HDFS性能测试

一、集群信息

```
角色：NameNode+ResourceManager+DFSZKFailoverController
设备数：3台
设备资源：256G+64核
10.104.4.41
10.104.5.12
10.104.7.155
角色：Zookeeper+JournalNode
设备数：3台
设备资源：128G+32核
10.104.2.145
10.104.1.174
10.104.1.16
角色：DataNode
设备数：3台
设备资源：128G+32核+4T*12
10.104.5.37
10.104.5.30
10.104.5.33
```

二、吞吐量压测

1、文件=30MB

并发：200

文件大小：30MB

BlockSize：128MB

读：50%

写：50%

硬盘使用率20%压测结果如下

总写入量≈2.1GB/s

总读取量≈0.7GB/s

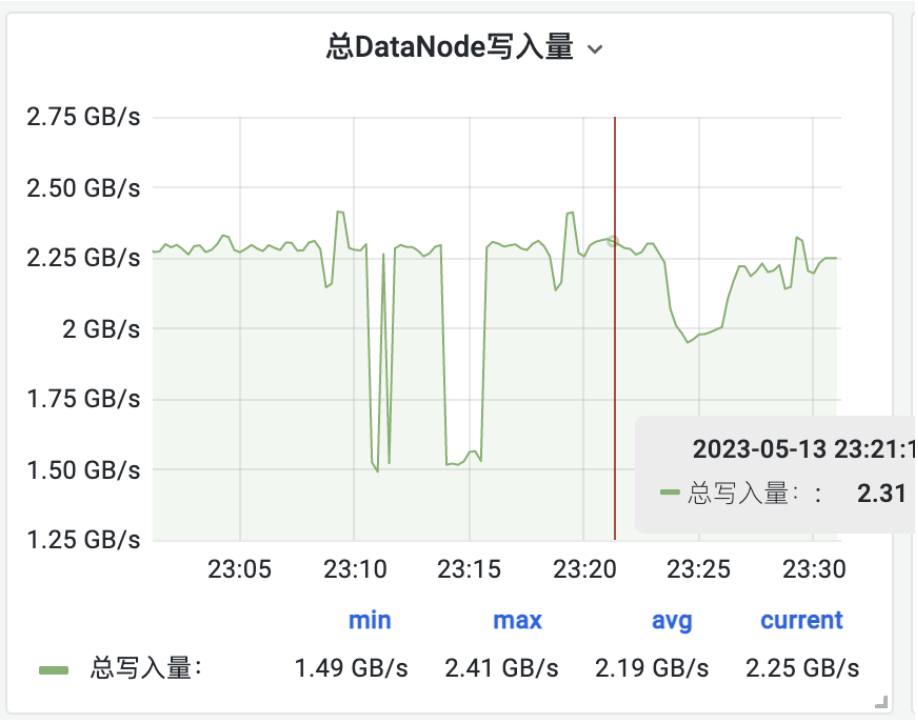
监控中的写入量和读取量中的掉坑，当时在调整客户端并发，请求量掉了下来，不是业务处理能力不够导致的掉坑。

