NameNode性能诊断及优化

阿里集团核心系统研发部-长仁

目录

- 专用计算组介绍
- 阿里HDFS的介绍及NameNode性能挑战
- NameNode性能诊断及优化
  - RPC框架
  - DFS实现
- 随后工作
- 问答

阿里核心系统研发部专用计算组

- 针对特定领域问题及不同硬件架构，以计算性能、效能为导向的优化。
- Java虚拟机定制、优化，基于OpenJDK VM，定制打造Taobao JVM并开源。
- 长期、持续为阿里Hadoop开发、运维团队提供服务，持续打酱油。
- 关于我们的更多信息请访问 jvm.taobao.org

那些年我们打过的Hadoop酱油

- GC invisible heap（GCIH）用于Hadoop，在Worker进程间直接共享Java对象。
- 实现利用SSE4.2 crc32指令的JVM intrinsic加速校验。
- 实现利用Packed Comparison SIMD指令的JVM intrinsic加速Map sort过程并优化prefixKey。
- 在Hadoop集群推广TaobaoJVM替换OpenJDK VM。
- 压缩算法优化及基于FPGA，MIPS的压缩卡在Hadoop上应用预研。
- Hadoop集群上JVM相关疑难杂症解决：JVM crash，GC异常等。
- 更多信息访问 jvm.taobao.org。

