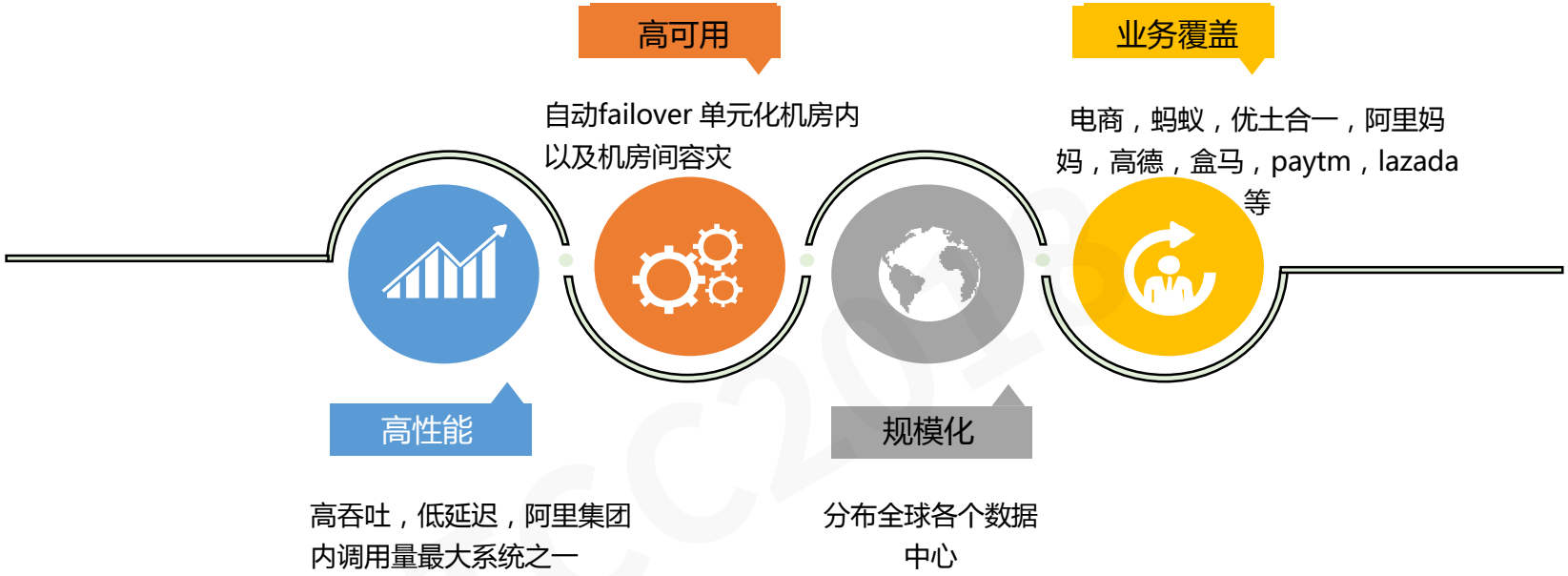
我们是怎么支撑双11万亿流量的
—— 阿里分布式缓存(Tair)技术分享

姜志峰

Agent

- 阿里自研大规模分布式缓存 – TAIR
- 技术挑战
- 性能与成本
  - 单机能力提升
  - 客户端吞吐
  - 业务方案
- 热点问题

大规模分布式缓存 – TAIR



2010.04 诞生，MDB内存存储产品，满足缓存需求
2012.06 LDB持久化产品，满足持久化存储需求
2012.10 RDB缓存产品，满足复杂数据结构的存储需求
2013.03 Fastdump产品， 应对全量导入场景，大幅度降低导入时间和访问延时
2014.07 专注于性能提升
2016.11 智能运维，单元弹性，千亿流量
2017.11 性能飞跃，热点散列，资源调度，支持万亿流量

TAIR在阿里的应用

- 缓存，降低对后端数据库的访问压力，
  会员，session，库存，购物车，优惠等
- 数据存储，允许部分数据丢失

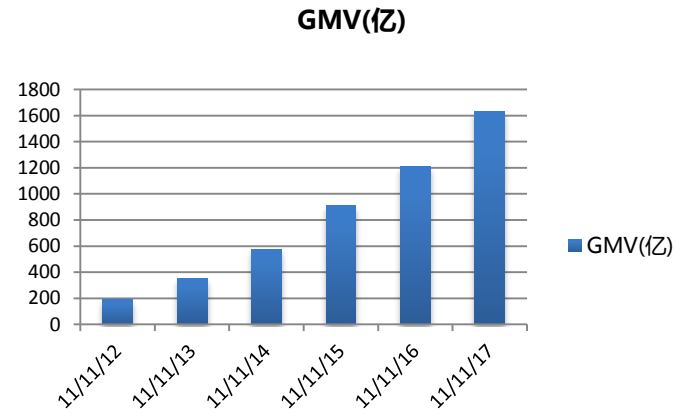
- 离线数据导入，在线访问
- 读取低延迟，不能有毛刺

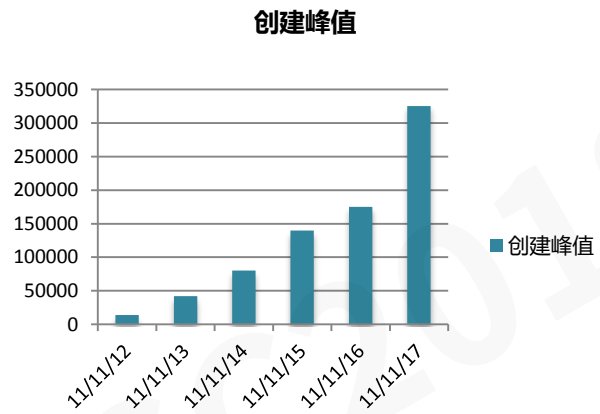
- 通用kv存储，交易快照，安全风控等
- 存储黑白单数据，读qps很高
- 计数器功能，更新非常频繁，数据不可丢失

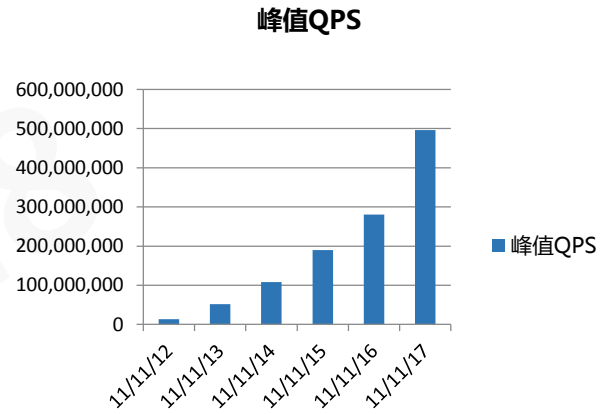
- 复杂的数据结构的缓存与存储
  - exhash， vers-string， bloom，增强的GIS/GEO
- 更灵活的异地多活

