HBase在阿里的应用与优化

邓明鉴

2012.7.1

Agenda

- HBase介绍
- HBase在阿里的发展
- 遇到的问题及优化
- 未来的工作
- Q & A

HBase介绍

- HBase是什么
  - HBase is the Hadoop database. Think of it as a distributed, scalable, big data store.

HBase介绍

- HBase的特点
  - 支持海量数据
  - 拥有良好扩展性
  - 高性能读写
  - 快速分析
  - 满足强一致性要求
  - schema灵活多变
  - 列存储
  - 良好易用的JAVA接口

HBase在阿里的发展

- 2011.3月开始研究
- 2011.5月上线第一个应用
- 截止2012.1，线上部署:
  - 2个机房
  - 150台服务器
  - 应用约12个（核心应用2个）
  - 总tps约100k /s
  - 总数据量约60TB

HBase在阿里的发展

- 截止2012.1
  - 线上只部署有0.90.2RC3的版本

HBase在阿里的发展

- 接下来...
  - 应用持续增加
  - 数据量成倍增长
  - tps成倍增长
  - 应用类型及要求更加复杂
  - 应用方对服务稳定性要求增加
  - 应用方对响应时间要求增加
  - 应用方对服务透明性要求增加

遇到的问题及优化

- 数据量增大
  - 对写无影响
  - 影响compact
  - 影响gc
  - 影响读性能
  - 网络带宽
  - region数量上升？ → HFileV2

遇到的问题及优化

- 稳定性要求提高
  - 随意宕机
  - multi assign
  - 慢响应
  - hang

遇到的问题及优化

- 内存增大，gc问题突出
  - YGC占用时间长
    - Eden大小控制在2GB以内
  - FULL GC会导致节点crash
    - 使用CMS
    - mslab谨慎使用
  - 频繁cms导致load升高
    - 合理的CMSInitiatingOccupancyFraction
    - 合理的SurvivorRatio

遇到的问题及优化

- 毛刺与慢请求
  - compact算法
  - 多线程compact
  - 读写分离
  - java.lang.Class.getMethod

遇到的问题及优化

- HDFS实时性问题
  - blockreport → cdh3u3解决
  - datanode 参数
    - dfs.socket.timeout
    - dfs.datanode.socket.write.timeout
    - dfs.datanode.failed.volumes.tolerated
    - dfs.client.cached.conn.retry
    - dfs.datanode.max.xcievers
  - 调度算法，如 FSNamesystem.commitBlockSynchronization

遇到的问题及优化

- 性能优化
  - 0.94为什么可以极大提高性能？
    - group sync
    - 及时清理cache
    - 改进的compact算法
    - lazy-seek
    - HLog Compress
    - 前缀压缩

遇到的问题及优化

- 0.94版本的优势与劣势
- 优势
  - 性能明显提升
  - 测试框架加快
  - 节省占用空间
- 劣势
  - 不稳定，BUG较多

遇到的问题及优化

- 宕机恢复时间要求短
  - 之前的状况：100k region的集群大约需要15-30分钟才能完全恢复读写

遇到的问题及优化

- 修改脚本，crash即时感知
- 跳过不必要的hlog
- 扫描meta带上cache
- 批量操作，减少rpc
- 优化zk操作中锁竞争
- tcpnodelay
- bulkassign
- zk串行改并行
- hdfs创建元数据操作并行

