饿了么分布式KV设计架构与实践

赵子明

个人简介

饿了么高级架构师，曾在大麦网、京东工作

曾用5年时间做到互联网架构师

应用架构:单应用->微服务，小型电商->大型电商

基础组件:Zookeeper多活,分布式KV,API网关等

目录

contents

背景

互联网快速发展，传统DB在扩展性和性能方面制约

NoSQL相对于传统DB放弃了join,ACID等特性来追求性能、扩展性、低成本

NoSQL百花齐放，各种表格,kv都在某些场景下用的很好

只要我们需要的

对比

HBase	Dynamo	pika
建在HDFS之上的分布式、面向列的表格系统。	经典的分布式KV，高可用性，高扩展性。	可持久化的大容量redis存储服务
region	hash	hash
中心架构M/S	去中心架构p2p	M/S
强一致性	Quorum NRW,可配置	
基于HDFS延迟略大	本地存储读写性能较高	高性能
部署维护较复杂	部署维护简单	

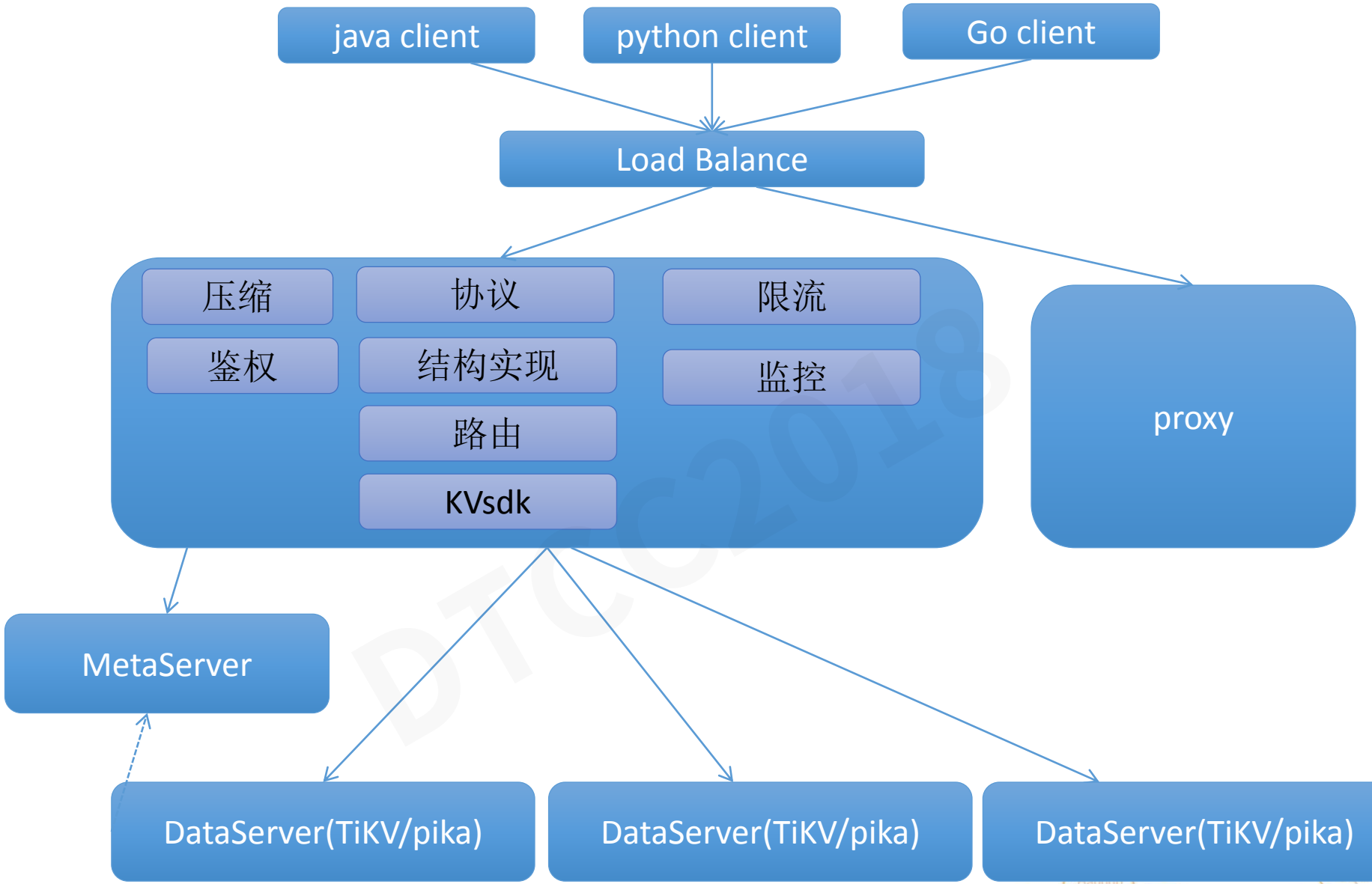
背景

- 饿了么数据量快速增长
- 部分大数据业务使用redis占用大量内存
- kv数据分布在mysql、redis、mongo、cassandra等

扩展性差，性能不达标，运维成本高，数据可靠...