阿里体系下缓存的技术挑战

增长：缓存TAIR峰值 >> 交易峰值 >> 总GMV





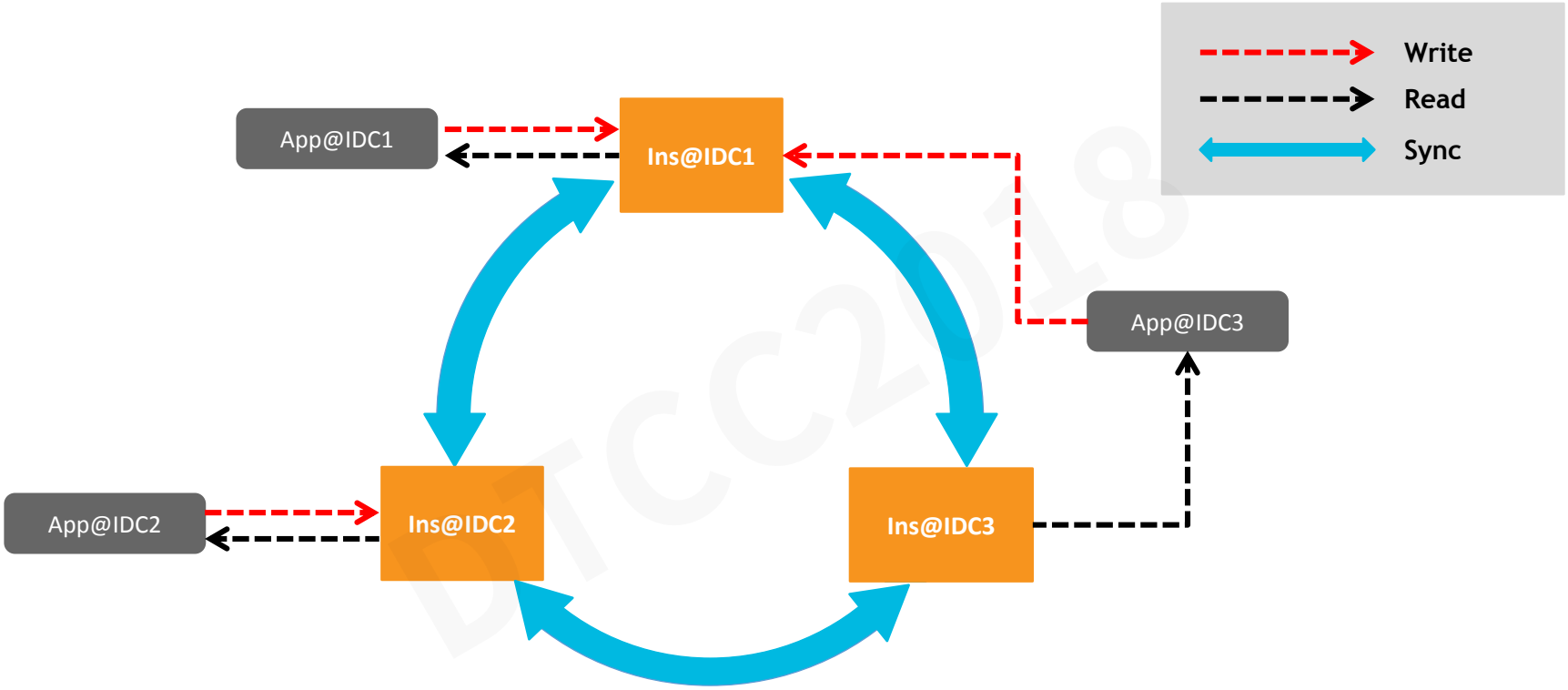


- 更强的容灾能力 → 异地多活与单元化
- 持续服务能力 → 24*365的稳定性
- 极致的RT需求 → 为了更好的体验
- 成本上的要求
  - 性能提升
  - 弹性扩展与资源调度

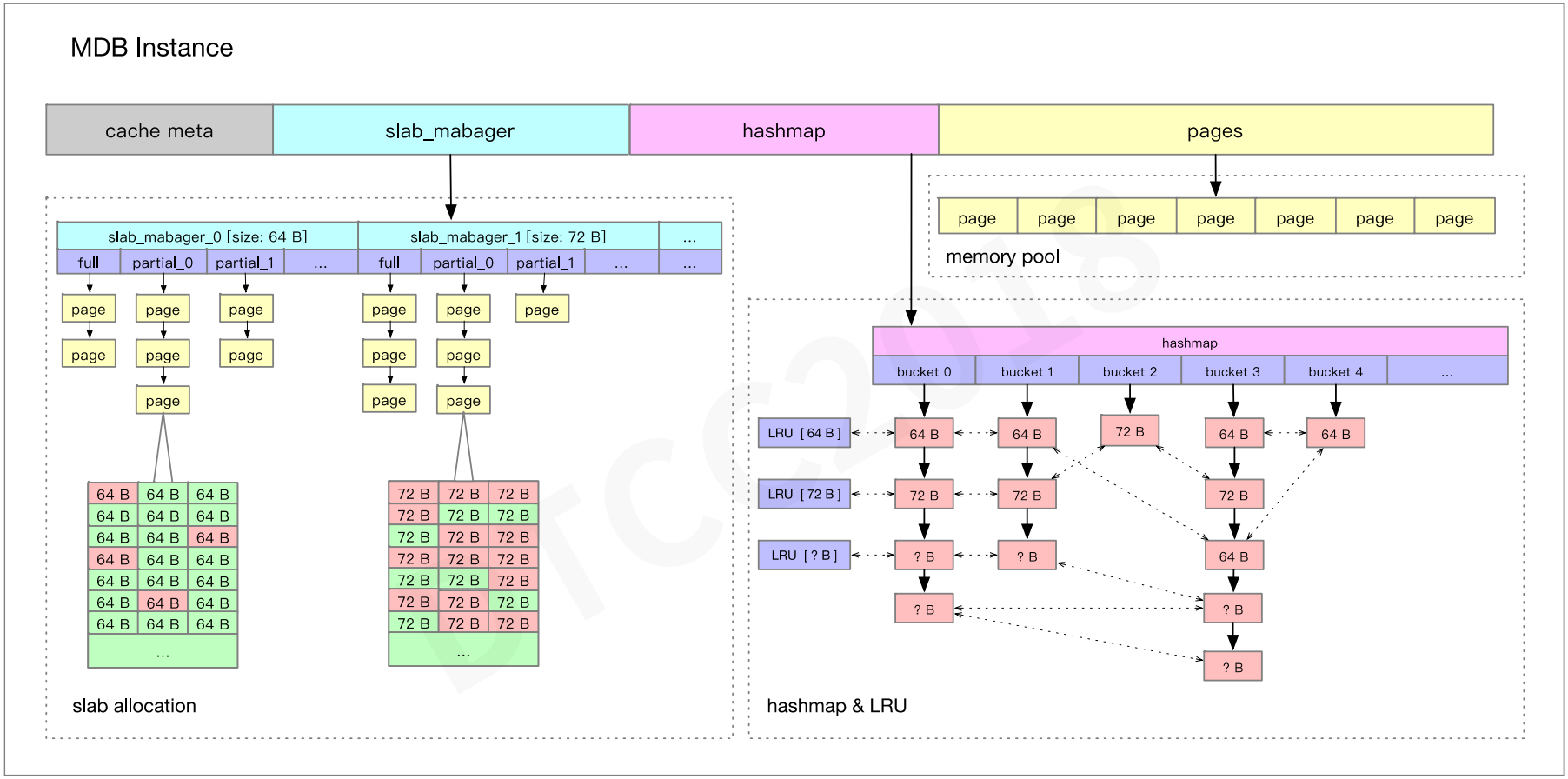
异地多活与单元化


单元A
机房
统一接入层
应用层
中间件
TAIR
DB
中心机房
机房
统一接入层
应用层
中间件
TAIR
DB
机房
统一接入层
应用层
中间件
TAIR
DB
单元B
机房
统一接入层
应用层
中间件
TAIR
DB