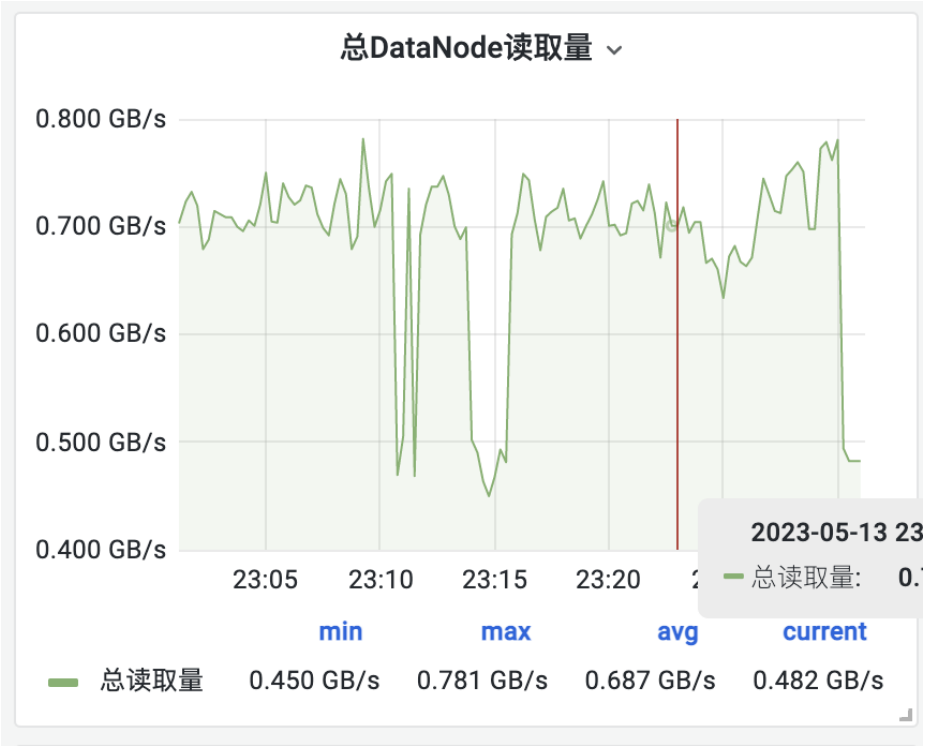
磁盘IO在60%以上，瓶颈在于客户端请求资源不足，增加并发，并不会增加请求量。

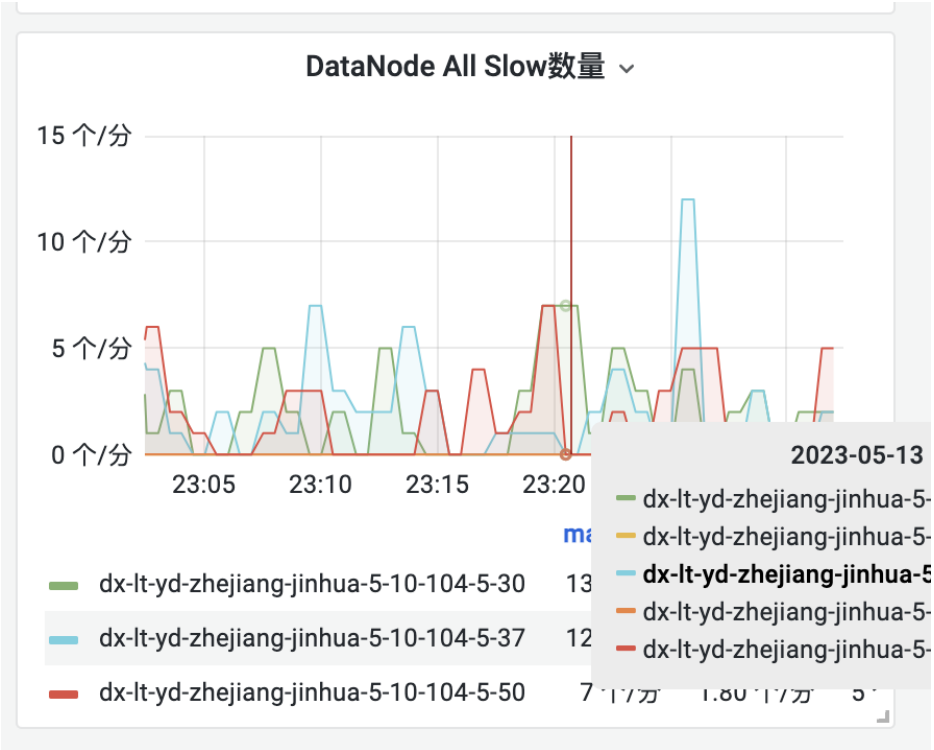
慢节点日志少于10个左右/分

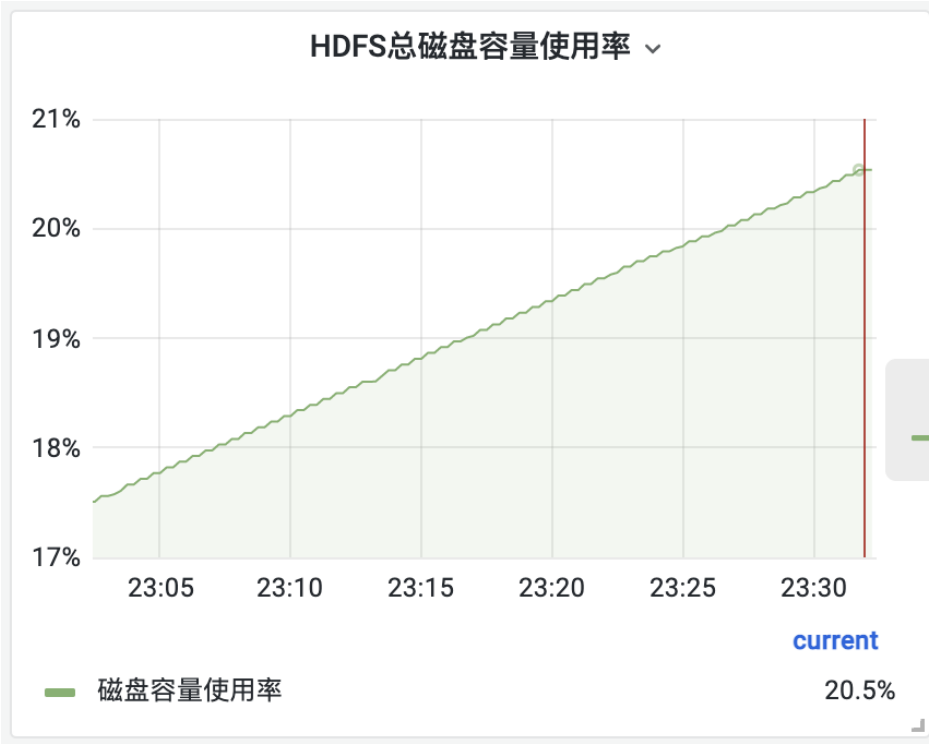
平均每块硬盘写入性能≈59MB/s

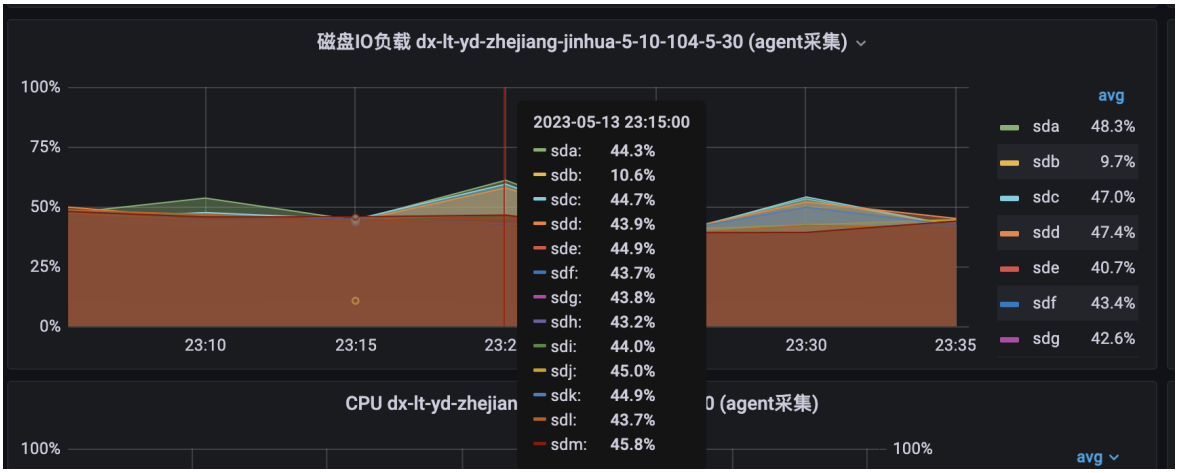
平均每块硬盘写入性能≈19MB/s

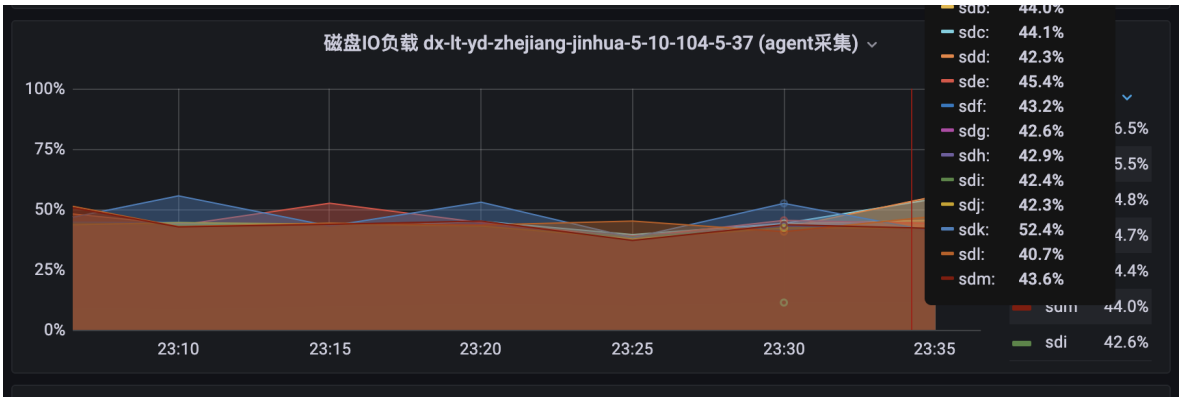


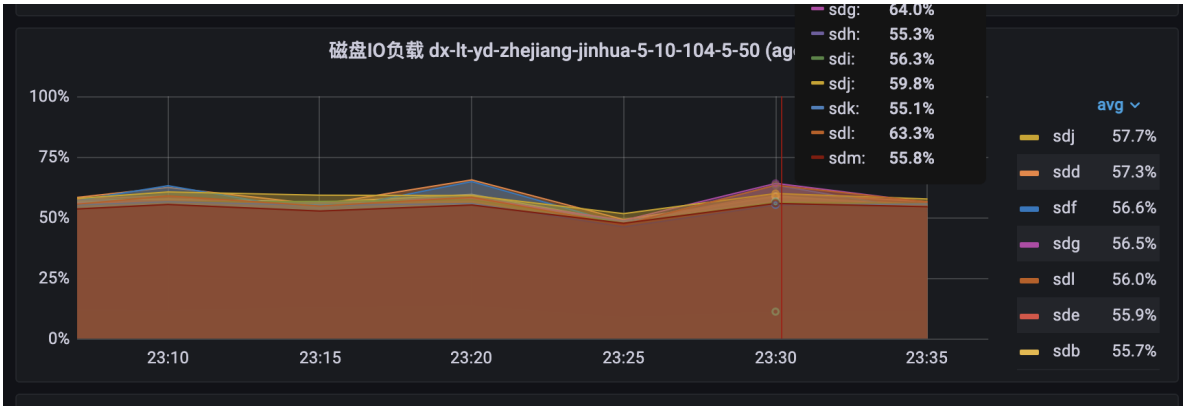

总DataNode读取量
0.800 GB/s
0.700 GB/s
0.600 GB/s
0.500 GB/s
0.400 GB/s
23:05
23:10
23:15
23:20
2023-05-13 23
总读取量:
min
max
avg
current
总读取量
0.450 GB/s
0.781 GB/s
0.687 GB/s
0.482 GB/s


DataNode All Slow数量
15 个/分
10 个/分
5 个/分
0 个/分
23:05
23:10
23:15
23:20
2023-05-13
dx-lt-yd-zhejiang-jinhua-5-10-104-5-30
dx-lt-yd-zhejiang-jinhua-5-10-104-5-37
dx-lt-yd-zhejiang-jinhua-5-10-104-5-50


HDFS总磁盘容量使用率
21%
20%
19%
18%
17%
23:05
23:10
23:15
23:20
23:25
23:30
current
磁盘容量使用率
20.5%


