HBase异构存储测试报告

1. 测试总结
2. HBase采用nvme+HDD硬盘的异构存储方式，读写性能与全SSD的服务器性能对比差异相对较小，但因测试环境及时间限制，建议后续需要补充以下复杂场景测试：

大数据量铺底（集群总存储使用量达到60%）场景读写性能测试

线上环境长时间疲劳测试（两周以上）

单条大记录测试（50K、100K、1M等）读写并发测试

网络流量临界条件下（网络带宽打满）的性能测试

使用NVME做L2缓存的读测试

1. 采用2nvme+4HDD、2nvme+8HDD的架构分别测试，读写性能初步没有明显变化。
2. HBase采用wal写nvme的策略，读写性能明显好于全HDD的存储方式，与全nvme的方式性能相当。
3. 在使用HDD硬盘的情况下，当发生major\_compact时，compact线程数量过高（8线程以上）对性能影响较大，建议最佳配置整理线程数为2+2（长整理+短整理）线程数
4. 单节点4\*HDD与单节点8\*HDD在整理线程数为2+2的情况下，读写性能差别不大，但是观察到在4\*HDD配置下，major\_compact的性能会有明显下降，考虑到major\_compact会对后续性能有所影响，故不建议减少HDD磁盘配置，另外需要补充在全表major compact场景下与全SSD服务器配置的性能对比
5. 将异构集群加入线上环境执行大业务量写入测试，可以看到性能能维持稳定，平均5W的请求量，依然能够保持tp99指标。

目前测试还存在的问题如下：

测试总时长仍然较短（少于四天），若可能希望能延长测试时间

需要更大批量的写入（不同大小的数据记录）以验证请求峰值的性能

增加线上读性能、读写混合场景测试

1. 测试目的
   1. 目的：

本测试要验证HBase采用硬盘异构存储方式的可行性，主要考察性能变化情况，在满足性能的条件下，寻找降低成本的服务器配置方案，初步验证点如下：

第一，验证HBase采用wal写nvme的异构存储模式架构、全nvme配置架构、全HDD配置架构的性能对比

第二，验证HBase在不同HDD配置数量情况下的性能变化测试详情

第三，验证major compact在不同配置下对HBase的读写性能的影响

第四，验证不同磁盘通道配置下（4\*HDD与8\*HDD），在major\_compact下是否由较大性能变化

第五，验证线上环境下使用异构存储的性能情况

* 1. 测试1：基本读写性能对比

测试集群配置：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 集群节点数量 | 硬件配置 | 操作系统 | jdk | 节点数量 | Hadoop | 测试工具 | 发压机硬件配置 |
| Wal写nvme模式 | CPU：32C  内存：256G  硬盘：2\*4T nvem+8\*10THDD | CENTOS7.2 | Jdk1.7 | 6 | Hadoop2.7.1  Hbase1.1.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |
| 全nvme模式 | 同上 | 同上 | Jdk1.7 | 6 | Hadoop2.7.1  Hbase1.1.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |
| 全HDD模式 | 同上 | 同上 | Jdk1.7 | 6 | Hadoop2.7.1  Hbase1.1.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |
| 全SSD模式 | CPU：32C  内存：256G  硬盘：12\*1T SSD | 同上 | Jdk1.7 | 6 | Hadoop2.7.1  Hbase1.1.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |

1. DN需要配置nvme数量不少于2块、SSD数量为12块
2. Hadoop、HBase的JVM相关配置使用京东256G服务器最新线上环境配置，关闭bucketcache以充分测试读写性能
   1. 测试准备

测试表：预切分300个region,每个RS各50region；

测试铺底10亿条数据；

Standard YCSB测试结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模式 | Read（随机读1亿数据） | | Insert(写入1亿条数据) | | Read/update各50%混合（读写2亿条数据） | |
| TPS(k/s) | Tp99 | TPS(k/s) | Tp99 | TPS(k/s) | Tp99 |
| Wal写nvme模式 | 93 | 502 | 433 | 92 | 129 | 读：1000  写：66 |
| 全nvme模式 | 92 | 841 | 436 | 45 | 128 | 读:1012  写：65 |
| 全HDD模式 | 91 | 728 | 374 | 218 | 124 | 读：722  写：37 |
| 全SSD模式 | 87 | 697 | 416 | 20 | 131 | 读：898  写：40 |