跨单元同步和多活(LDB/RDB)

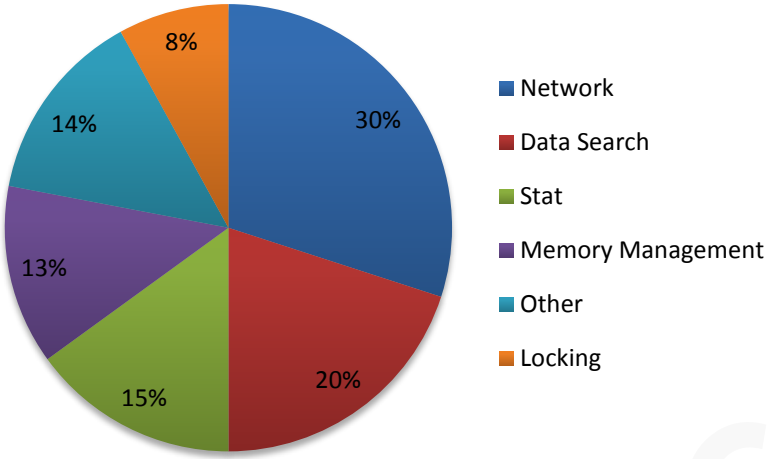


分析服务端性能瓶颈

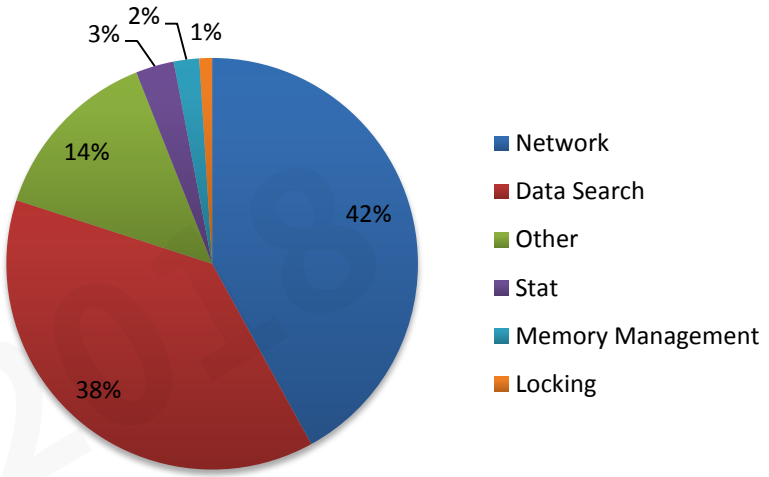


锁是阻止系统性能提升的绊脚石

服务端去锁优化



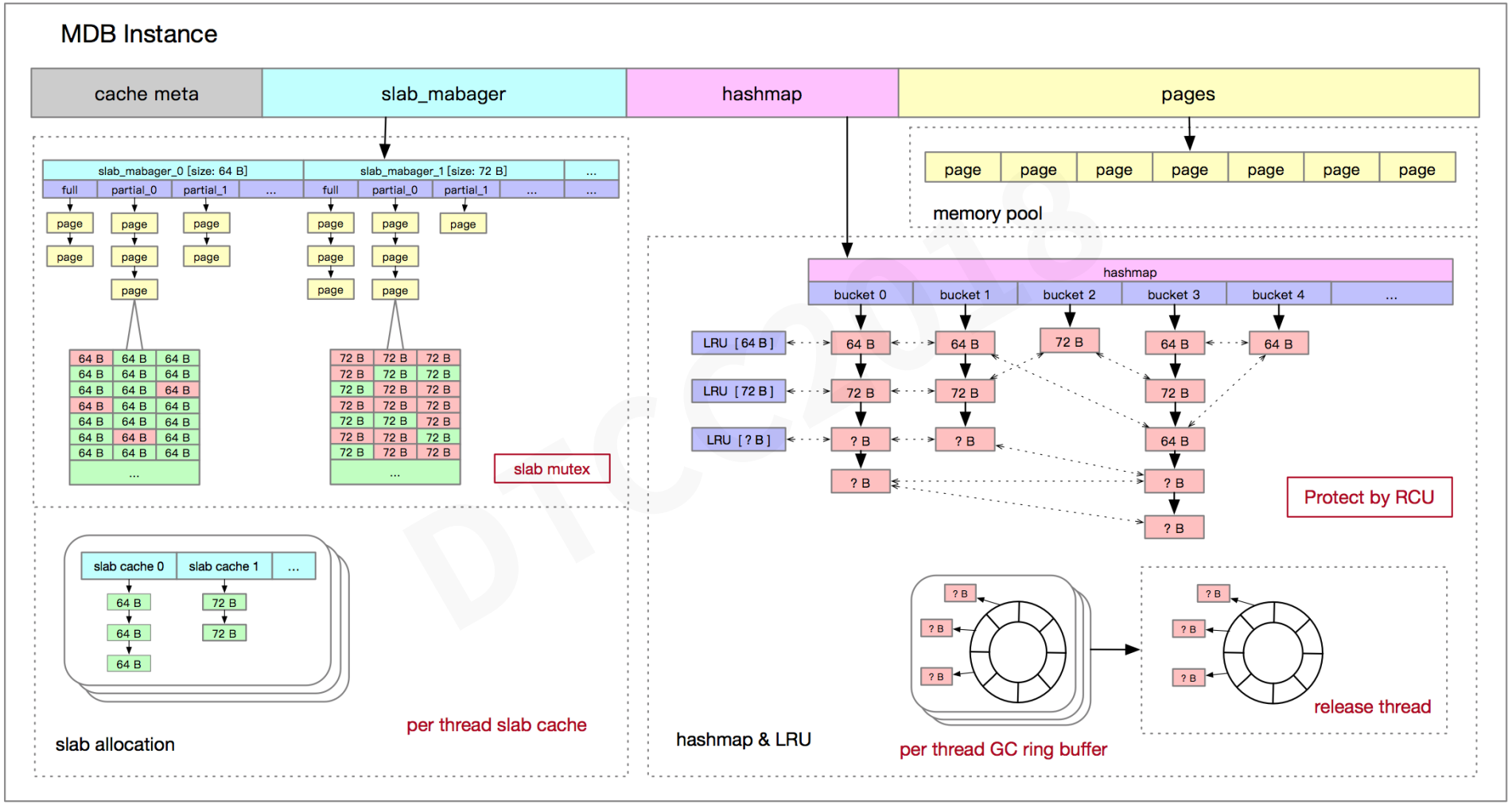
优化前



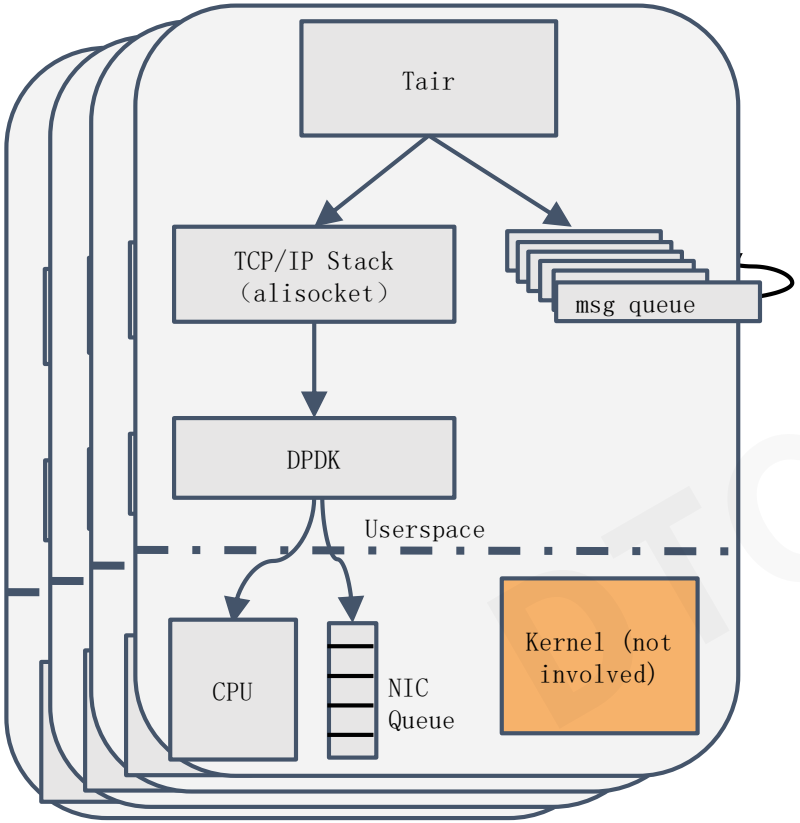