磁盘IO负载 dx-lt-yd-zhejiang-jinhua-5-10-104-5-30 (agent采集)
100%
75%
50%
25%
0%
23:10
23:15
23:30
23:35
2023-05-13 23:15:00
sda: 44.3%
sdb: 10.6%
sdc: 44.7%
sdd: 43.9%
sde: 44.9%
sdf: 43.7%
sdg: 43.8%
sdh: 43.2%
sdi: 44.0%
sdj: 45.0%
sdk: 44.9%
sdl: 43.7%
sdm: 45.8%
avg
sda 48.3%
sdb 9.7%
sdc 47.0%
sdd 47.4%
sde 40.7%
sdf 43.4%
sdg 42.6%
CPU dx-lt-yd-zhejian
(agent采集)
100%
100%
avg


磁盘IO负载 dx-lt-yd-zhejiang-jinhua-5-10-104-5-37 (agent采集)
100%
75%
50%
25%
0%
23:10
23:15
23:20
23:25
23:30
23:35
sdc: 44.1%
sdd: 42.3%
sde: 45.4%
sdf: 43.2%
sdg: 42.6%
sdh: 42.9%
sdi: 42.4%
sdj: 42.3%
sdk: 52.4%
sdl: 40.7%
sdm: 43.6%
sdi 42.6%