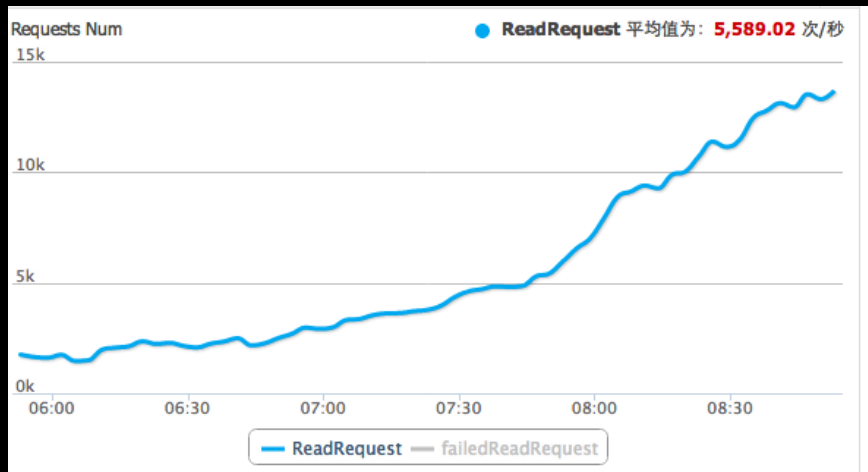
遇到的问题及优化

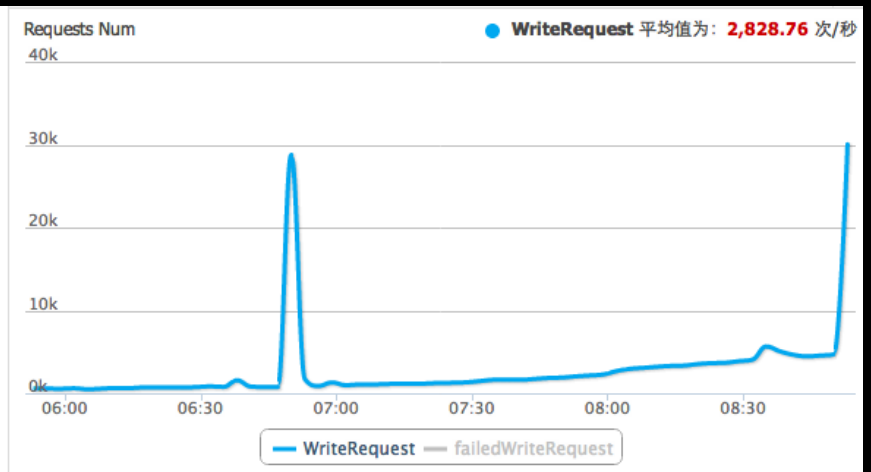
- 宕机恢复时间缩短
  - 现状(单台server crash):15-30分钟
  - 优化为：小于1分钟
- ddl恢复时间缩短
  - 现状(集群重启)：数小时
  - 优化为：数分钟