优化后

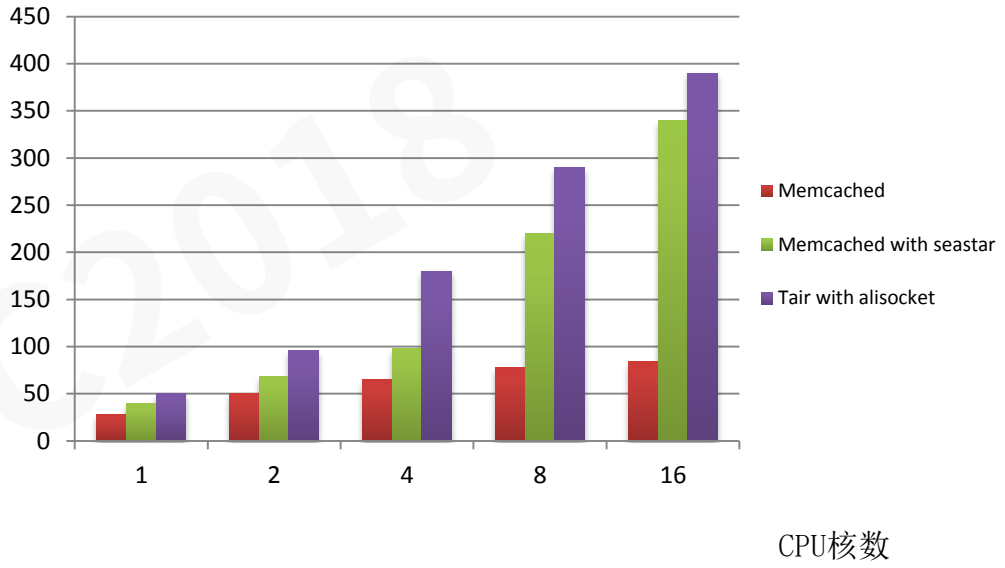
- 细粒度锁（fine-grained locks）
- 无锁数据结构（lock-free data structures）
- CPU本地数据结构（per-CPU data structures）
- 读拷贝更新（RCU）

优化后



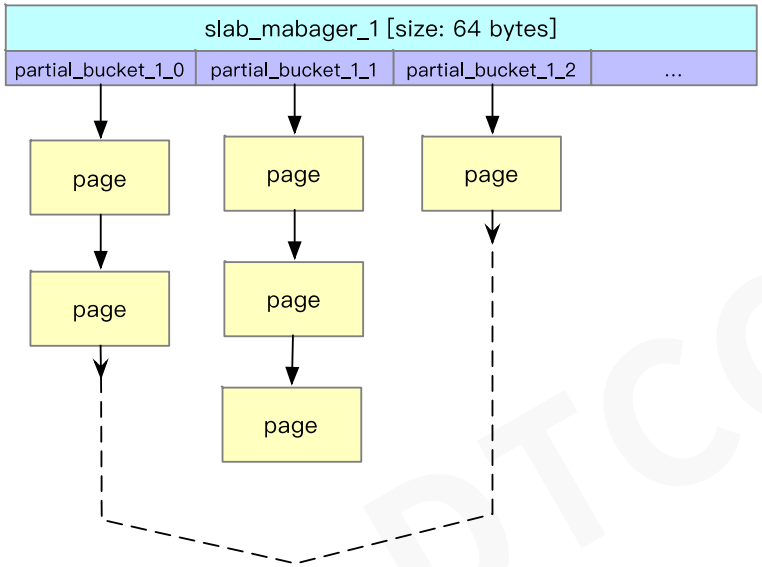
用户态协议栈(Tair + Alisocket)



