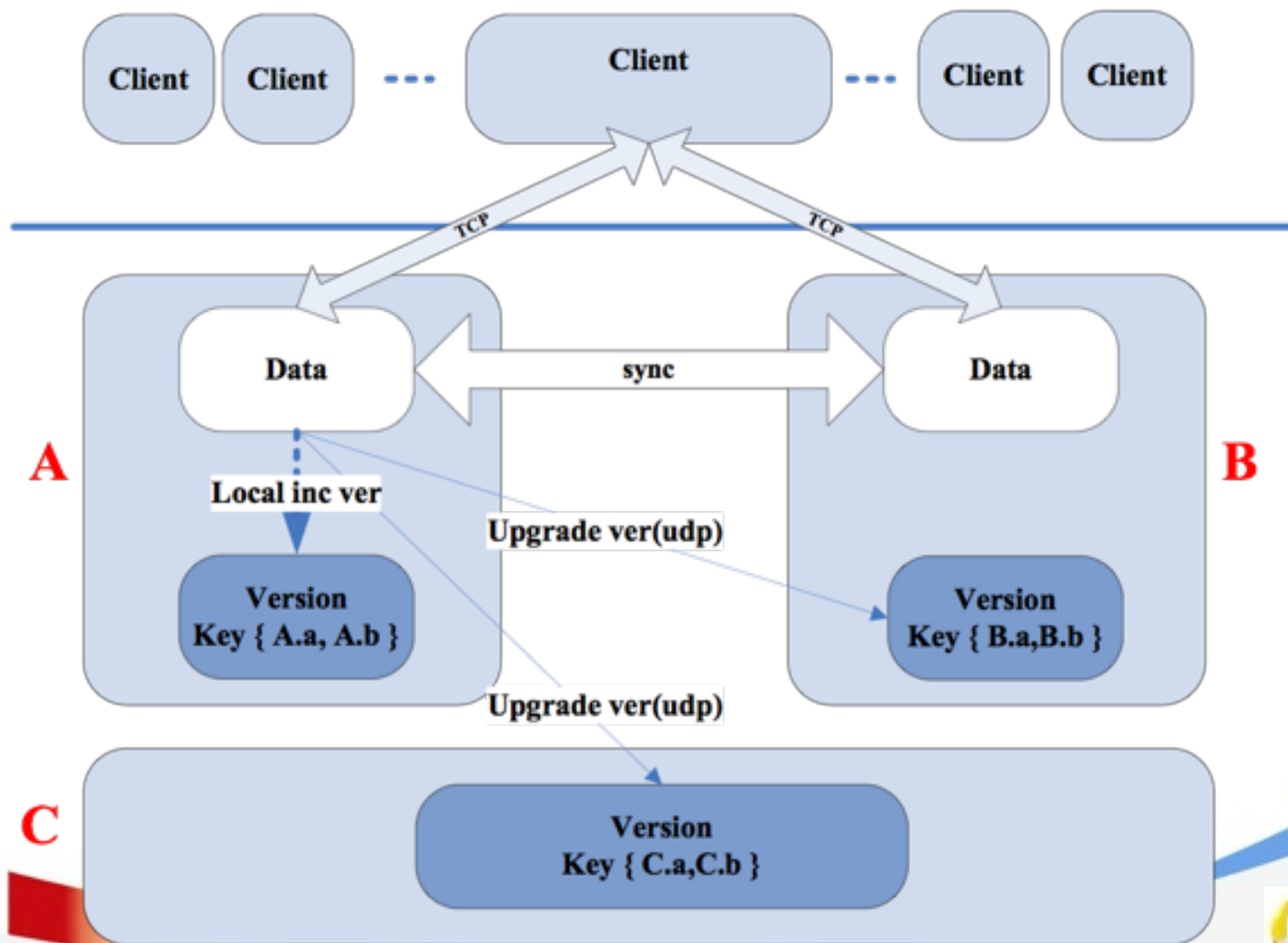
系统架构 — 分布算法设计

为每次变更选举(by key)