测试结论：

1. Wal写nvme模式对比全nvme模式性能没有明显差别
2. 全HDD模式对比前两种模式，写性能有15%左右下降，读性能无明显变化，读写混合情况下没有明显性能下降
   1. 测试2：不同HDD配置读写性能对比

测试集群配置：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 集群节点数量 | 硬件配置 | 操作系统 | jdk | 节点数量 | Hadoop | 测试工具 | 发压机硬件配置 |
| Wal写nvme模式（8HDD） | CPU：32C  内存：256G  硬盘：2\*4T nvem+8\*10THDD | CENTOS7.2 | Jdk1.7 | 6 | Hadoop2.7.1  Hbase1.1.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |
| Wal写nvme模式（4HDD） | CPU：32C  内存：256G  硬盘：2\*4T nvem+4\*10THDD | 同上 | Jdk1.7 | 6 | Hadoop2.7.1  Hbase1.1.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |

* 1. 测试准备1

测试表：预切分300个region,每个RS各100region；

测试铺底10亿条数据；

Standard YCSB测试结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模式 | Read（随机读1亿数据） | | Insert(写入1亿条数据) | | Read/update各50%混合（读写2亿条数据） | |
| TPS(k/s) | Tp99 | TPS(k/s) | Tp99 | TPS(k/s) | Tp99 |
| Wal写nvme模式（8HDD） | 93 | 728 | 433 | 218 | 129 | 读：722  写：37 |
| Wal写nvme模式（4HDD） | 92 | 541 | 428 | 94 | 127 | 读：592  写：34 |

测试结论：

1. 4\*10T HDD对比8\*10T HDD，读、写、读写混合性能均相当
   1. 测试3：8通道HDD模式下，配置不同compact线程数对读写性能的影响

测试集群配置：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 集群节点数量 | 硬件配置 | 操作系统 | jdk | 节点数量 | Hadoop | 测试工具 | 发压机硬件配置 |
| Wal写nvme模式 | CPU：32C  内存：256G  硬盘：2\*4T nvem+8\*10THDD | CENTOS7.2 | Jdk1.7 | 6 | Hadoop2.7.1  Hbase1.1.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |

1. DN需要配置nvme数量不少于2块
2. 测试过程中HBase的长短整理线程数量各自调整为2、4、5
   1. 测试准备

测试表：预切分300个region,每个RS各50region；

测试铺底10亿条数据；

Standard YCSB测试结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模式 | Read（随机读1亿数据） | | Insert(写入1亿条数据) | | Read/update各50%混合（读写2亿条数据） | |
| TPS(k/s) | Tp99 | TPS(k/s) | Tp99 | TPS(k/s) | Tp99 |
| 长整理线程：2  短整理线程：2  执行major\_compact | 94 | 1199 | 421 | 510 | 130 | 读：763  写：33 |
| 长整理线程：4  短整理线程：4  执行major\_compact | 79.5 | 910 | 331 | 214 | 127 | 读:780  写：21 |
| 长整理线程：5  短整理线程：5  执行major\_compact | 80 | 373 | 202 | 400 | 117 | 读：966  写：2 |
| 8\*HDD  不执行major\_compact | 93 | 502 | 433 | 92 | 129 | 读：1000  写：66 |

测试结论：

1. 执行major\_compact的情况下，HDD硬盘线程数越多，对读写性能影响越大，整理线程总数10的情况下，写吞吐量下降达到60%，读吞吐量下降13%，读写综合性能下降10%；整理线程总数8的情况下，写吞吐量下降达到23%，读吞吐量下降13%，读写综合性能受影响较小；在整理线程总数4的情况下，读写性能基本未受到影响，故使用HDD的情况下，设置总整理线程数4为比较合理的值
   1. 测试4：8通道HDD和4通道HDD在同样配置的compact条件下，性能变化对比

：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 集群节点数量 | 硬件配置 | 操作系统 | jdk | 节点数量 | Hadoop | 测试工具 | 发压机硬件配置 |
| Wal写nvme模式 | CPU：32C  内存：256G  硬盘：2\*4T nvem+8\*10THDD | CENTOS7.2 | Jdk1.7 | 6 | Hadoop2.7.1  Hbase1.1.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |
| Wal写nvme模式 | CPU：32C  内存：256G  硬盘：2\*4T nvem+4\*10THDD | CENTOS7.2 | Jdk1.7 | 6 | Hadoop2.7.1  Hbase1.1.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |

1. DN需要配置nvme数量不少于2块
2. 测试过程中HBase的长短整理线程数量各自调整为2、4
   1. 测试准备

测试表：预切分300个region,每个RS各50region；

测试铺底10亿条数据；

测试开始之前执行多次major\_compact命令确保读写过程中始终在进行major\_compact

Standard YCSB测试结果：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 模式 | Read（随机读1亿数据） | | Insert(写入1亿条数据) | | Read/update各50%混合（读写2亿条数据） | |
| TPS(k/s) | Tp99 | TPS(k/s) | Tp99 | TPS(k/s) | Tp99 |
| 8\*HDD  长整理线程：2  短整理线程：2  执行major\_compact | 94 | 1199 | 421 | 510 | 130 | 读：763  写：33 |
| 4\*HDD  长整理线程：2  短整理线程：2  执行major\_compact | 91 | 841 | 436 | 45 | 128 | 读:1012  写：65 |
|  |  |  |  |  |  |  |