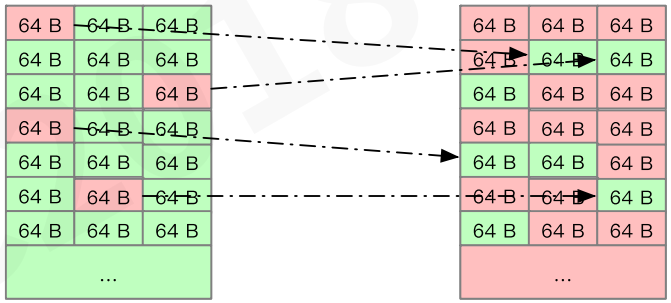


内存碎片合并提升使用率

MDB Instance

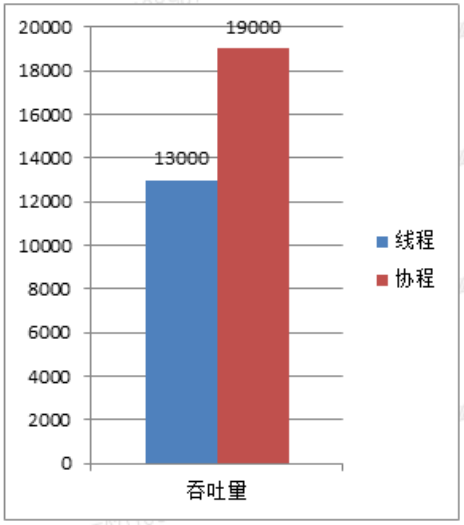


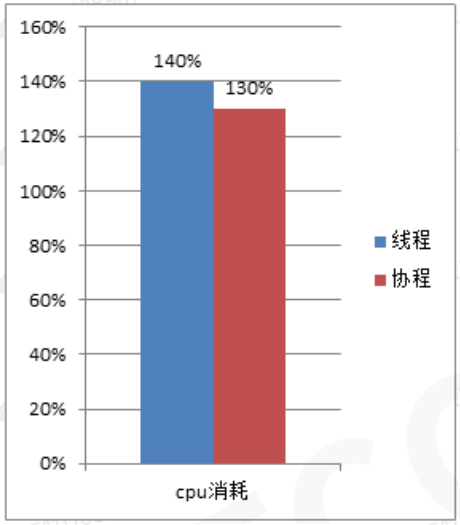
从两端开始向中心做数据移动

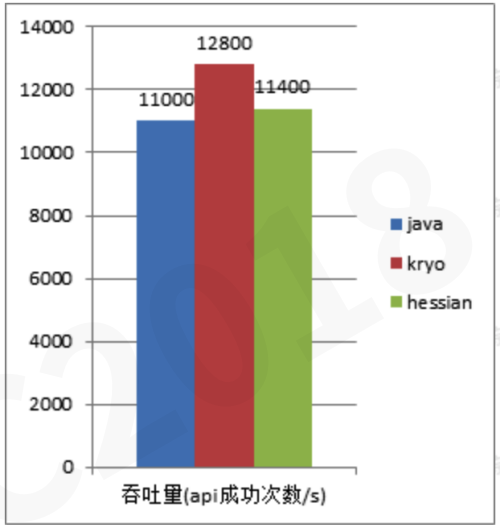


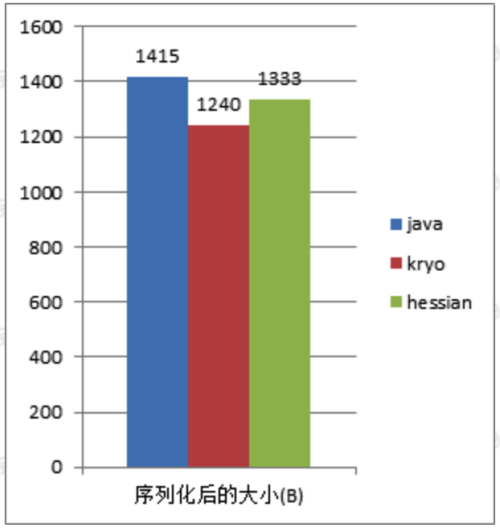
数据移动后修改所有数据结构的关联关系，释放空的内存页

提升客户端的吞吐







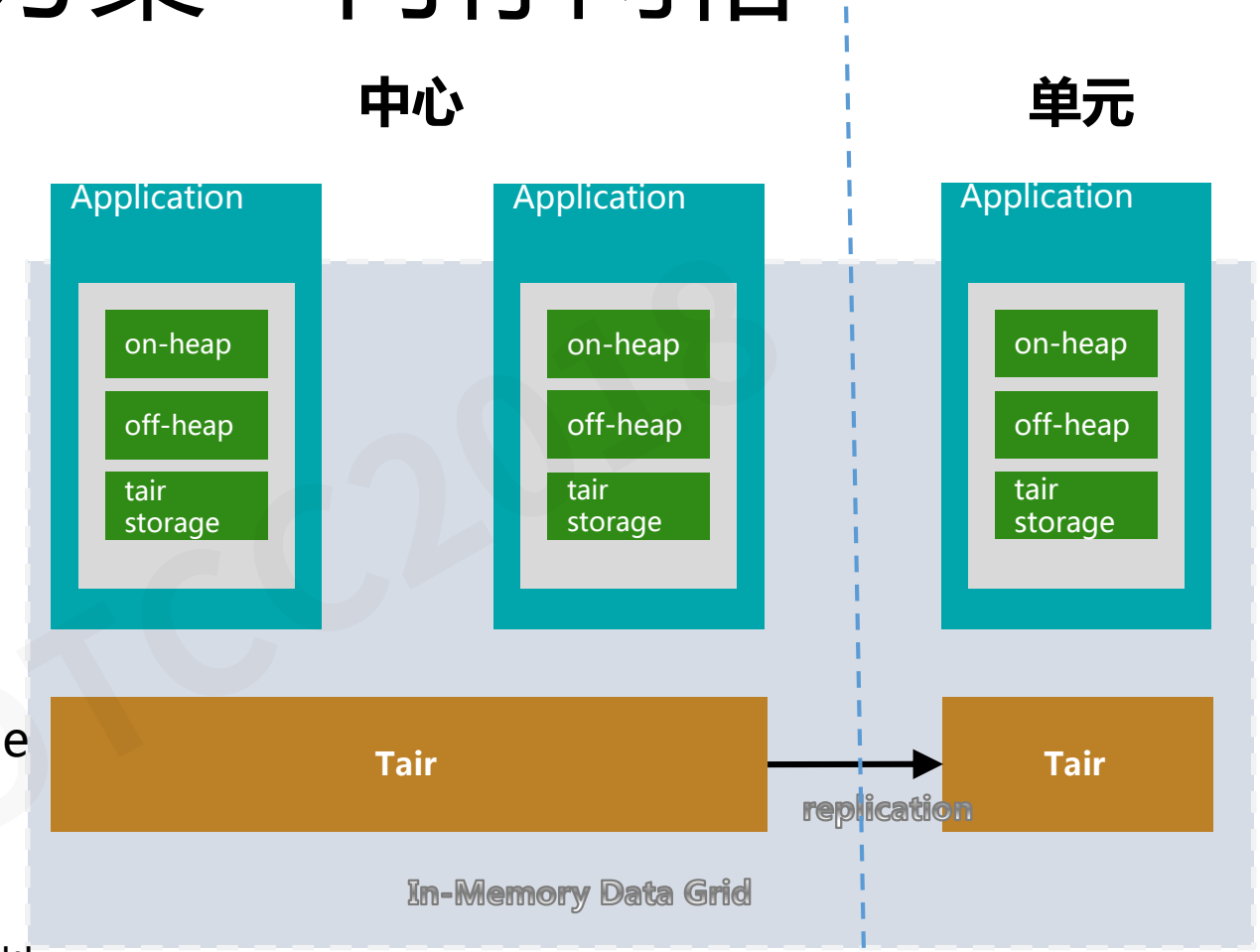


- 网络框架替换，适配ajdk协程
  - **mina ⇒netty**