我们对KV的要求

- 高性能（在线场景）
- 大数据量
- 高可靠
- 高可用
- 易运维
- 强一致（数据不丢）



- 多协议支持（目前支持redis）
- 易扩展
- 数据一致&高可用(raft)
- 可定制

使用场景

- ✓ 搜索
- ✓ 推荐
- ✓ 促销
- ✓ 智能调度
- ✓ 商品品类、排序

应用情况

十几套集群,分布在四个机房

几十TB数据

每天百亿级调用

单集群高峰>10w qps

平均响应在1~2ms

大KEY写入读取慢问题

Net I/O

Disk I/O

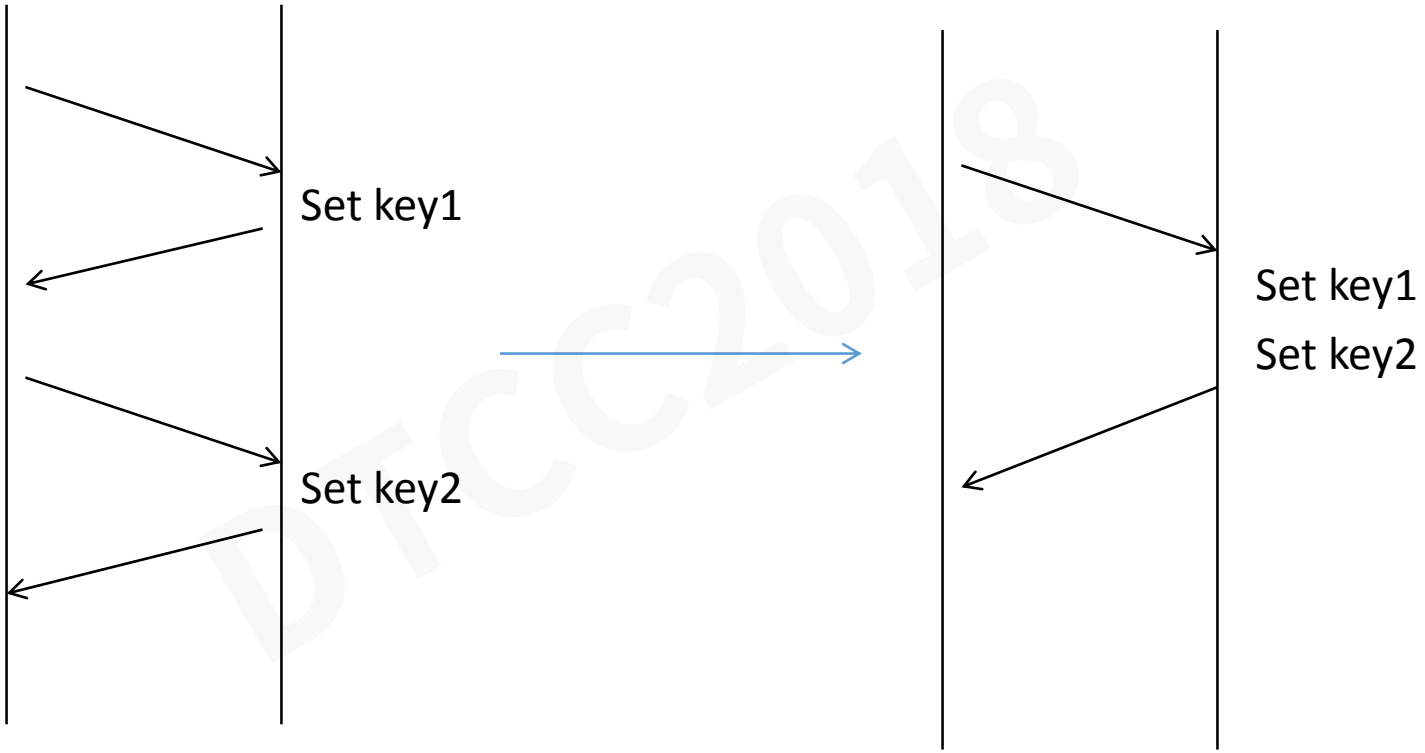
Storage

CPU Time

Compressor	Ratio	Compression	Decompression
memcpy	1.000	13100 MB/s	13100 MB/s
LZ4 default (v1.8.2)	**2.101**	**730 MB/s**	**3900 MB/s**
LZO 2.09	2.108	630 MB/s	800 MB/s
QuickLZ 1.5.0	2.238	530 MB/s	720 MB/s
Snappy 1.1.4	2.091	525 MB/s	1750 MB/s
Zstandard 1.3.4 -1	2.877	470 MB/s	1380 MB/s
LZF v3.6	2.073	380 MB/s	840 MB/s
zlib deflate 1.2.11 -1	2.730	100 MB/s	380 MB/s
LZ4 HC -9 (v1.8.2)	**2.721**	40 MB/s	**3920 MB/s**
zlib deflate 1.2.11 -6	3.099	34 MB/s	410 MB/s

跨机房操作很慢！

磁盘存储的KV有没有必要pipeline？