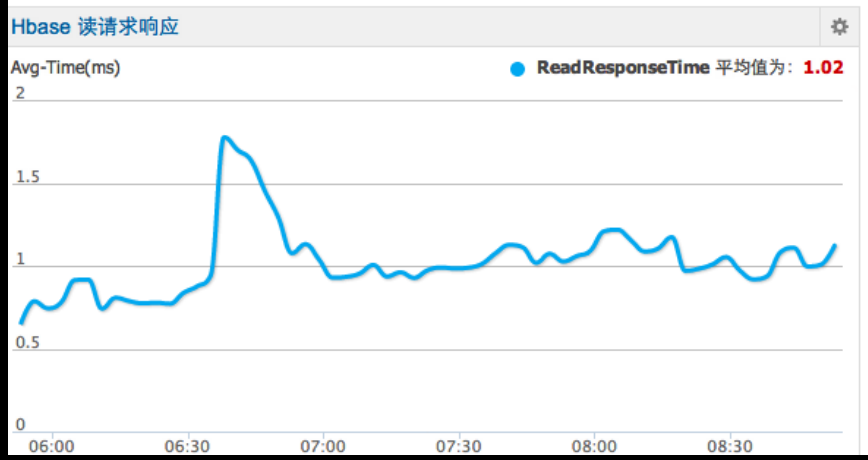
遇到的问题及优化

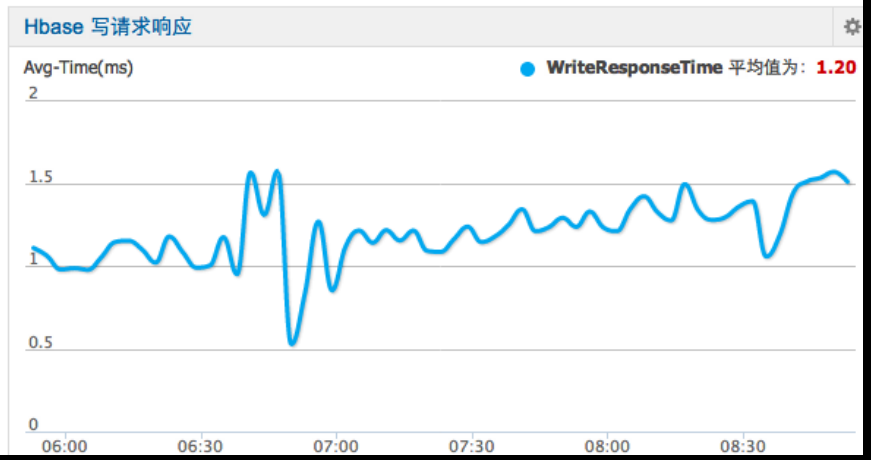
- 更加完善的监控

遇到的问题及优化

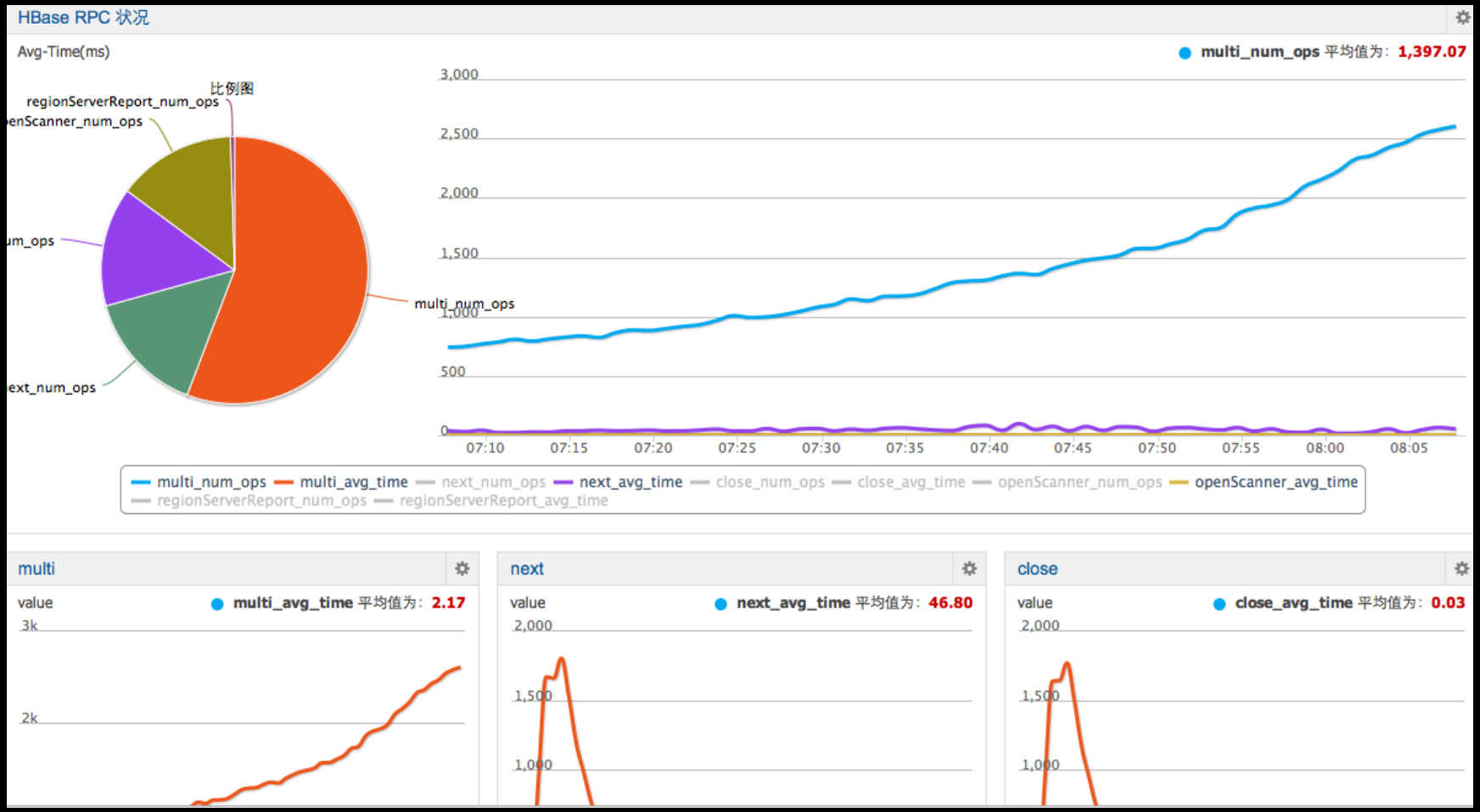

Requests Num
ReadRequest 平均值为：5,589.02 次/秒
15k
10k
5k
0k
06:00
06:30
07:00
07:30
08:00
08:30
ReadRequest
failedReadRequest


Requests Num
WriteRequest 平均值为：2,828.76 次/秒
40k
30k
20k
10k
0k
06:00
06:30
07:00
07:30
08:00
08:30
WriteRequest
failedWriteRequest


Hbase 读请求响应
Avg-Time(ms)
ReadResponseTime 平均值为：1.02
2
1.5
1
0.5
0
06:00
06:30
07:00
07:30
08:00
08:30


Hbase 写请求响应
Avg-Time(ms)
WriteResponseTime 平均值为：1.20
2
1.5
1
0.5
0
06:00
06:30
07:00
07:30
08:00
08:30

遇到的问题及优化


HBase RPC 状况
Avg-Time(ms)
multi_num_ops 平均值为: 1,397.07
比例图
regionServerReport_num_ops
multi_num_ops
3,000
2,500
2,000
1,500
1,000
500
0
07:10
07:15
07:20
07:25
07:30
07:35
07:40
07:45
07:50
07:55
08:00
08:05