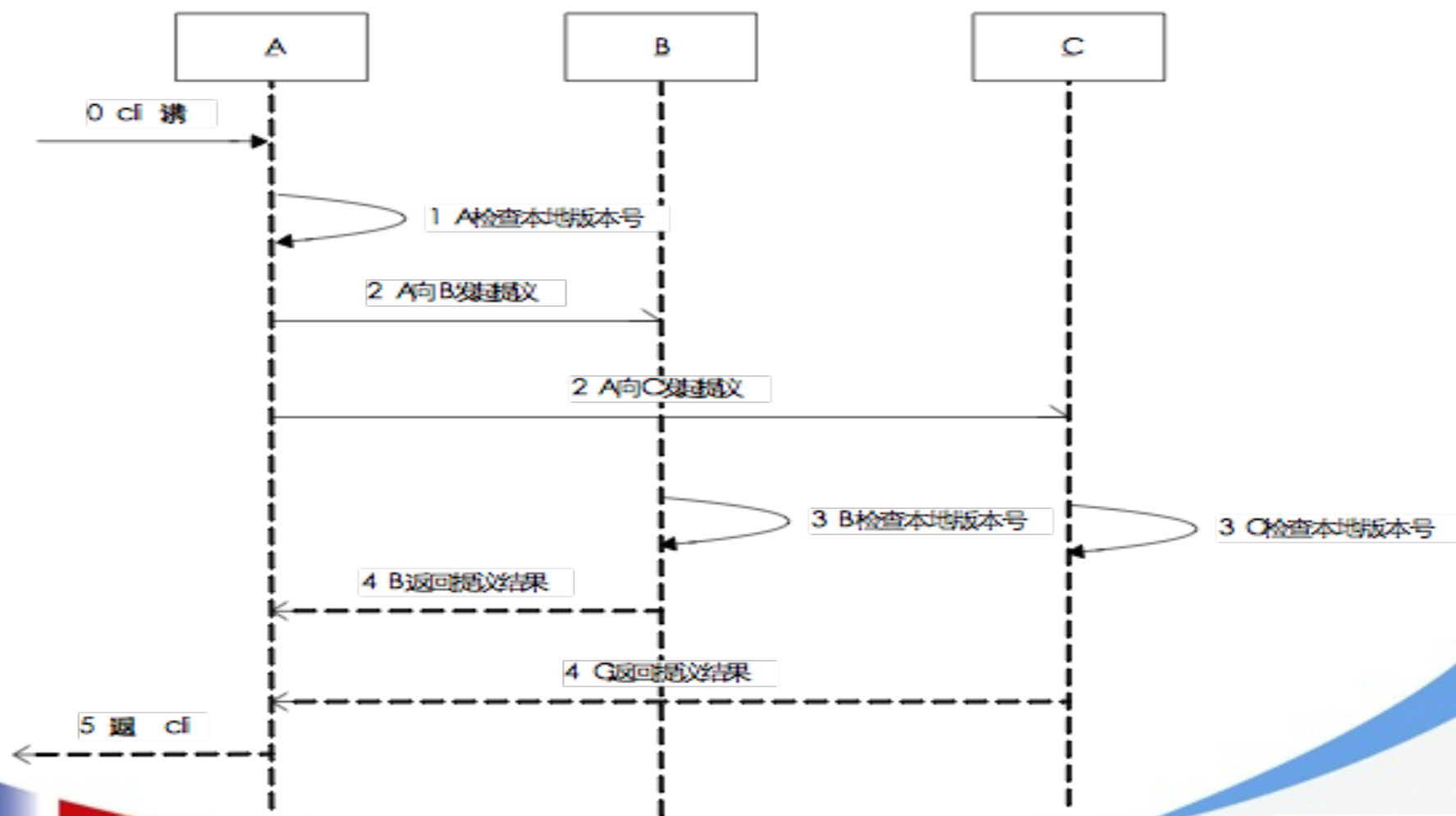
算法过程

提议 / 变更 / 同步 / 广播

系统架构



系统架构 — 写流程



系统架构 — Replication & Sharding

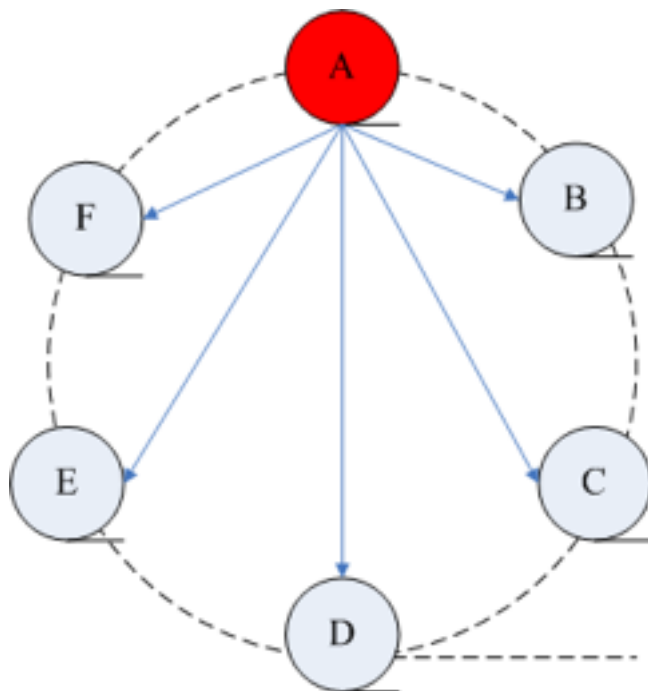
权衡点

自治，负载均衡，扩散控制
replication -> relation

