Alibaba Developer Conference

轻量级分布式key/value存储系统在360的应用

杨康
yangkang@360.cn

2012/07/06

360 www.360.cn

议程

- 背景介绍
- 发展历程
- 单机方案
- 分布式方案
- 多集群方案
- 经验心得

背景介绍 – 应用场合

- 主要应用场合
  - 数据离线流式生产（写）、海量在线检索（读）
    - 木马/恶意网址云查杀
    - 推荐系统

背景介绍 – 系统需求

- 系统需求
  - key/value的数据存储，平均value长度小于1K
  - 低延迟
  - 7*24小时在线
  - 性能要求高
  - 一致性要求
    - 无节点失效时：强一致性
    - 节点失效后：可以暂时停止写操作

发展历程 – KVDB 0. 1

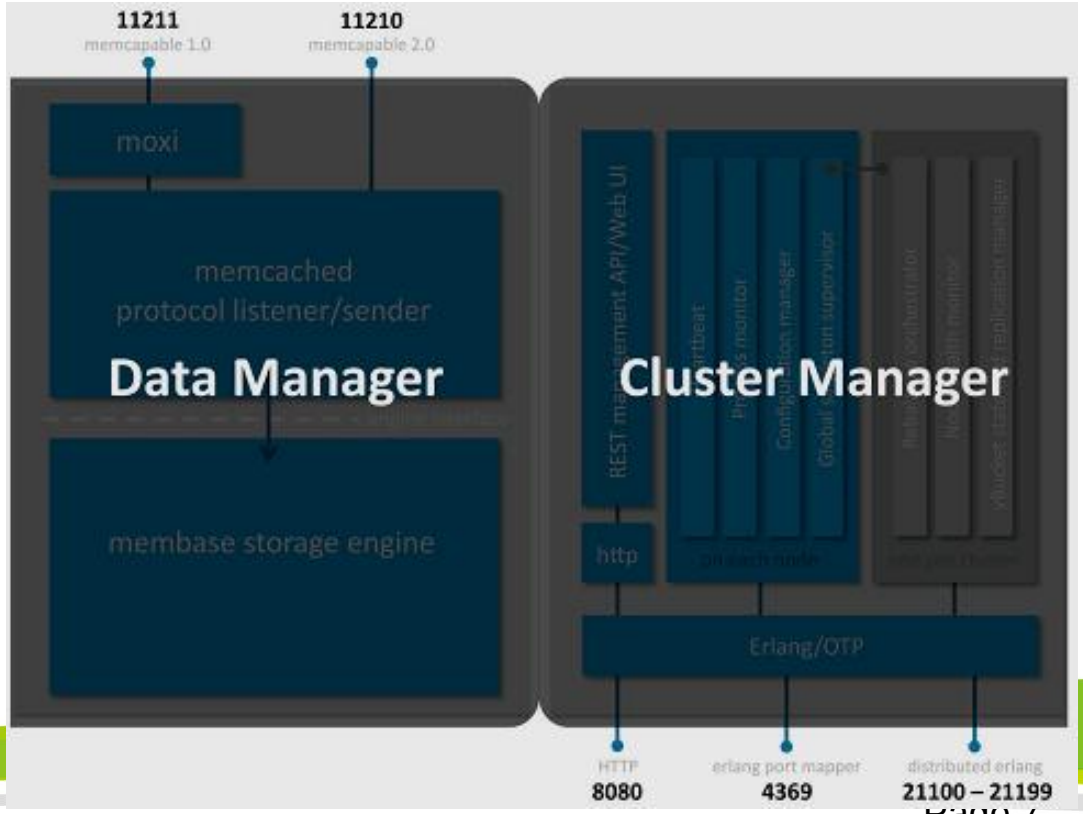
- 09年初 – 10年底
- 10亿数量item

发展历程 – KVDB 0. 1

- 单机内存哈希表存储引擎
  - 在Memcached基础上做二次开发
    - 内存利用率（将LRU特性做成可选：item结构中的time、exptime等成员）
    - 比较多个关联容器实现：std::map、ext/hash_map、stlport::hash_map、google sparse_hash
    - hash_map的rehash问题
    - 协议扩展：dump、load支持
    - 大量只存储有无标识item：使用布隆过滤器支持，写时拷贝