测试结论：

1.在设置整理线程总数为4的情况下，4\*HDD和8\*HDD的性能变化不大

* 1. 测试5：线上测试

：

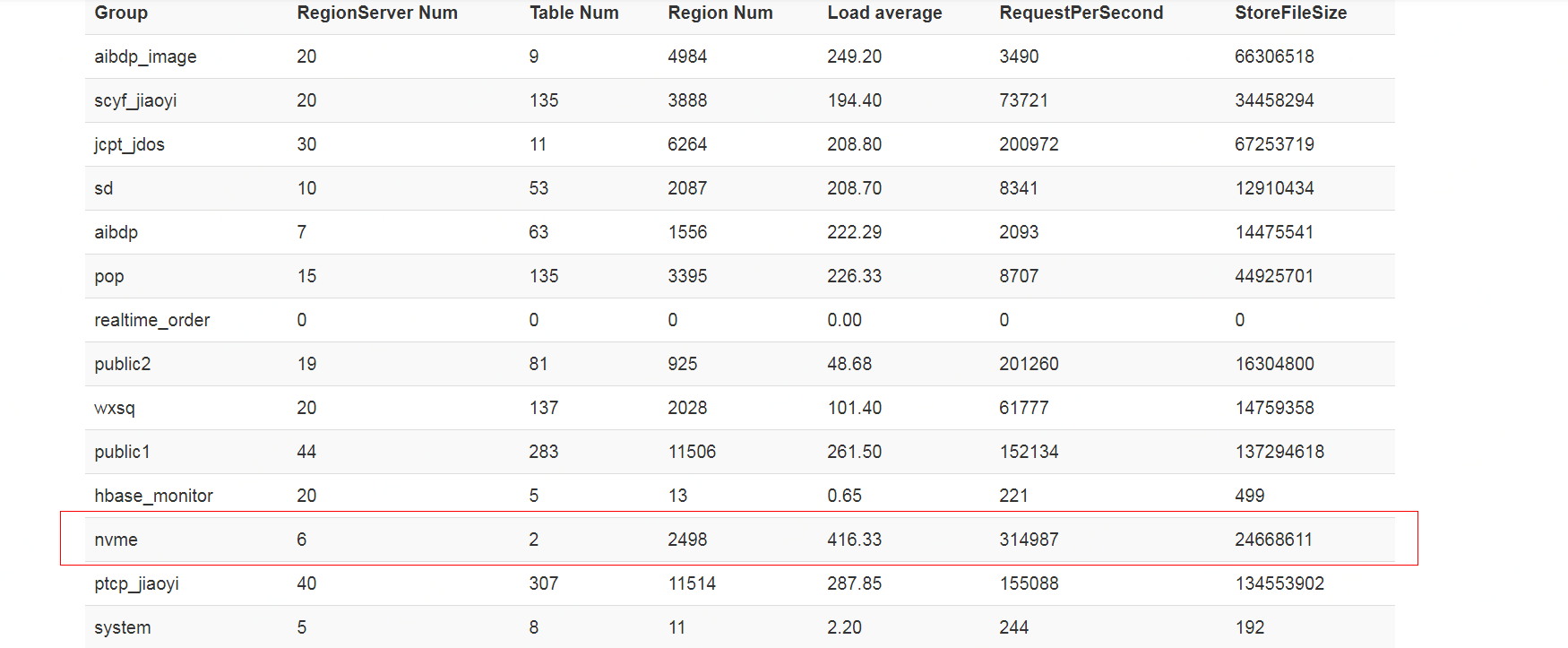
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 集群节点数量 | 硬件配置 | 操作系统 | jdk | 节点数量 | Hadoop | 测试工具 | 发压机硬件配置 |
| Wal写nvme模式 | CPU：32C  内存：256G  硬盘：2\*4T nvem+8\*10THDD | CENTOS7.2 | Jdk1.7 | 6 | Hadoop2.7.1  Hbase1.1.6 | Standard YCSB | CPU：32C  内存：256G  （非数据节点） |

1. 线上测试环境选择将节点加入MAIA集群，此集群为ADAM集群的从集群，将6个节点单独分为nvme组，为了避免对其他业务的影响，测试选择用表为hbase\_monitor:tsdb，此表为hbase的监控信息表
   1. 测试准备