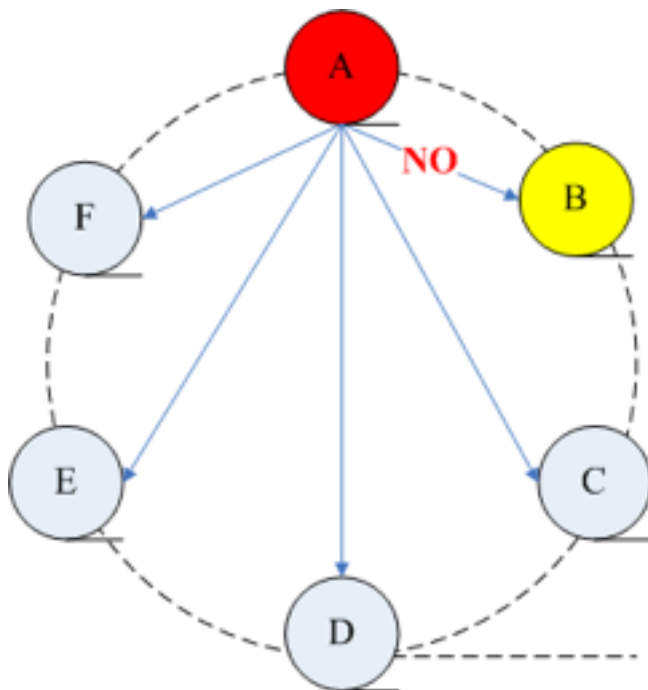
容灾低消

同城（上海）多数派存活
三园区（独立供电，独立。。）

系统架构 — Replication



系统架构 — Replication



系统架构 — Sharding

一组kv6为一个单位

1.人工分段

局部扩容，影响收敛

2.均匀分布 指定分段 hash32(string)

翻倍扩容

3.一致性哈希

具体实现?