发展历程 – KVDB 1. 0

- 10年12月 – 11年9月
- 超过30亿数量item
- 分布式内存哈希表存储引擎
  - 考虑Membase



发展历程 – KVDB 1. 0

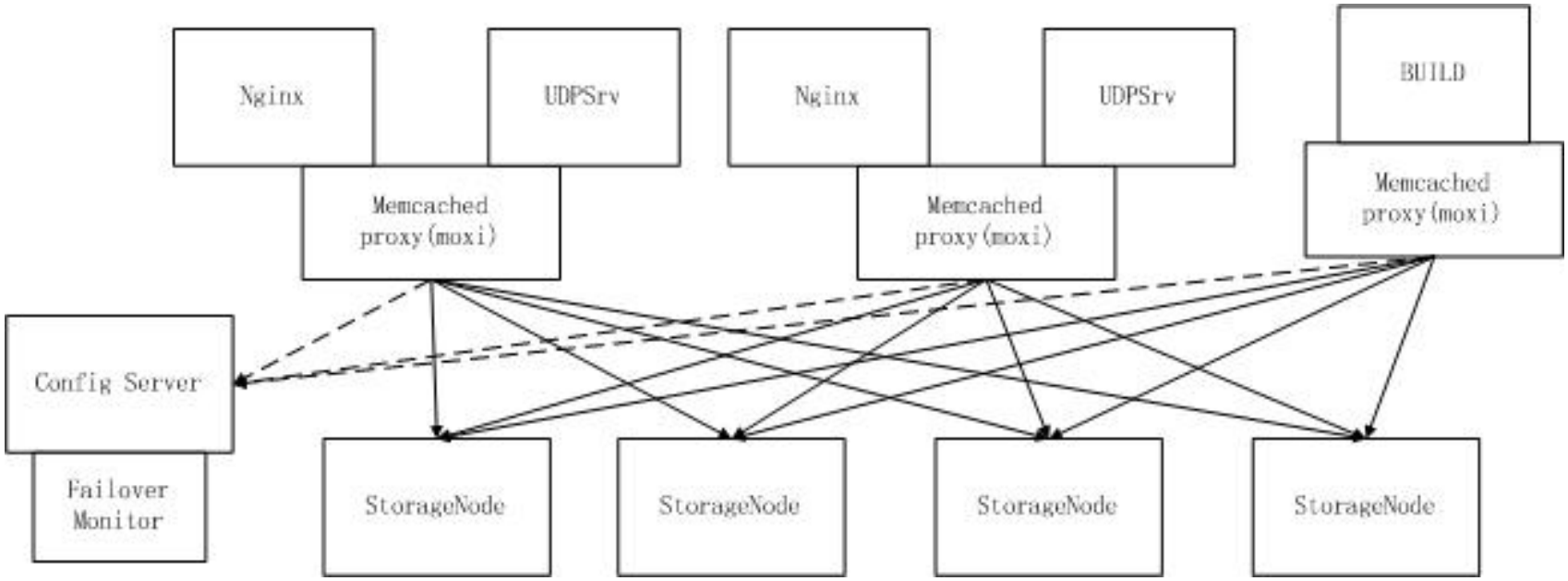
- 分布式内存哈希表存储引擎
  - 考虑Membase
    - 问题：
      - 存储引擎满足不了要求（SQLite）
        - » 改进：存储引擎可定制化
      - Memcached Proxy（Moxi）
        - » 问题1：对下游存储节点的超时处理过于暴力
        - » 问题2：没有短暂屏蔽失效节点的功能，下游存储节点故障会阻塞client

分布式Key/Value存储 – 性能

- 5台机器
- key长度32bytes，value长度为32bytes。随机选取key，测试单机的情况：
  - 单次读：get 45000
  - 批量读（每笔100个key）：gets 400000
  - 写：set 42000
- 访问延迟：1万次查询（每笔100个key），平均延迟0.5ms