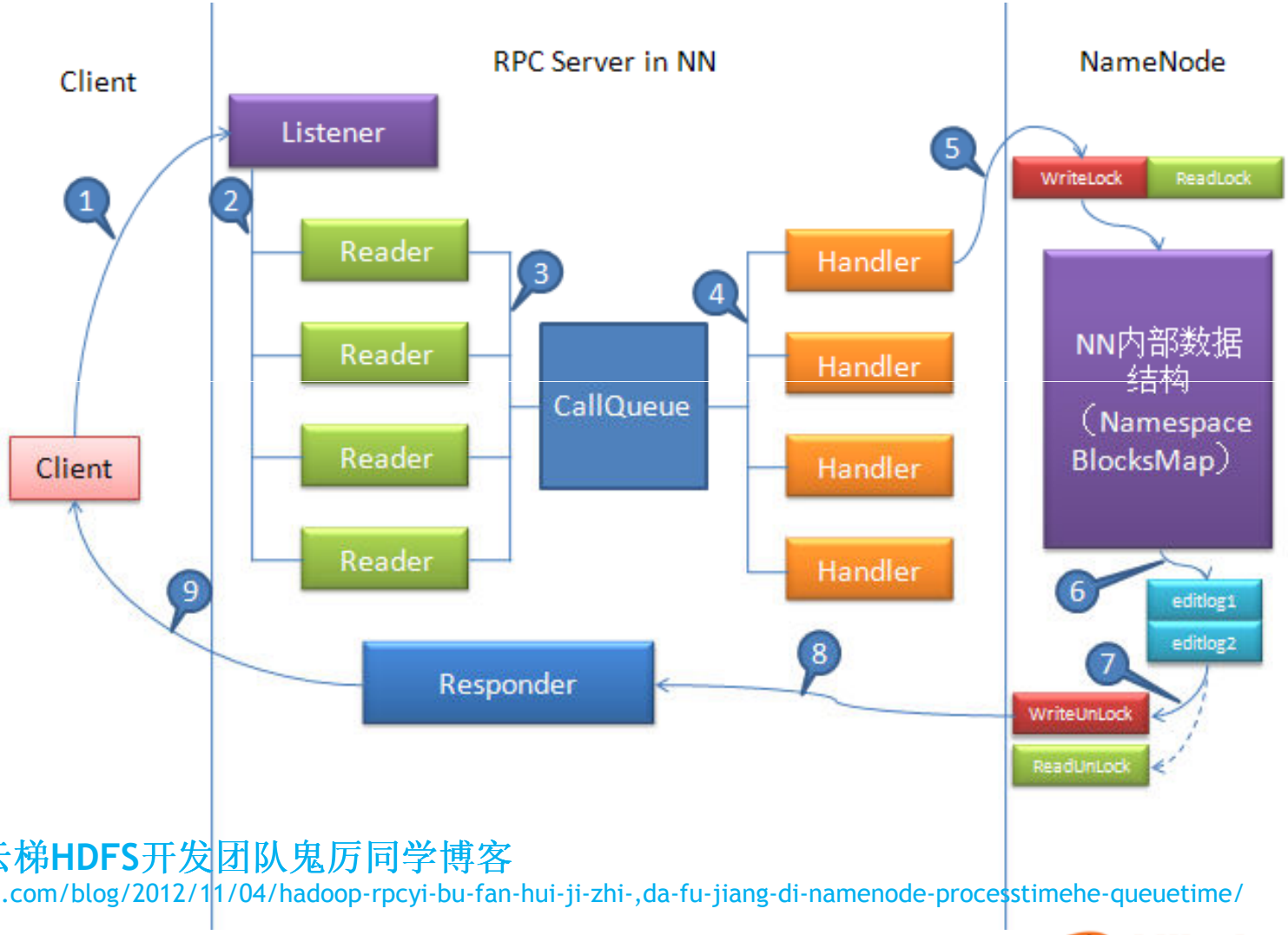
阿里hadoop集群HDFS现状

- 内部名称“云梯”。
- 规模约3200台。
- 世界最大的单namespace集群。
- NameNode配置
  - 双路Xeon E5520
  - 196G内存
- 目前NameNode性能指标：
  - CPU利用率30%
  - FileOps+Heartbeat 1.5万每秒
  - 平均RpcProcessingTime 3ms
  - 平均RpcQueueTime 60ms
- 云梯开发、运维团队持续应对性能挑战：监控，优化

HDFS性能挑战

- 开发、运维团队反映的问题：
  - 机器增加，queue time提高，影响集群性能。
  - CPU利用率不高，感觉有潜力可挖。
- 可疑瓶颈点：
  - RPC框架部分?
  - DFS实现部分？
  - IO部分随后优化，能提升多少，心里有数。
- 我们的目标：快速定位解决问题，锁重点怀疑，搂草打兔子，顺便看看有没有CPU热点。