2、文件=100MB

并发：200

文件大小：100MB

BlockSize：128MB

读：50%

写：50%

硬盘使用率25%压测结果如下

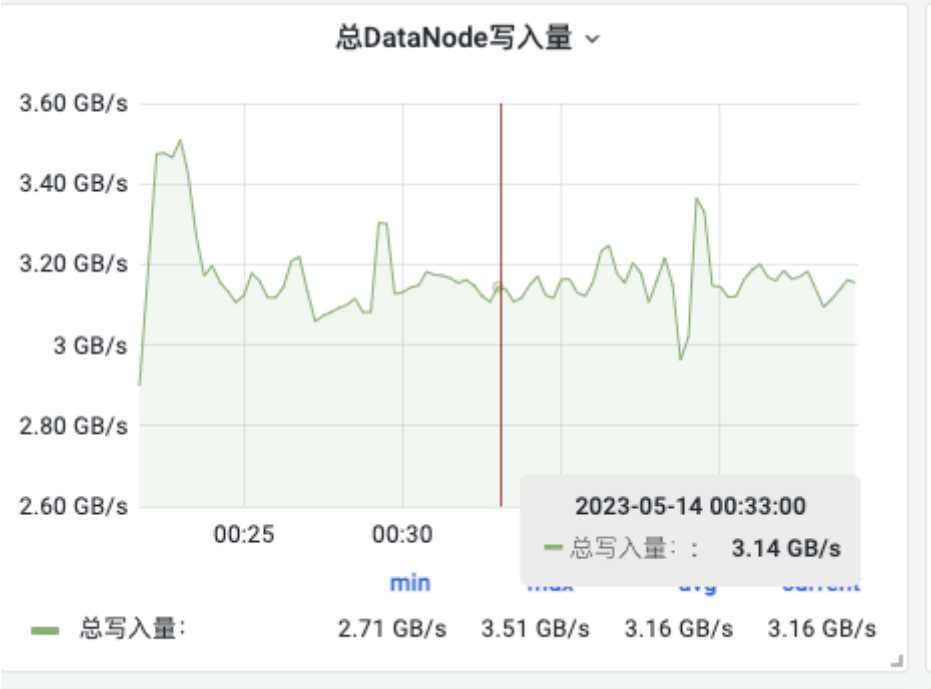
总写入量≈3.0GB/s

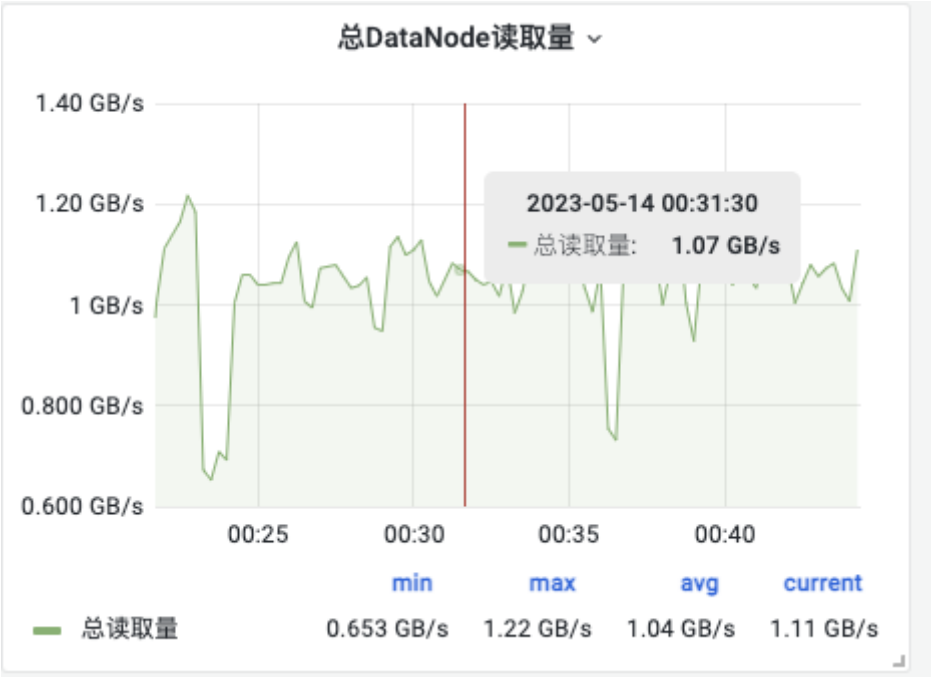
总读取量≈1.0GB/s

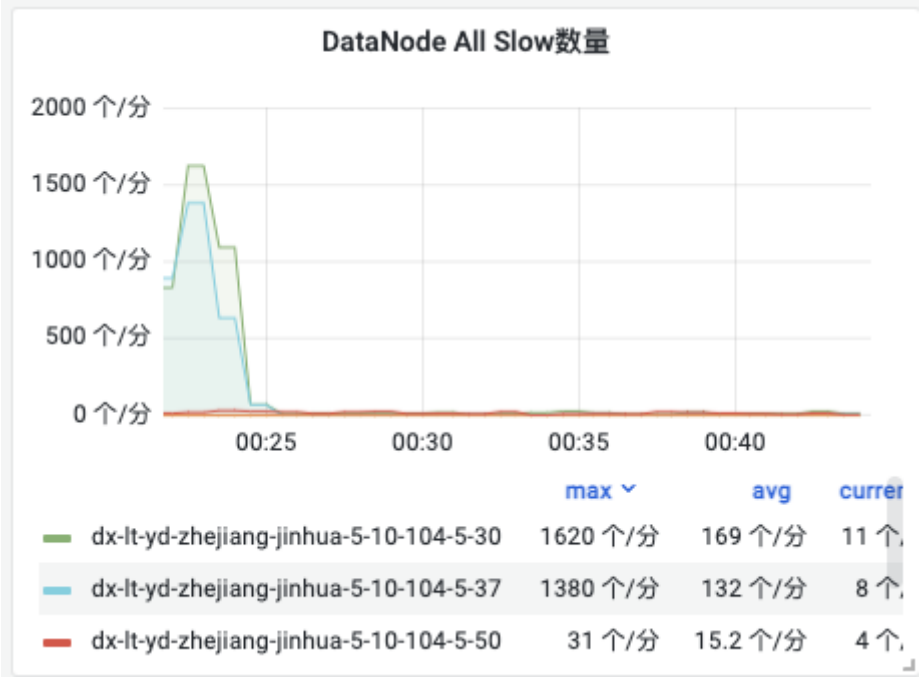
磁盘IO基本占满，瓶颈在于磁盘，慢节点日志少于10个左右/分

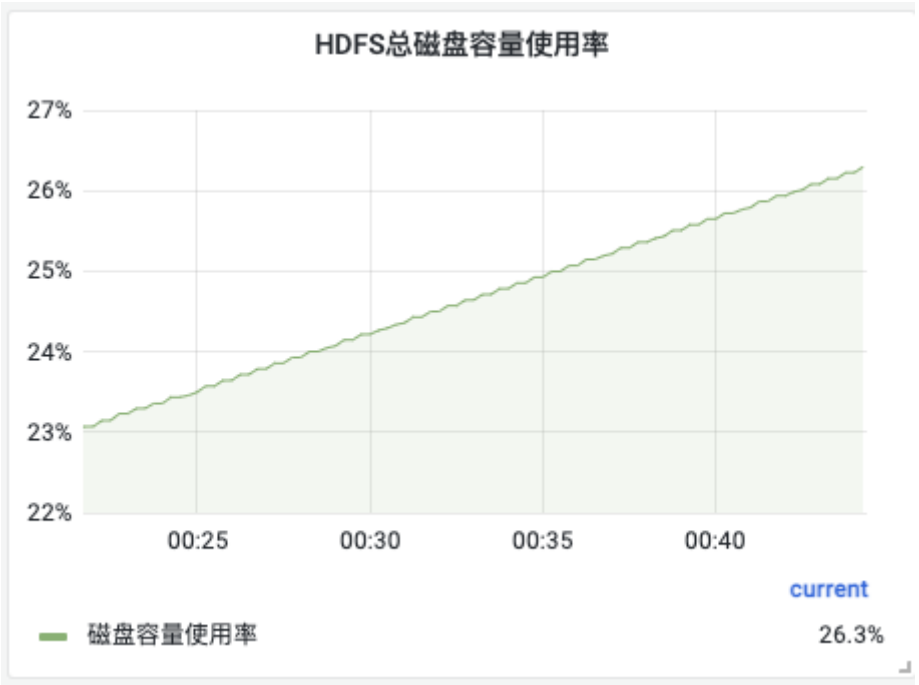
平均每块硬盘写入性能≈85MB/s

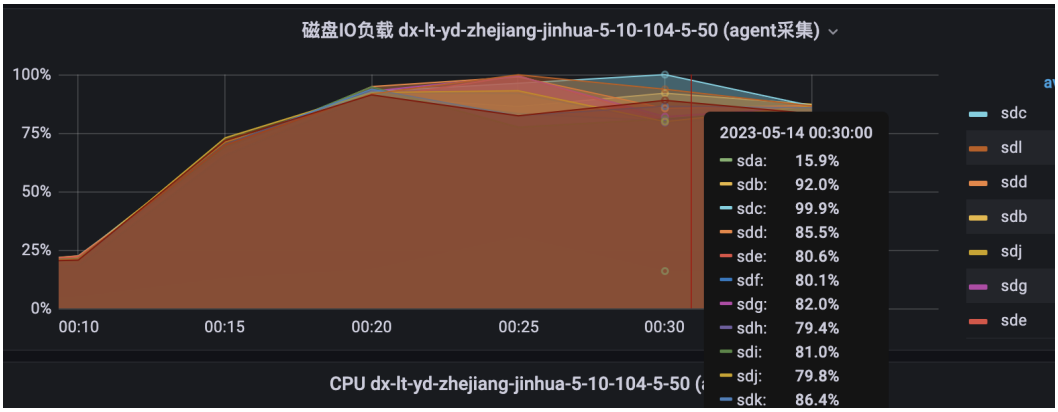
平均每块硬盘写入性能≈28MB/s

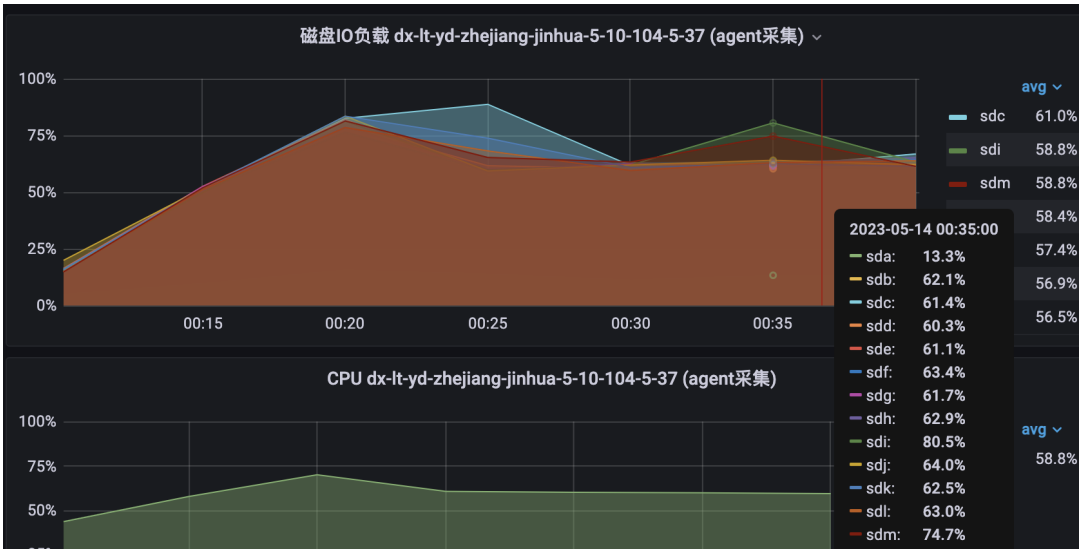

总DataNode写入量
3.60 GB/s
3.40 GB/s
3.20 GB/s
3 GB/s
2.80 GB/s
2.60 GB/s
00:25
00:30
2023-05-14 00:33:00
总写入量: : 3.14 GB/s
min
总写入量: 2.71 GB/s 3.51 GB/s 3.16 GB/s 3.16 GB/s

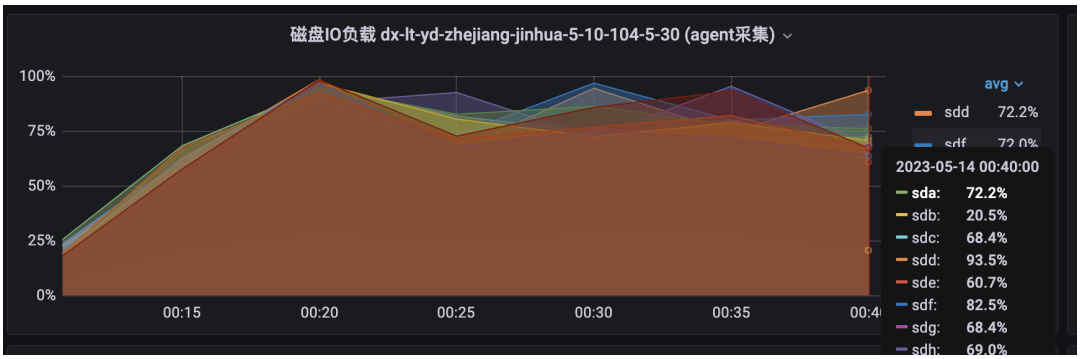

总DataNode读取量
1.40 GB/s
1.20 GB/s
1 GB/s
0.800 GB/s
0.600 GB/s
00:25
00:30
00:35
00:40
2023-05-14 00:31:30
总读取量: 1.07 GB/s
min
max
avg
current
总读取量 0.653 GB/s 1.22 GB/s 1.04 GB/s 1.11 GB/s


DataNode All Slow数量
2000 个/分
1500 个/分
1000 个/分
500 个/分
0 个/分
00:25
00:30
00:35
00:40
max
avg
curren
dx-lt-yd-zhejiang-jinhua-5-10-104-5-30
1620 个/分
169 个/分
11 个
dx-lt-yd-zhejiang-jinhua-5-10-104-5-37
1380 个/分
132 个/分
8 个
dx-lt-yd-zhejiang-jinhua-5-10-104-5-50
31 个/分
15.2 个/分
4 个


HDFS总磁盘容量使用率
27%
26%
25%
24%
23%
22%
00:25
00:30
00:35
00:40
current
磁盘容量使用率
26.3%