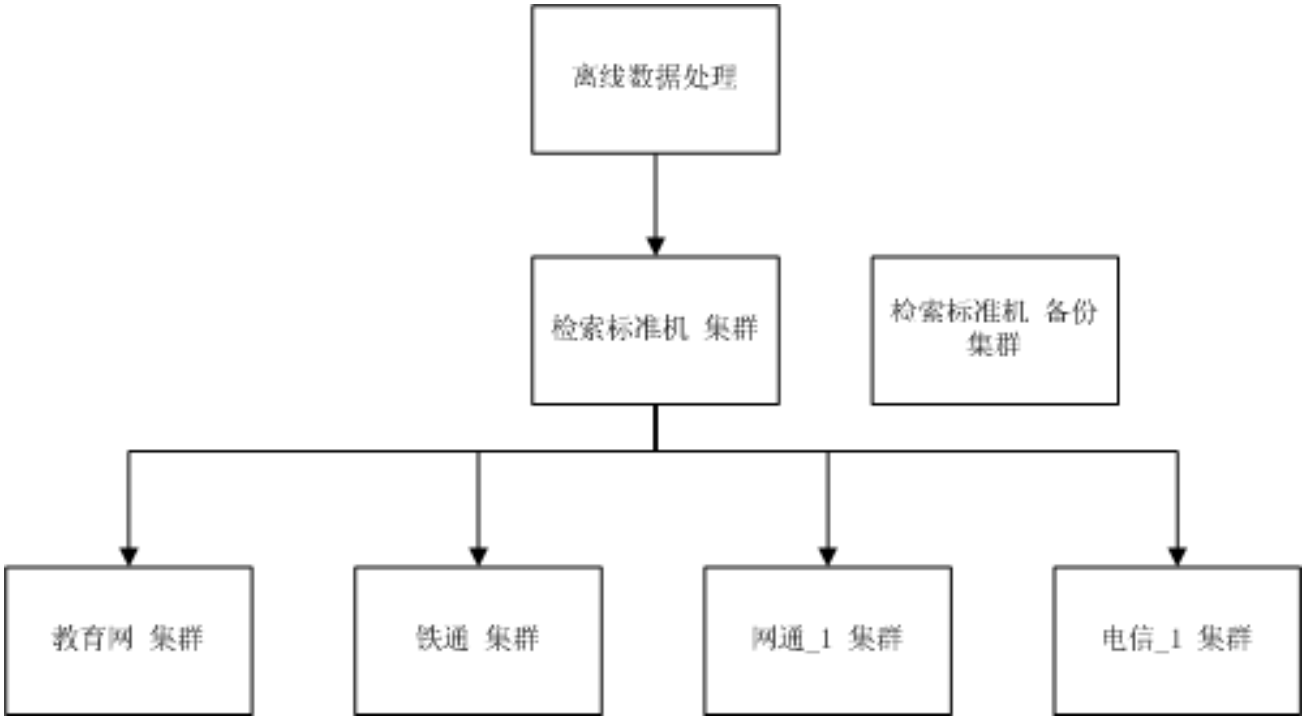
分布式Key/Value存储 – 单集群架构

- 单集群架构



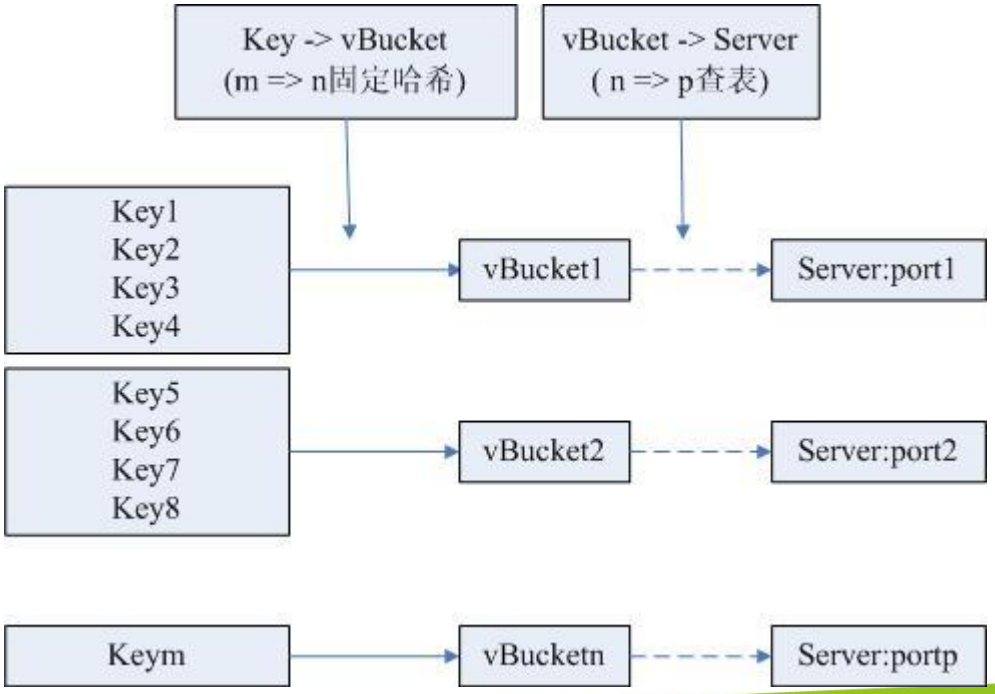
分布式Key/Value存储 – 多IDC部署

- 多IDC部署



分布式机制

- 全局维护一个路由表，中心化管理
- 支持数据的多份冗余
- 路由策略



存储扩容和缩减

- 增加存储节点
  - 新增部分vbucket的备份到新节点上
  - 等待同步结束，将新节点相关vbucket状态设置为active或者replica
  - 修改全局路由表这部分vbucket和server的对应关系
  - 释放旧节点上的相关vbucket占用的空间

备份和恢复

- 冗余热备
- 在线的故障转移
  - 全局Heartbeat程序
  - 监控集群内的节点是否存活，在某个节点失效后修改全局路由表，将尚存活的节点上对应vbucket状态更新为active
  - 集群内秒级切换
  - 后续数据Rebalance工作由运维工程师来手动完成若集群无可用active vbucket，则通知DNS下线整个IDC
- 冷备

发展历程 – KVDB 1. 1

- 11年10月 – 至今
- 100亿数量item
- 分布式持久化存储引擎
  - LevelDB on SSD

	INTEL SSDSA2CW300G3	INTEL SSDSA2SH064G1GC	SAS 300G(xen9 /data3)	SAS 300G(xen9 /data3)
fillseq	470302 op/s	457898 op/s	415552 op/s	438916 op/s