  - 吞吐量提升40%+

- 序列化优化
  - 集成kryo和hessian
  - 吞吐量提升16%+

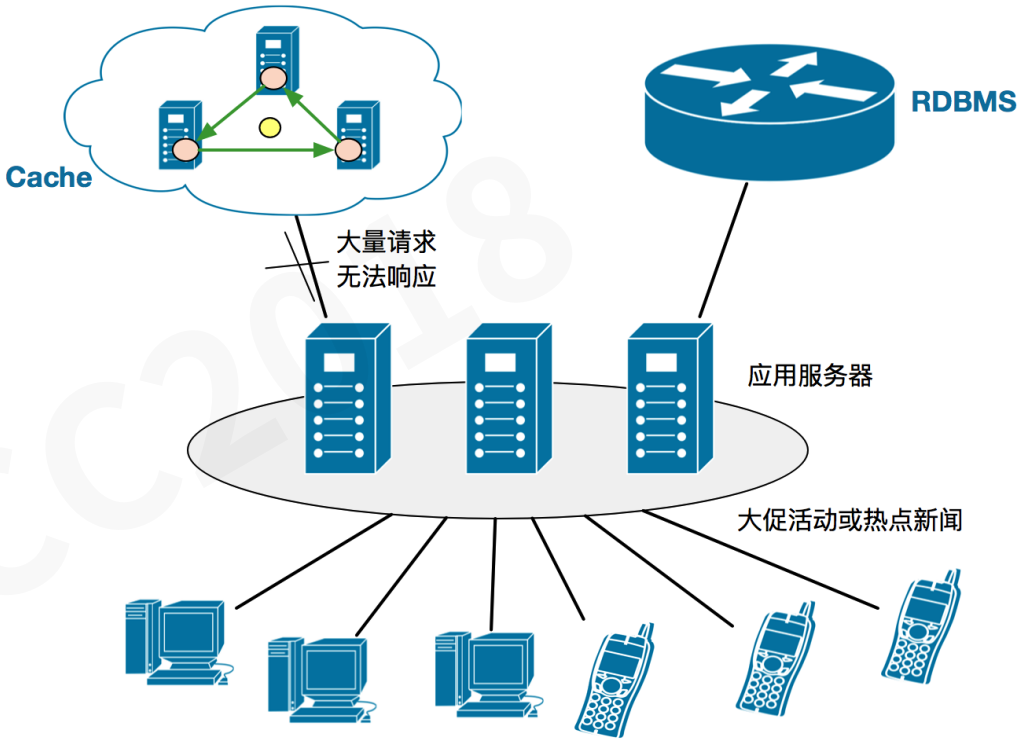
业务解决方案—内存网格

- 场景
  - 读写量超大
  - 大量本地计算
  - 高性能计算快速IO
- 方案
  - 数据本地性
  - 读穿透
  - Write Through
  - Write Behind/merge
  - 多单元replication
- 效果
  - 50%以上的计算在本地



热点导致集群能力短板

- 突发流量
  - 热门商品，店铺
  - 时事新闻
  - 各类压测
  - ...
- 缓存
  - 数据分片仍是单点
  - 单机单key能力是有限的
  - 被击穿
- 结局
  - 全系统崩溃
- 根源
  - 突增的访问热点