硬盘使用率75%压测结果如下

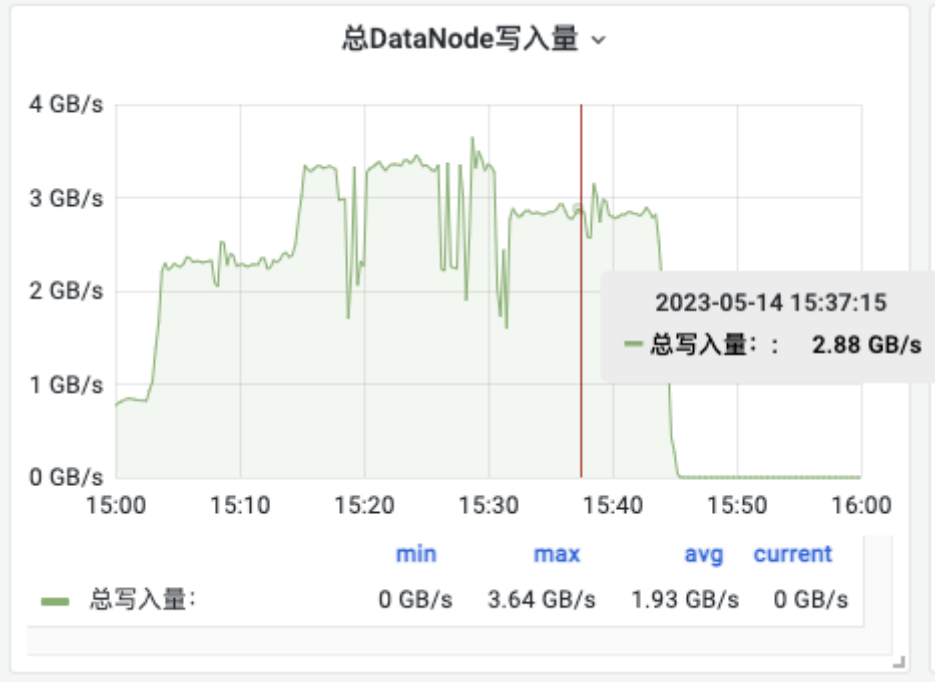
总写入量≈2.8GB/s

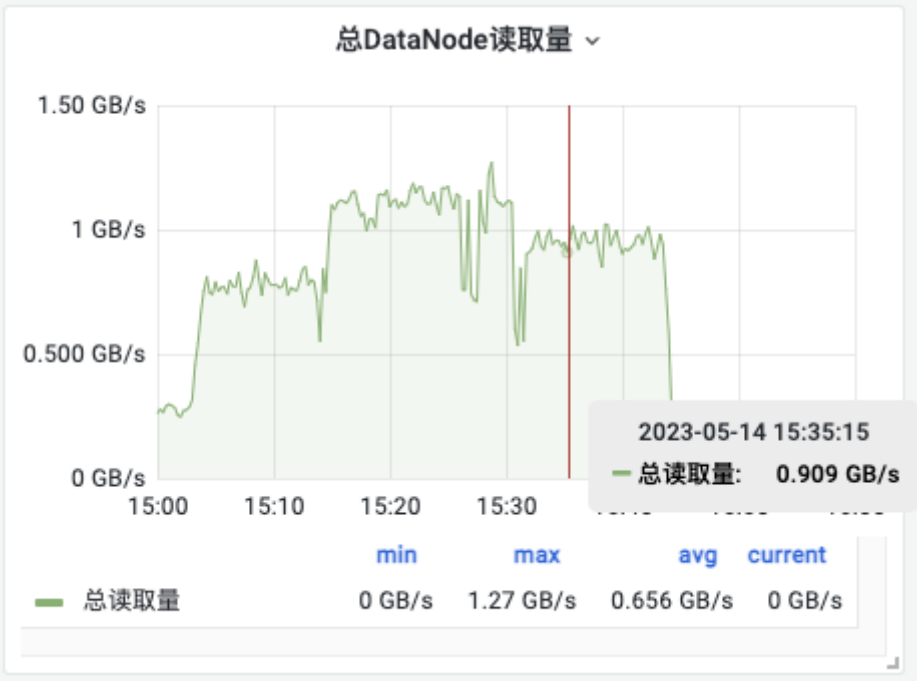
总读取量≈0.9GB/s

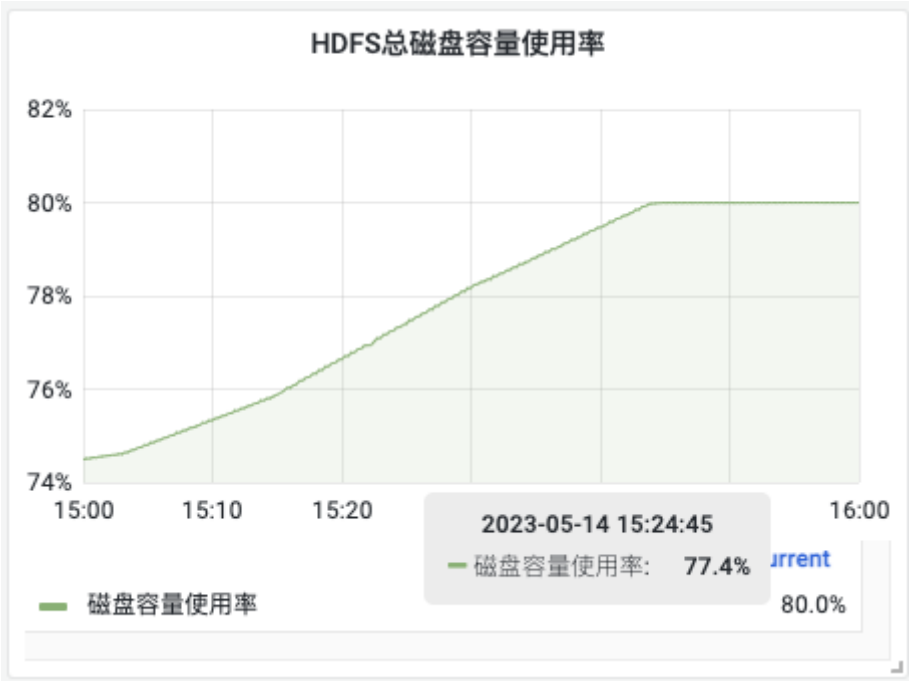
磁盘IO基本占满，瓶颈在于磁盘，慢节点日志少于10个左右/分

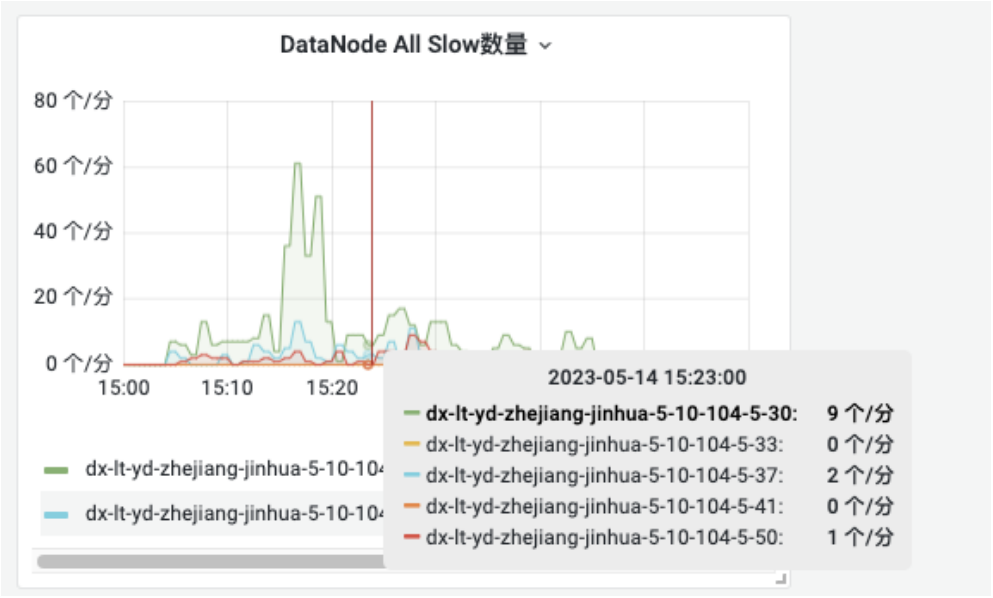
平均每块硬盘写入性能≈79MB/s

平均每块硬盘写入性能≈25MB/s









三、RPS压测

DataNode节点扩容至 15 台

```