系统架构 — 概览

1.业务侧快速开发

存储需要提供**强一致性**

丰富的数据模型支持 (结构化/类SQL/KV)

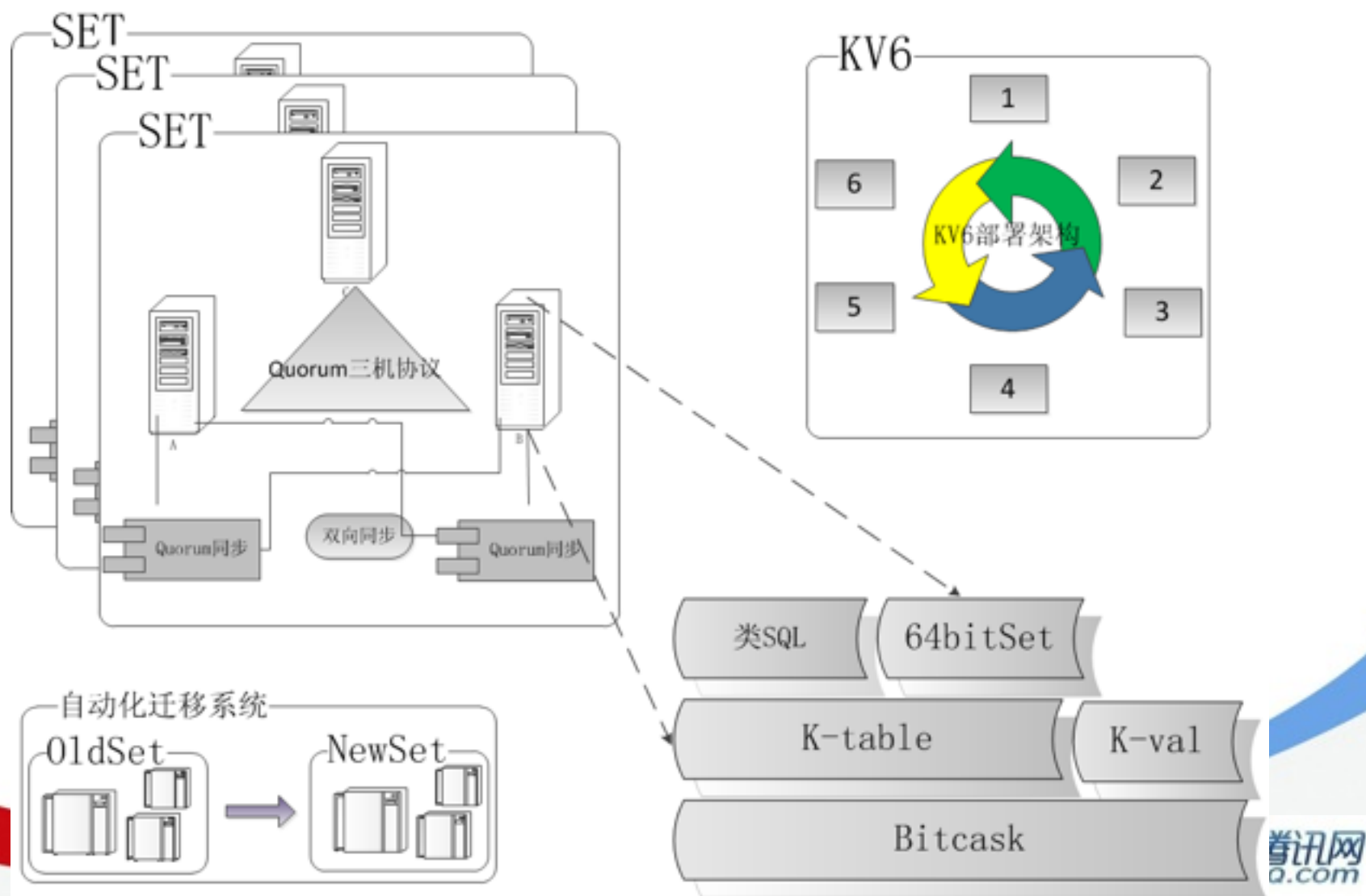
条件读，条件写

2.业务增长迅速，系统要能够方便地横向扩容

3.设备故障/短时节点失效成为常态，容灾自动化，主备可写无需人工介入

4.小数据

系统架构 — 概览



系统架构 — 存储模型

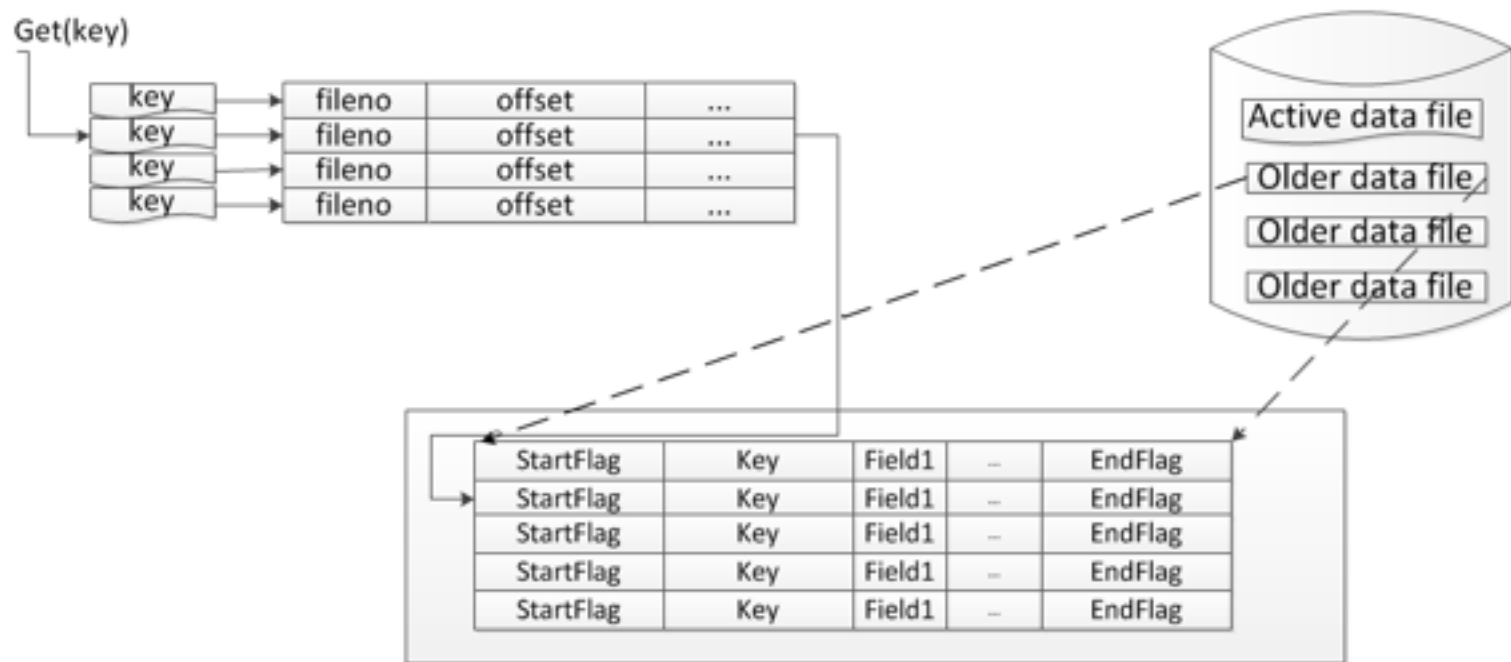
纯内存

Bitcask

小表系统

LSM-tree

系统架构 — bitcask



系统架构 — 小表系统

解决写放大问题

数据按变更聚集存储

Affected 1

ChangeTable

$(1+2+\dots+n-1 + \text{total}) / n$

分裂与合并

系统架构 — 数据流动

 自动化迁移

 节点同时做代理

 合并磁盘io

真实系统 — 同步流量

 同步流量

 数据 vs 操作


 幂等


 保底策略

真实系统 — 通信包量

 动态合并

 100k qps

 200% — 10%

 权衡与估算

 设计要点

真实系统 — 吞吐量

 异步化

 复杂度

 libco

真实系统 — 自动修复系统

 不要让错误累积

 全量扫描

其他

bitcask的一些变化

内存限制

全内存



Q & A

sunnyxu@tencent.com

Brought by **InfoQ**