multi_num_ops multi_avg_time next_num_ops next_avg_time close_num_ops close_avg_time openScanner_num_ops openScanner_avg_time
regionServerReport_num_ops regionServerReport_avg_time
multi
value
multi_avg_time 平均值为: 2.17
3k
2k
next
value
next_avg_time 平均值为: 46.80
2,000
1,500
1,000
close
value
close_avg_time 平均值为: 0.03
2,000
1,500
1,000

遇到的问题及优化

- 经过以上工作，目前的系统：
  - 覆盖4个机房
  - 服务器扩展到500+
  - 总数据量约400TB
  - 总tps超过300k /s
  - 应用增加到30多个
  - 核心应用达到6个
  - 0.90/0.92/0.94版本都有线上应用

遇到的问题及优化

- 目前线上系统
  - 2月起每月会release 1-2个版本
  - 0.90系列：8个版本，灰度发布 (服务器占比87%)
  - 0.92系列：1个版本(服务器占比10%)
  - 0.94系列：1个版本(服务器占比3%)

遇到的问题及优化

- 与社区的互动
  - 直接提交patch: 约30
  - 间接提交patch: 接近20
  - 占比：0.90.4以来的patch占比2.5%左右
  - critical以上级别: 8
  - apache id: chunhui shen/xing shi/binlijin
- 对HBase的在线应用更加有信心

未来的工作

- 二级索引
- snapshot及replication
- 引入独立cache策略，适用于不同场合
- 实时化HDFS版本
- NameNode HA
- 全面运维自动化
- Hive in HBase优化
- 资源隔离
- 安全与权限
- 更深入的性能优化

Q & A