DataNode 设备数
5（4T*12 64核 128G）
10（6T*12 32核 64G）
```

1、文件写

并发：400

文件大小：4KB

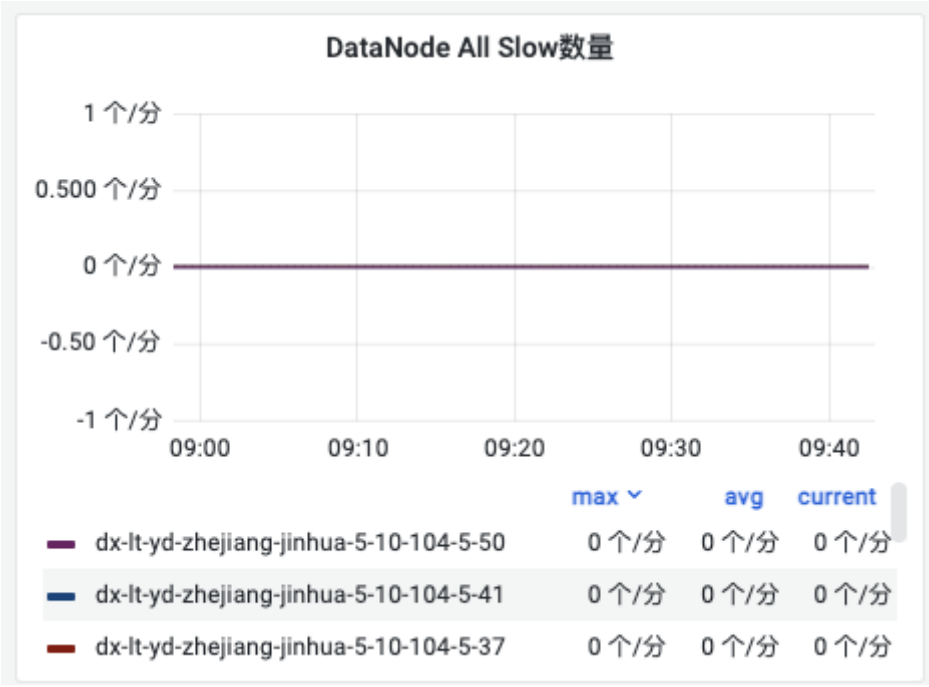
BlockSize：128MB

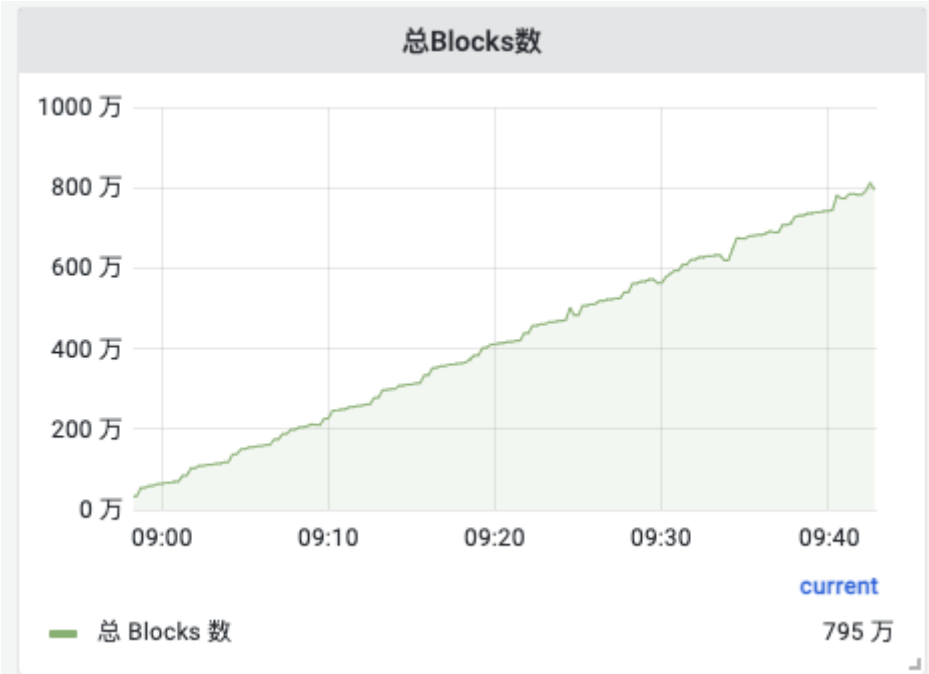
写：100%

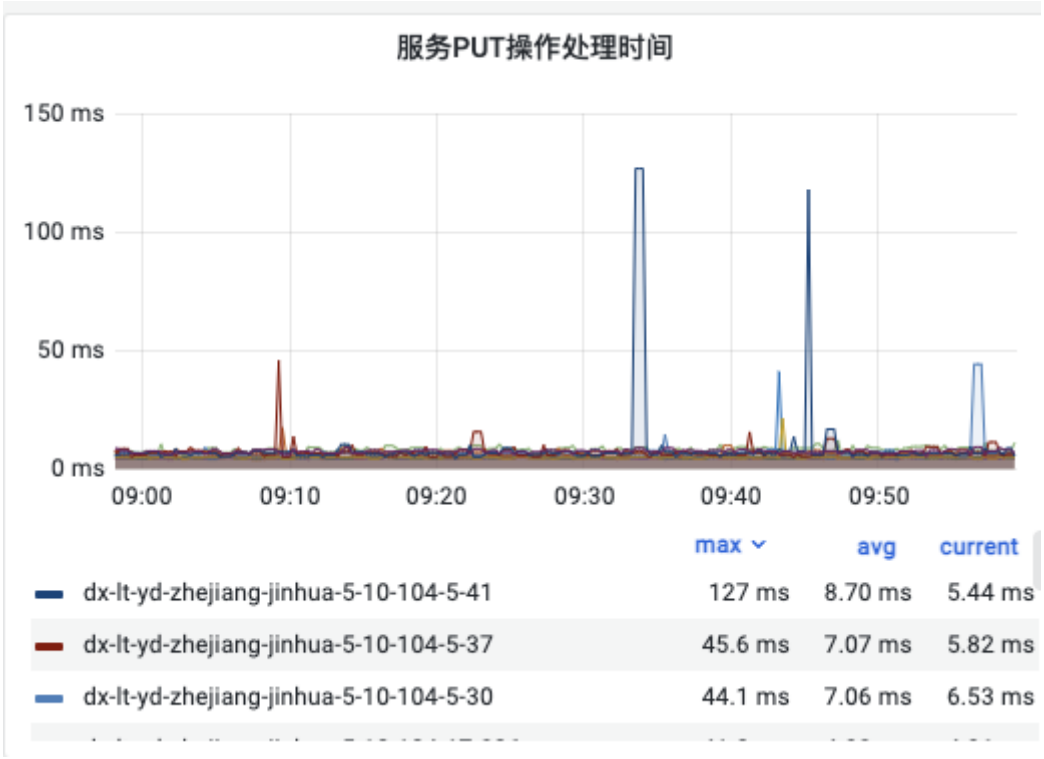
压测结果如下

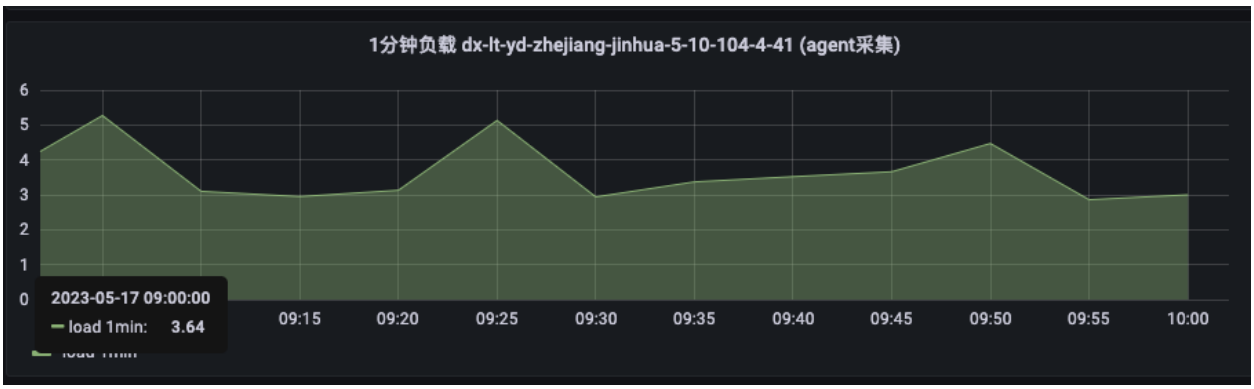
写RPS在 3000/s 左右，无Slow日志、NameNode资源使用不高、磁盘IO率不高