fillrandom	28327 op/s	30529 op/s	17308 op/s	17942 op/s
overwrite	24117 op/s	24923 op/s	9109 op/s	9014 op/s
readrandom	2471 op/s	2427 op/s	209 op/s	227 op/s
readseq	1414632 op/s	1403199 op/s	733732 op/s	698080 op/s
readreverse	131285 op/s	129349 op/s	156739 op/s	160487 op/s

发展历程 – KVDB 1. 1

- LevelDB持久化存储引擎
  - 多磁盘支持
  - 单块磁盘多实例引起的随机IO
  - 备份
    - 单机支持交错盘备份

LevelDB On SSD 问题若干

- compaction线程数
- 新版本LevelDB对布隆过滤器的支持
  - 不需要维护全局的布隆过滤器
- Intel 320 over provision技术
  - 牺牲20%的存储空间换取性能提升

LevelDB On SSD 问题若干 cont.

- 缺少的特性
  - 不支持快速做key的遍历
  - 不支持返回item总数
- 对大数据的支持不好
  - 读写都可以触发compaction
  - 多次读取
  - Bitcask存储引擎

正在做的改进 – KVDB X

- Config Server使用Zookeeper实现
  - 维护配置信息以及节点状态
- 自动化Rebalance
- 混合多级存储引擎（RAM/SSD/SAS）
- 友好的管理界面

谢 谢！