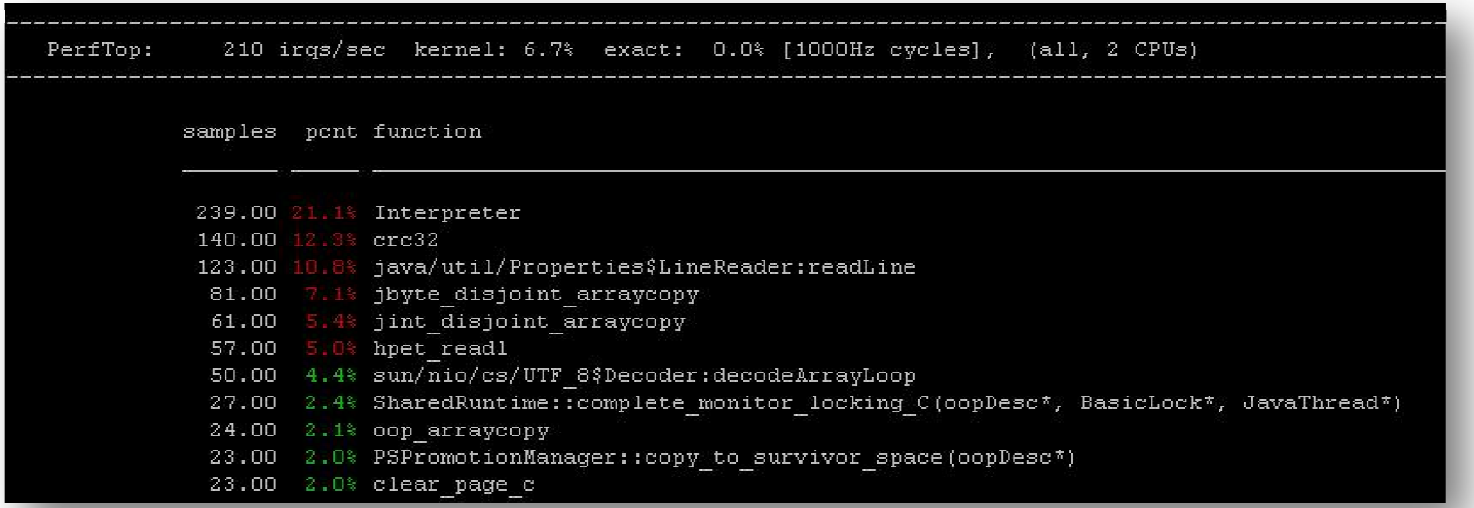
NameNode RPC框架



图片来自云梯HDFS开发团队鬼厉同学博客
http://luoli523.com/blog/2012/11/04/hadoop-rpcyi-bu-fan-hui-ji-zhi-,da-fu-jiang-di-namenode-processtimehe-queuetime/

RPC框架诊断

- 工具
  - 写个客户端造压力
    - 用尽量少的资源
    - 一定轻，要能产生多connection
  - jstack+脚本
  - Java能用的利用x86 PMU的profiling工具？Oprofile？Intel Vtune？用我们修改的perf 诊断NameNode

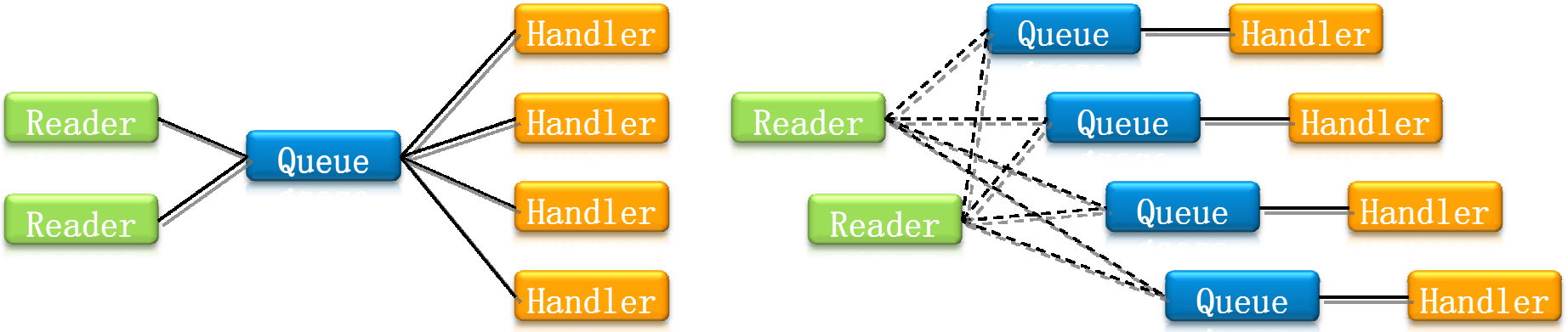


RPC诊断结果

- Queue实现：LinkedBlockingQueue
  - CPU Cache miss很高，多核，一个Queue，出入Queue的成本很大。
  - 性能考虑上可提高的点很多，这是用Java实现的。
- 明显CPU热点：UTF8/Unicode转码
- Log4j，性能大累赘