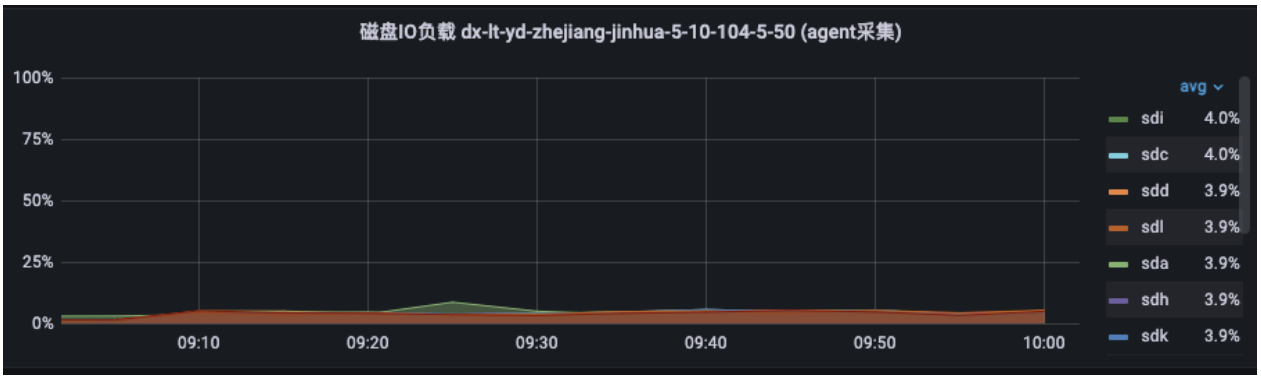
增加并发量后RPS无法增加，瓶颈应该在与NameNode写文件时锁机制


DataNode All Slow数量
1 个/分
0.500 个/分
0 个/分
-0.50 个/分
-1 个/分
09:00
09:10
09:20
09:30
09:40
max
avg
current
dx-lt-yd-zhejiang-jinhua-5-10-104-5-50
0 个/分
0 个/分
0 个/分
dx-lt-yd-zhejiang-jinhua-5-10-104-5-41
0 个/分
0 个/分
0 个/分
dx-lt-yd-zhejiang-jinhua-5-10-104-5-37
0 个/分
0 个/分
0 个/分


总Blocks数
1000 万
800 万
600 万
400 万
200 万
0 万
09:00
09:10
09:20
09:30
09:40
current
总 Blocks 数
795 万


服务PUT操作处理时间
150 ms
100 ms
50 ms
0 ms
09:00
09:10
09:20
09:30
09:40
09:50
max
avg
current
dx-lt-yd-zhejiang-jinhua-5-10-104-5-41 127 ms 8.70 ms 5.44 ms
dx-lt-yd-zhejiang-jinhua-5-10-104-5-37 45.6 ms 7.07 ms 5.82 ms
dx-lt-yd-zhejiang-jinhua-5-10-104-5-30 44.1 ms 7.06 ms 6.53 ms


1分钟负载 dx-lt-yd-zhejiang-jinhua-5-10-104-4-41 (agent采集)
6
5
4
3
2
1
0
2023-05-17 09:00:00
load 1min: 3.64
09:15
09:20
09:25
09:30
09:35
09:40
09:45
09:50
09:55
10:00


磁盘IO负载 dx-lt-yd-zhejiang-jinhua-5-10-104-5-50 (agent采集)
100%
75%
50%
25%
0%
09:10
09:20
09:30
09:40
09:50
10:00
avg
sdi 4.0%
sdc 4.0%
sdd 3.9%
sdl 3.9%
sda 3.9%
sdh 3.9%
sdk 3.9%