线上集群：为了避免对线上业务影响，6个测试节点加入从集群MAIA，并且单独建立nvme分组

测试表：使用hbase监控表hbase\_monitor:tsdb, 此表写入压力较大

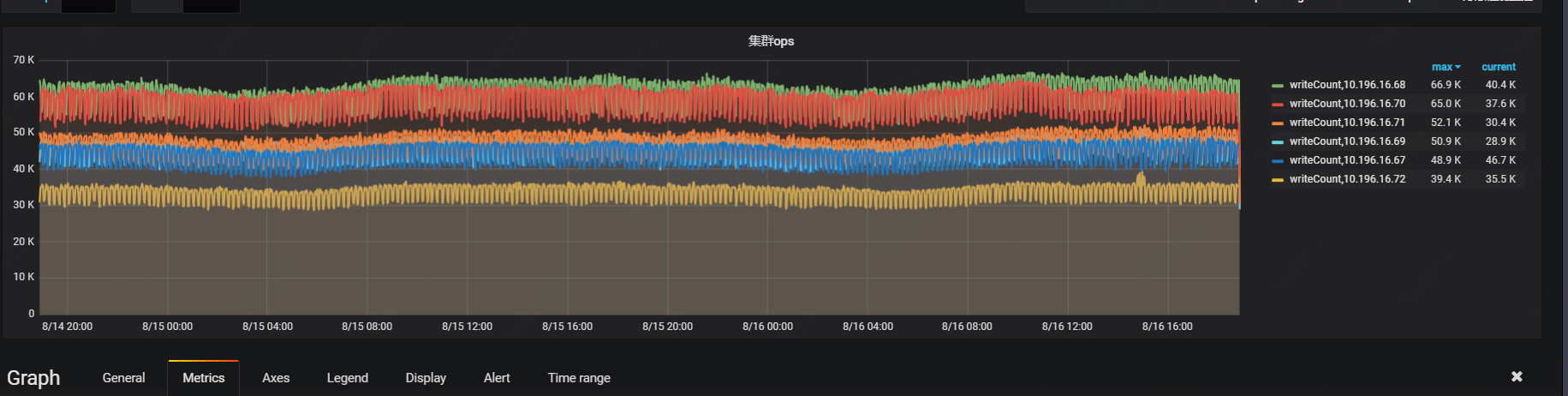
每秒请求次数结果：



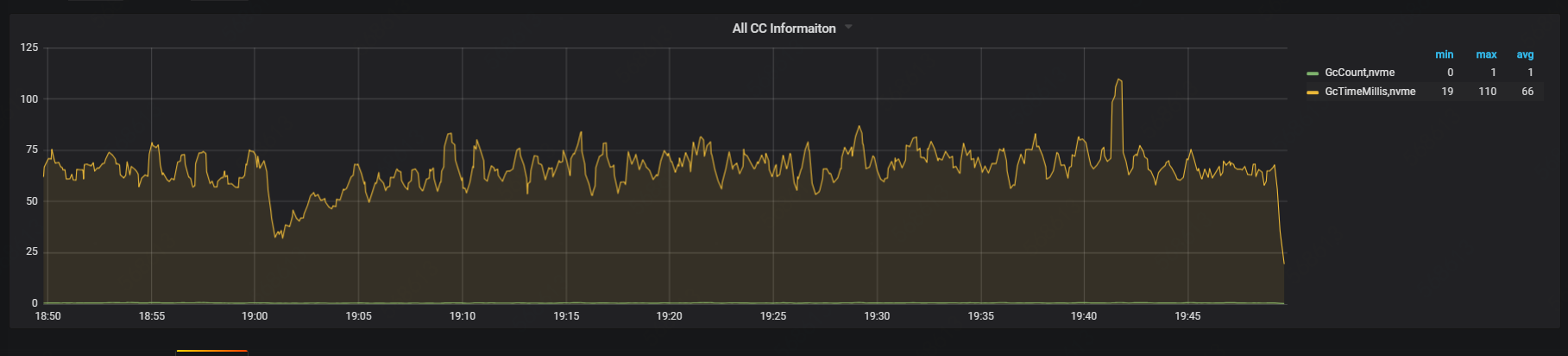
方法性能监控结果：



平均写入请求监控结果：



Gc监控结果：



测试尚存在的问题:

1. 在compact对性能影响的测试中，发现一个有趣现象，当单节点HDD硬盘数量为4的时候，compact的处理性能有明显下降，读写压力较大时，major\_compact进行速度较慢； 但是8\*HDD配置则没有此现象，尚需要进一步测试