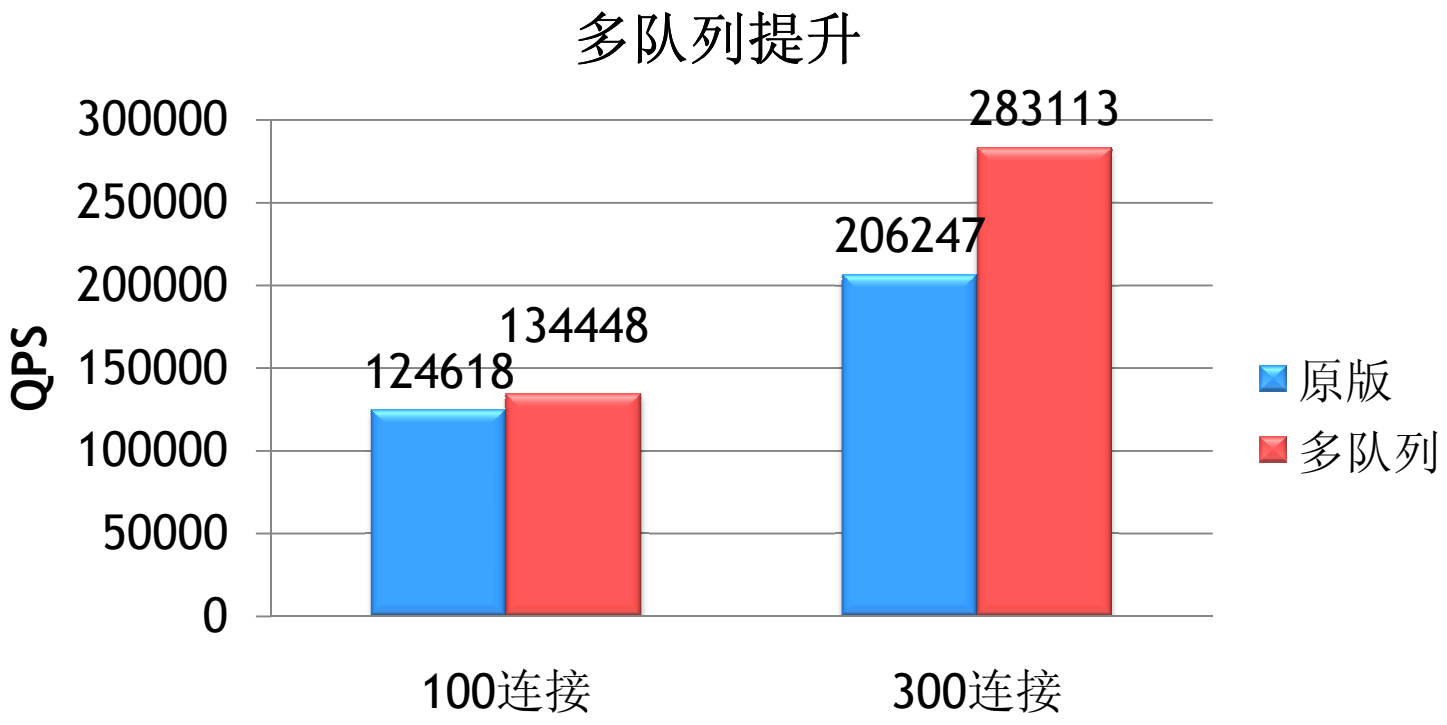
RPC多队列优化

- 多队列，一个Handler一个Queue，按callid分配，3行代码修改，最简单且有效。



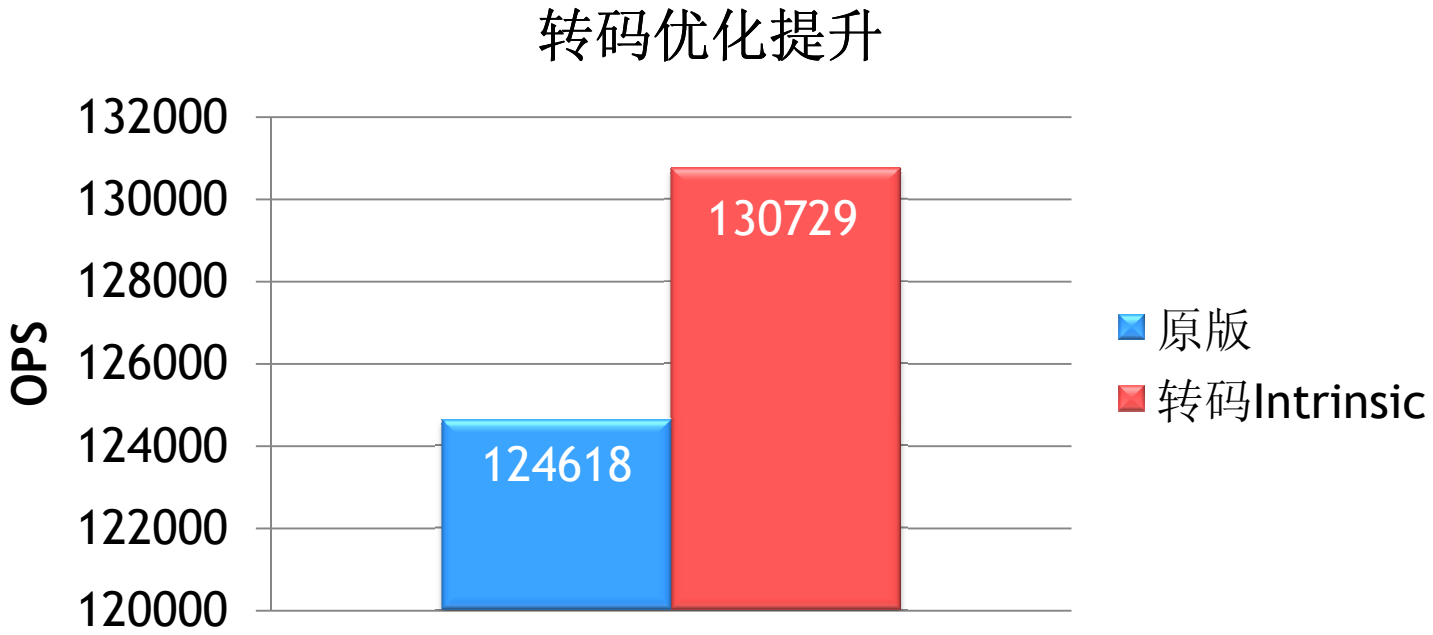
多队列优化效果

- 5 readers， 150 handlers，关闭log4j，调用空操作。
  - 1 client，100 connections， OPS提升7.9%；
  - 3 clients，300 connections ， OPS提升37%