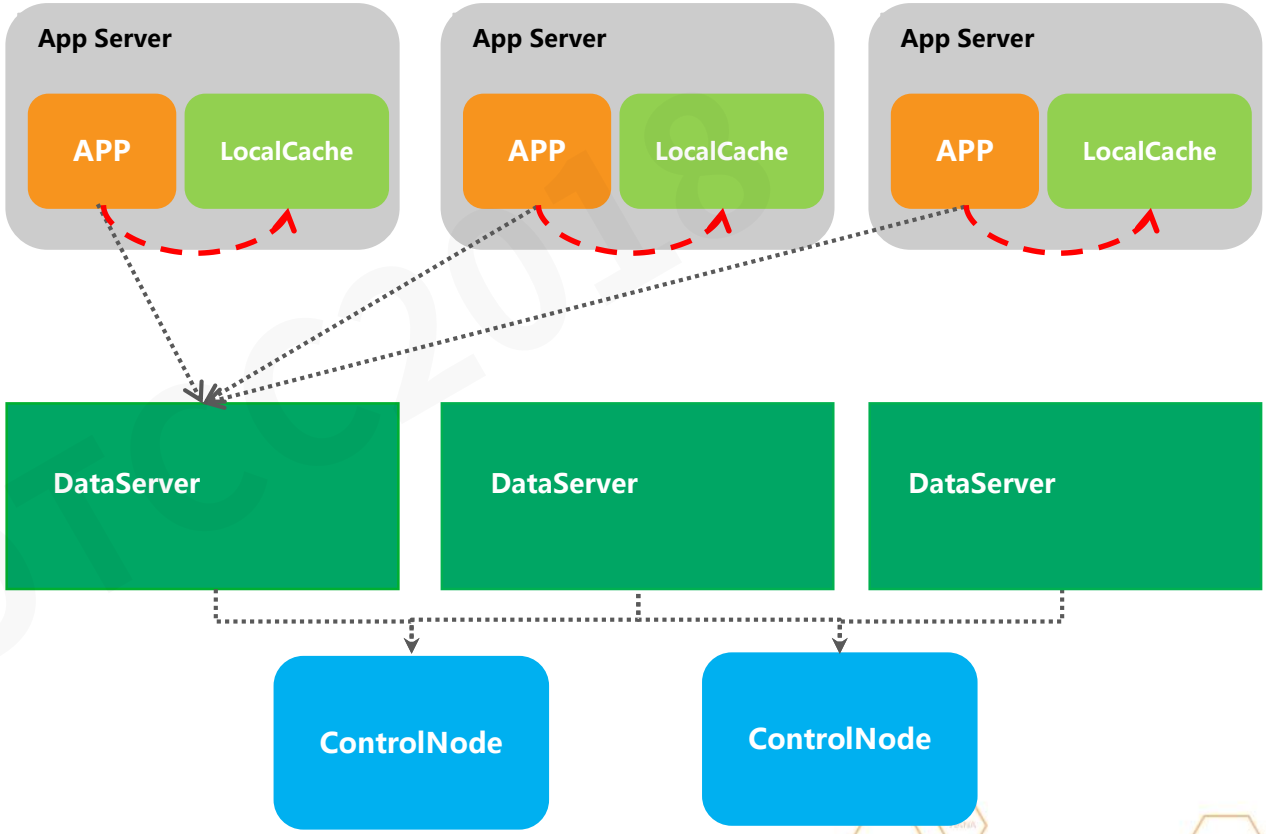


往年 – 热点预测

- 方案
  - 预测热点
  - Localcache
  - 热点拆分
- 效果
  - 如果能未卜先知还是可以的
  - 突发热点，就死扛吧

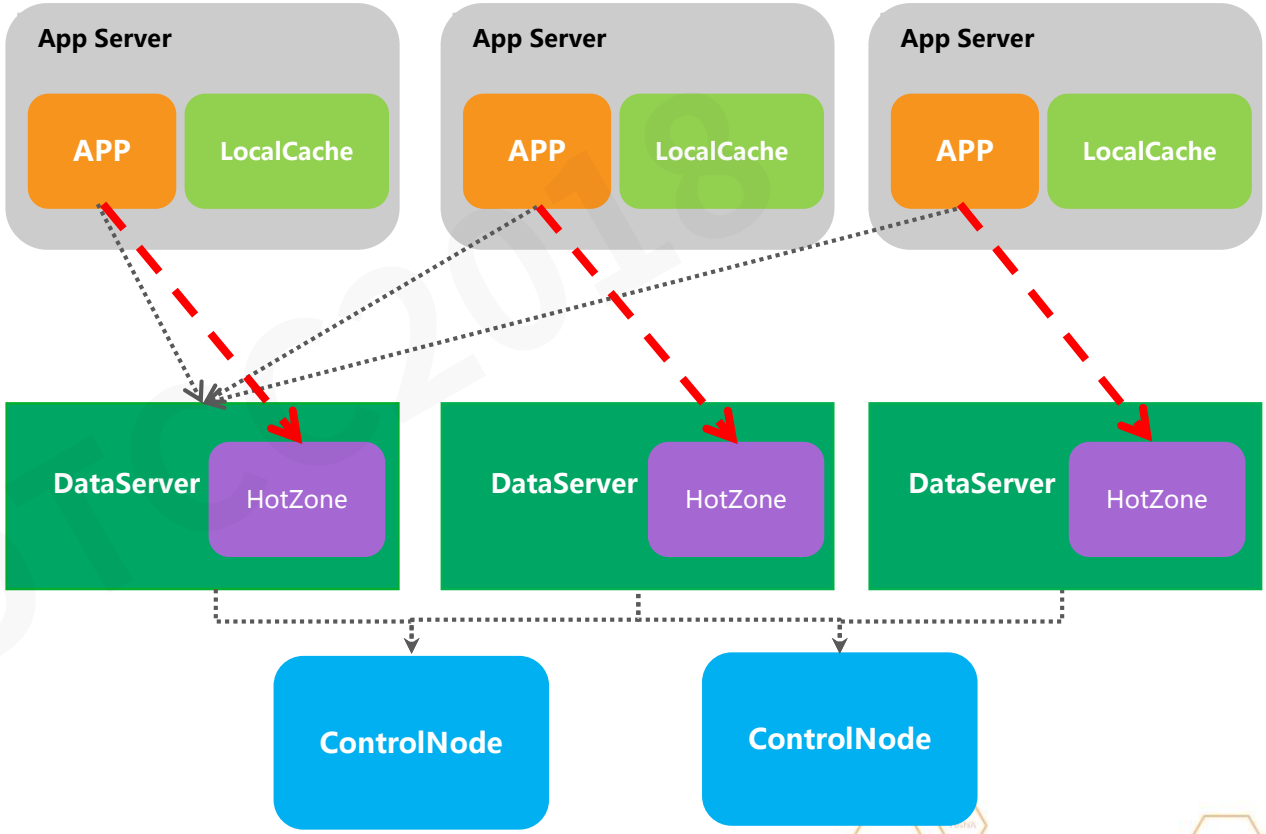
2016 - localcache + 热点识别

- 方案
  - 识别->localcache加速
  - localcache命中
- 效果
  - 解决了大部分场景下的热点问题
  - 带来了额外的客户端资源消耗
  - 应用机器能力总归是很赢弱的
  - prefix热点场合下的频繁置换
  - 并不完美的热点算法



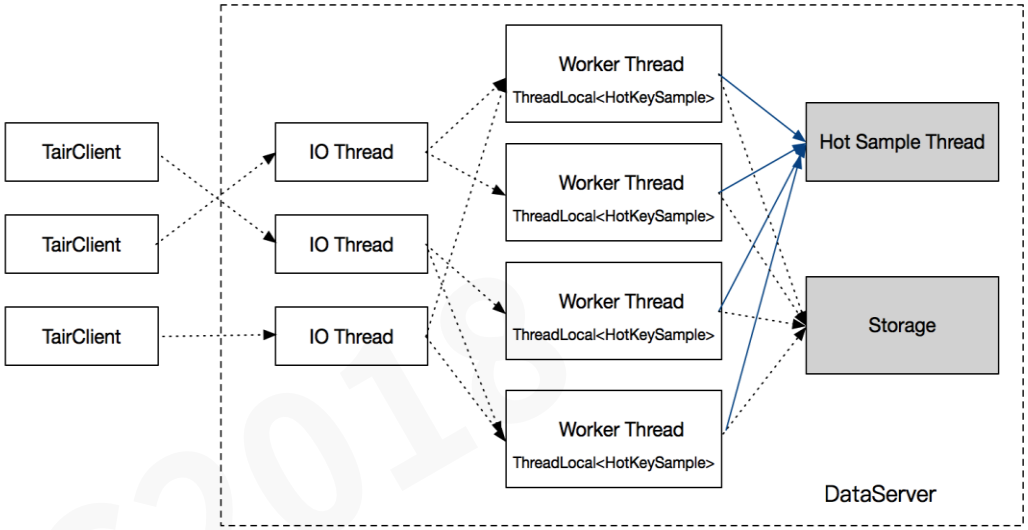
2017 – 热点散列

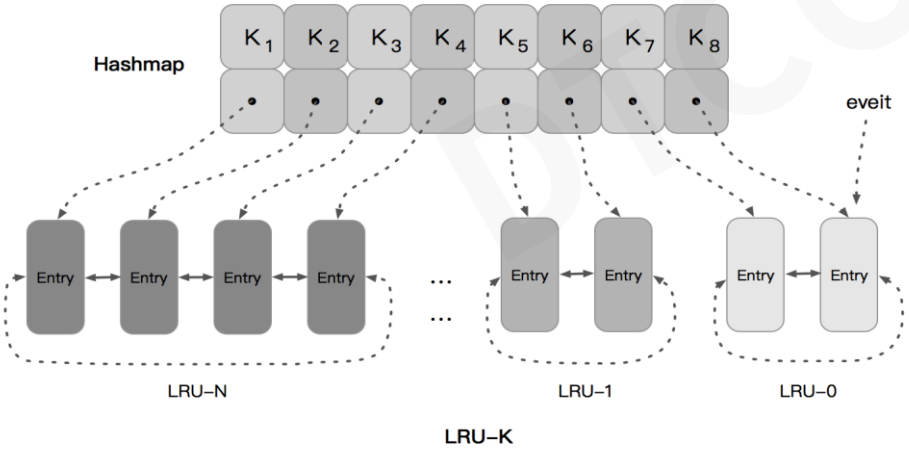
- 方案
  - 识别->散列(读)/合并（写）
  - HotZone缓存
- 效果
  - 服务端能力来支撑
  - 不消耗客户端额外资源
  - 无额外运维成本
  - 热点秒级识别
  - 热点能力水平扩展