Hash 多行操作超慢


type(h)
key长度
原始key
field长度
原始field
type(h)
key长度
原始key
field长度
原始field
type(h)
key长度
原始key
field长度
原始field
type(k)
长度
原始key

加了ttl性能损失30%

ttl开关，避免读取meta做ttl验证

展望

多协议支持

全自动化运维

异地多活

总结&思考

分布&均衡

实现成本&运维成本

合适的轮子

Java中间件高级开发工程师

20k-30k/月
饿了么（北京）

1、拥有3年以上服务器后端开发经验；
2、具有中间件、开发框架的开发经验；或者具有存储系统、nosql存储系统开发经验。
3、精通Java,go,C++其中2种语言；另外掌握python，Lua者更佳。
4、精通Web应用开发框架（Spring、MyBatis等）；
5、精通数据库设计和性能优化，熟悉MySQL及数据库编程（SQL、JDBC）；
6、理解io、多线程、集合等基础框架，了解JVM原理，了解软件性能分析、调优等相关方法；
7、有高并发、高可用系统开发经验者优先；
8、精通分布式系统的设计和应用，熟悉分布式、缓存、消息、负载均衡等机制和实现者优先；
9、熟悉各分布式开源项目，如zookeeper，Hbase，HDFS，Lucene，Redis,Cassandra,MongoDB，rabbitmq,dubbo,kafka等基础开源框架优先；
10、掌握或者精通nginx者优先
11、敏捷开发（Agile/Scrum）经验者优先
12、有基础架构开发经验者优先

简历请发:ziming.zhao@eleme

THANKS


HANA
Hadoop
SQL
MySQL
DB2
BigData
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