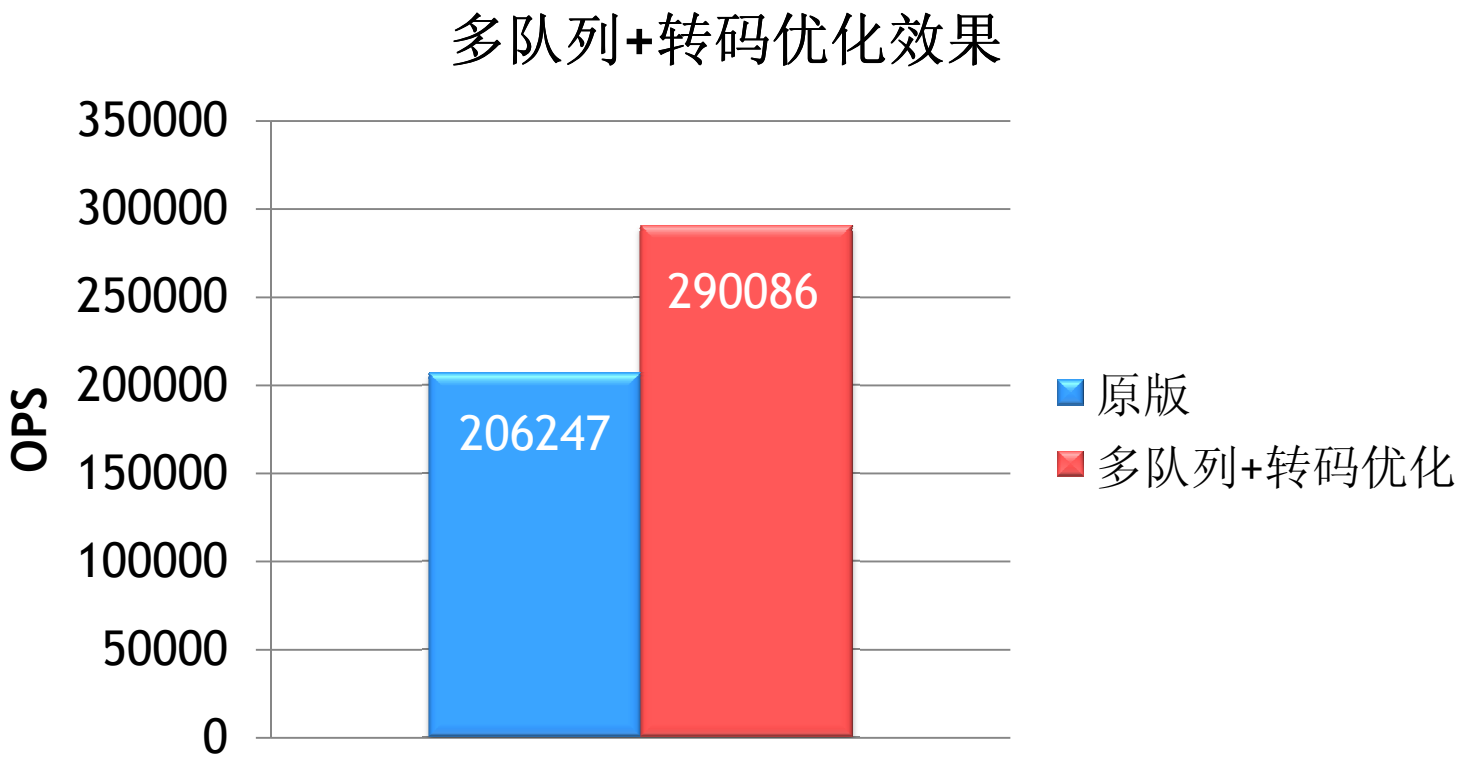
UTF8/Unicode转码优化及效果

- 用SIMD指令的JVM Intrinsic转码实现。
- 32 bytes比目前Hadoop内pure Java实现快4倍，64 bytes快十几倍。
- 5 readers， 150 handlers，1 client， 100 connections，关闭log4j，调用空操作。OPS提升4.9%



多队列+转码优化效果

- 3 clients，300 connections ， OPS提升40%



DFS诊断

- 工具
  - 自带Benchmark：NNThroughputBenchmarkMixed
  - 线上看看DFS部分锁竞争情况，jstack+脚本
- 结果
  - 并发惨不忍睹
  - 写锁里操作不够简练

DFS优化，简化写锁内操作

- 写锁内的log4j提出。
  - 效果：throughputmix测试提升11%。
- 写锁内editlog的logXXX动作提出。
  - Editlog的log操作放到写锁外记录顺序会不一致！写锁内生成serialNumber，按顺序输出。
  - 效果： throughputmix测试提升9.8%。

DFS优化，更新锁

- 读、写锁变为读、写、更新锁。由于目前JDK没有更新锁，Concurrency JSR-166 Interest邮件列表讨论，我们先实现了一个。
  - 效果：偏向读，throughputmix测试读提升近100倍，代价是写会变慢，最坏慢1倍。实际效果取决于读写比例。

随后工作

- 性能考量，NN内可优化点很多。
- 对于NN，如果云梯持续成为世界最大单Namespace集群，我们最关心offheap改造，应用GCIH，或者其他优化方法。

谢谢

- 问答时间