热点的识别

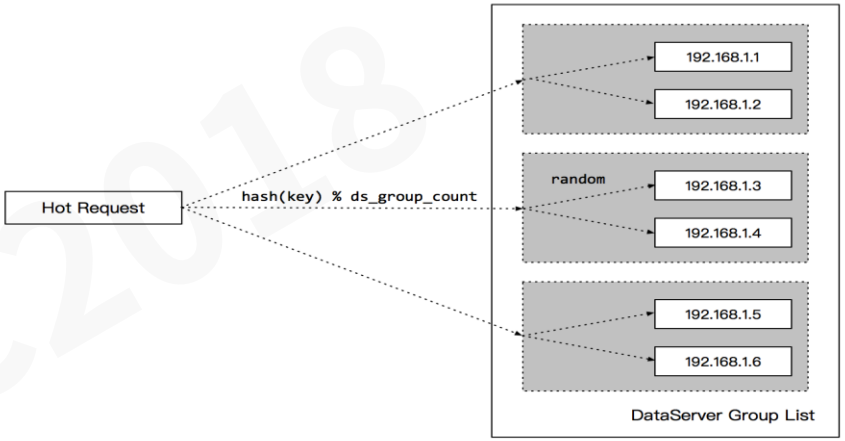
- 做到全样统计
- 流量热点的识别





读热点的处理

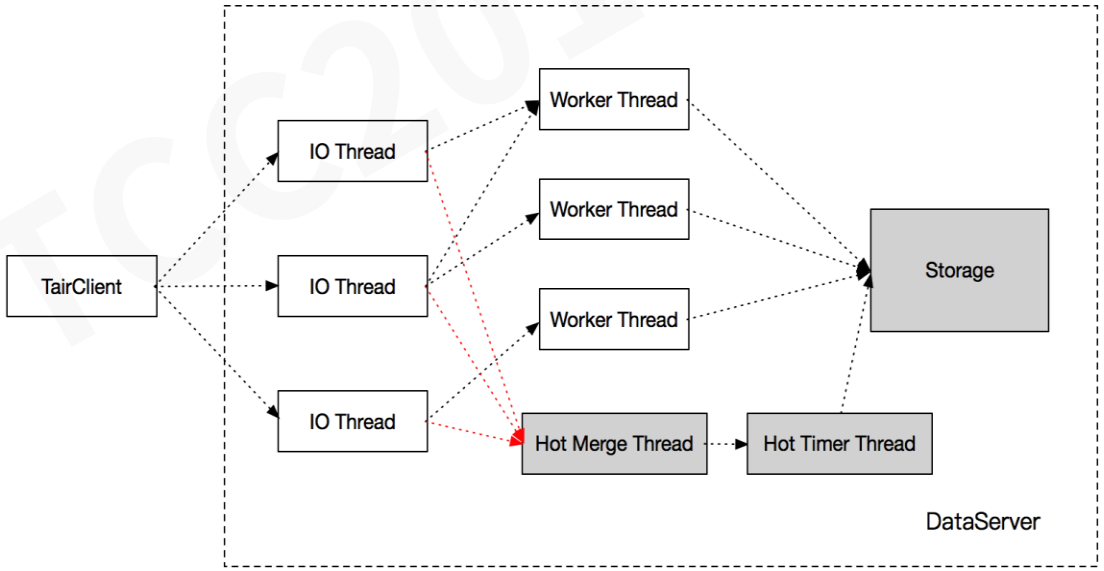
- 客户端选择HotZone还是本地Off-heap Cache
- 如何散列



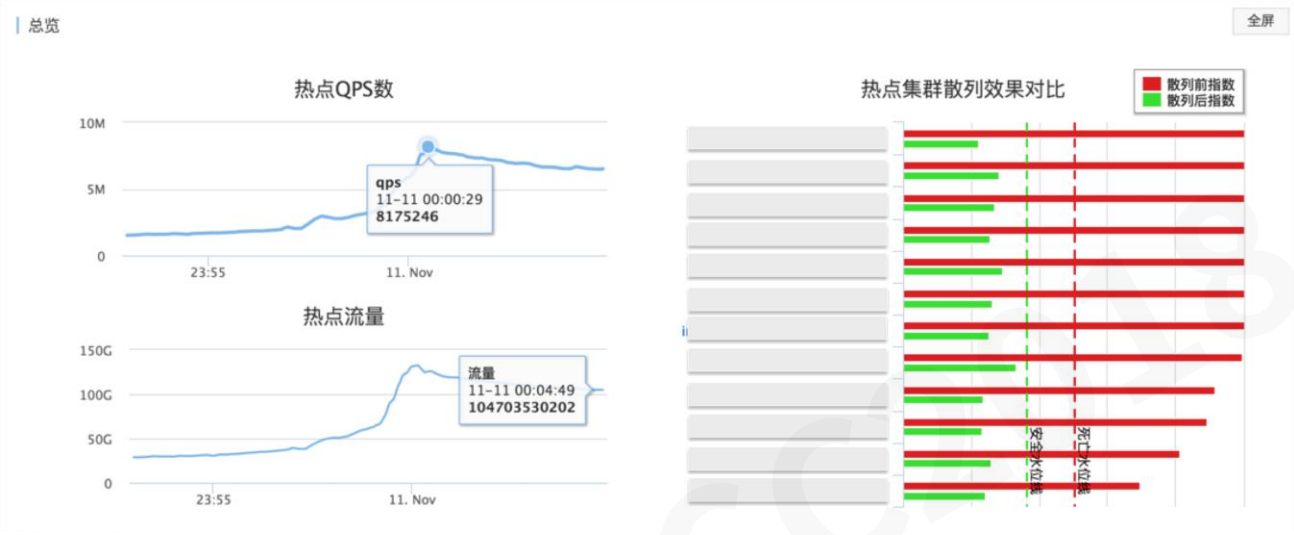
- 通过短暂的过期时间来确保数据一致性
- “数据不存在”时的处理 — 防止回源

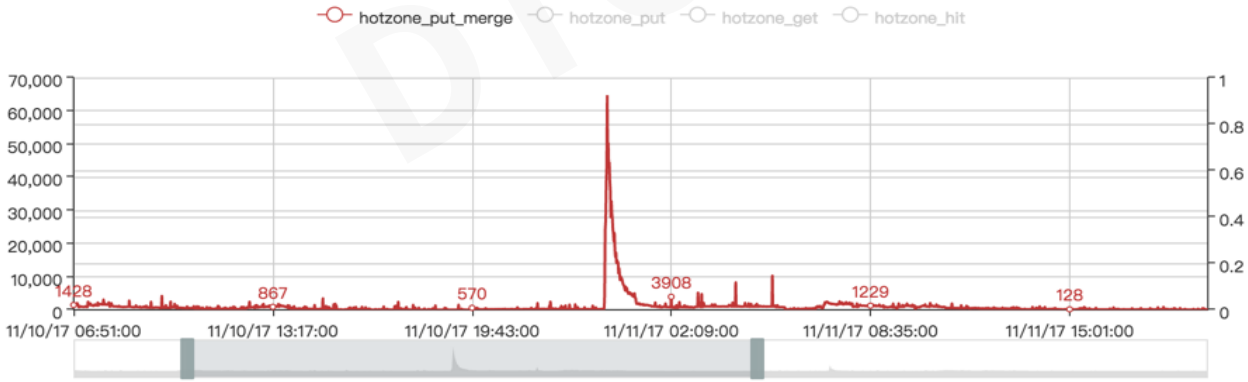
写热点的处理

- 方案的选择
  - 本地合并还是远端合并



2017双11时的热点保护





TAIR发展和愿景

- 从阿里体系走向世界
- 标准化和理论创新
- 贡献开源

- 专注于超大流量的在线访问
- 秒级数亿次的访问
- 提供极低延迟的服务响应

辰仪

在钉钉上扫一扫加我

THANKS


HANA
Hadoop
SQL
MySQL
DB2
BigDate
Spark
SQL Server
NewSQL
Oracle
MPP
NoSQL
OLAP

讲师申请
联系电话（微信号）：18612470168

关注“ITPUB”更多
技术干货等你来拿~

与百度外卖、京东、魅族等先后合作系列分享活动

微学堂

让学习更简单

微学堂是以ChinaUnix、ITPUB所组建的微信群为载体，定期邀请嘉宾对热点话题、技术难题、新产品发布等进行移动端的在线直播活动。

截至目前，累计举办活动期数60+，参与人次40000+。

ITPUB学院

ITPUB学院是盛拓传媒IT168企业事业部（ITPUB）旗下
企业级在线学习咨询平台
历经18年技术社区平台发展
汇聚5000万技术用户
紧随企业一线IT技术需求
打造全方式技术培训与技术咨询服务
提供包括企业应用方案培训咨询（包括企业内训）
个人实战技能培训（包括认证培训）
在内的全方位IT技术培训咨询服务

ITPUB学院讲师均来自于企业
一些工程师、架构师、技术经理和CTO
大会演讲专家1800+
社区版主和博客专家500+

培训特色

无限次免费播放
随时随地在线观看
碎片化时间集中学习
聚焦知识点详细解读
讲师在线答疑
强大的技术人脉圈

八大课程体系

基础架构设计与建设
大数据平台
应用架构设计与开发
系统运维与数据库
传统企业数字化转型
人工智能
区块链
移动开发与SEO

联系我们

联系人：黄老师
电　话：010-59127187
邮　箱：edu@itpub.net
网　址：edu.itpub.net
培训微信号：18500940168