2、读文件

并发：400

文件大小：4KB

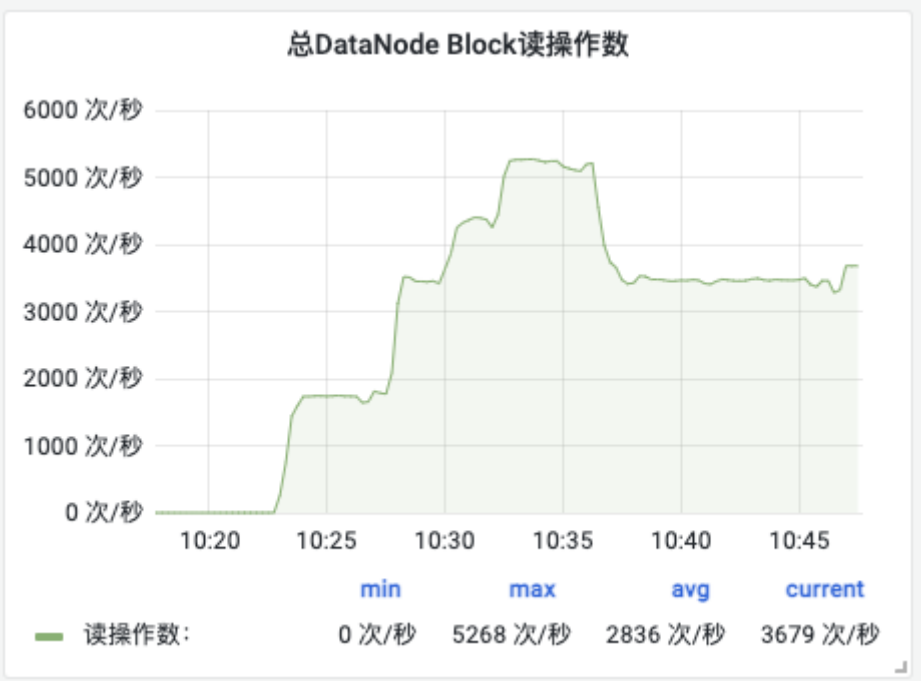
BlockSize：128MB

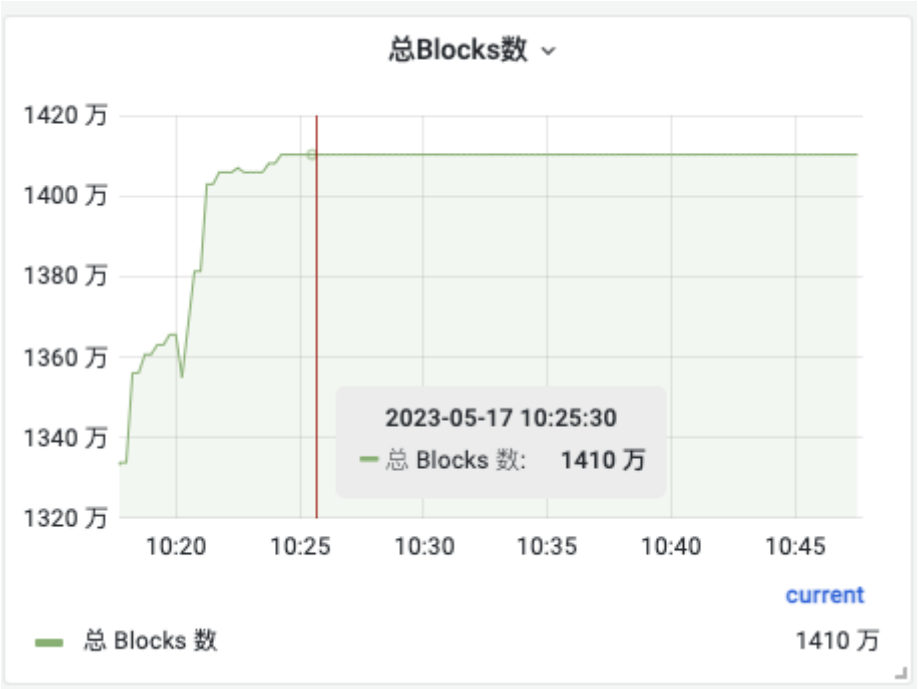
读：100%

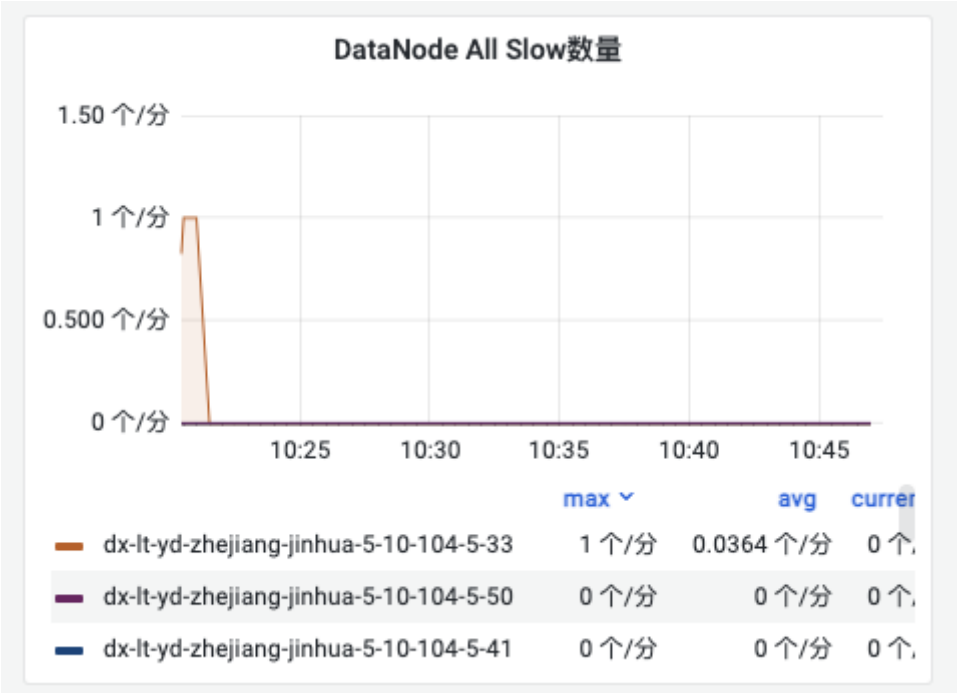
压测结果如下

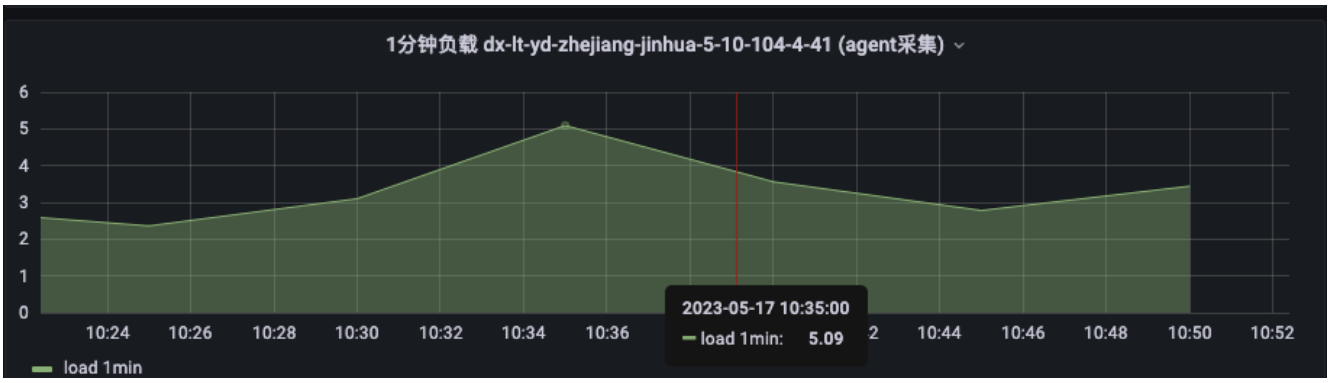
读RPS在 4000/s 左右，无Slow日志、NameNode资源使用不高、DataNode磁盘IO在10%左右

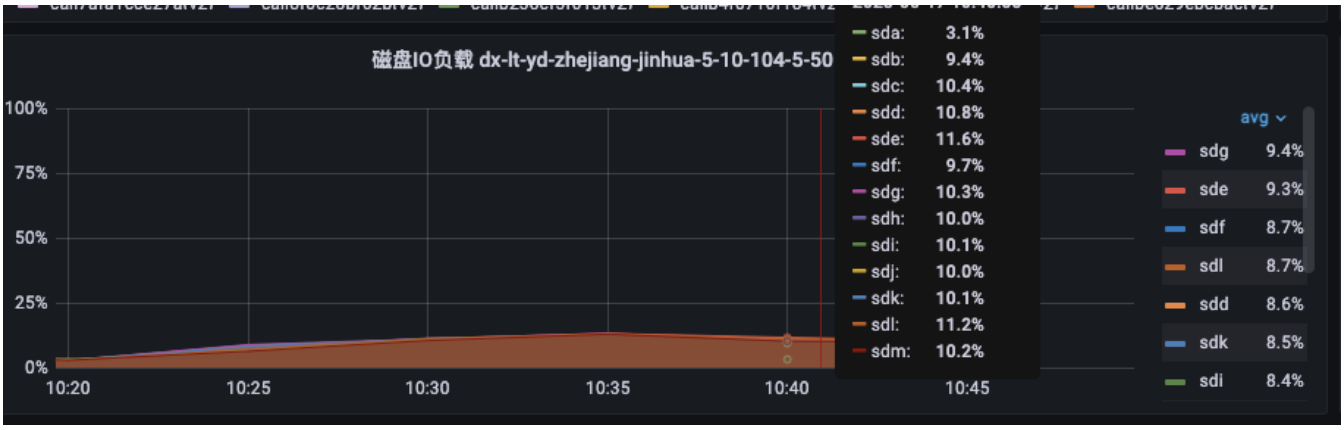
增加并发量后RPS，峰值可以达到8000/s，无失败读取操作，认为这个值还可以往上加，但高并发下DataNode的LogWarn日志特别多，RPC操作会被打断（不影响读取），尤其旧金华HDFS的 5 台设备（上面有K8S，负载较高）。

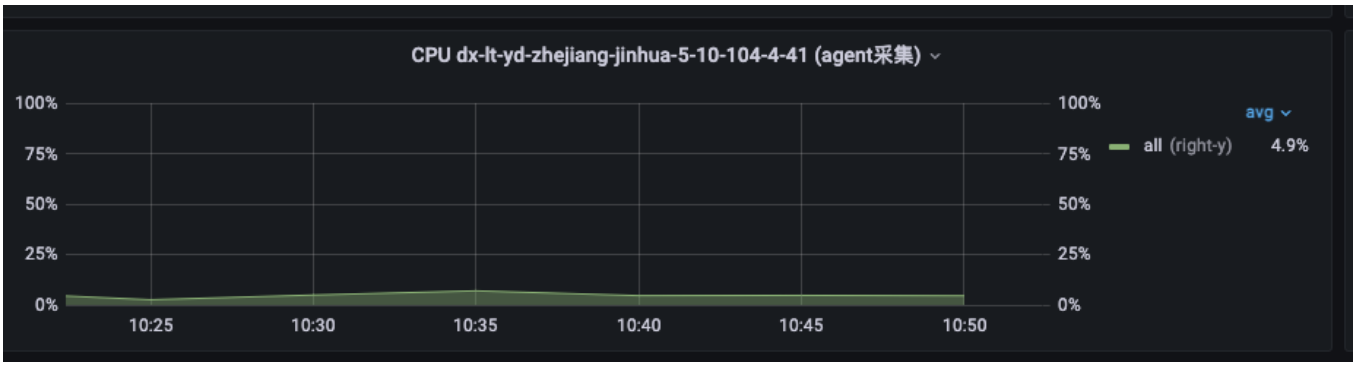


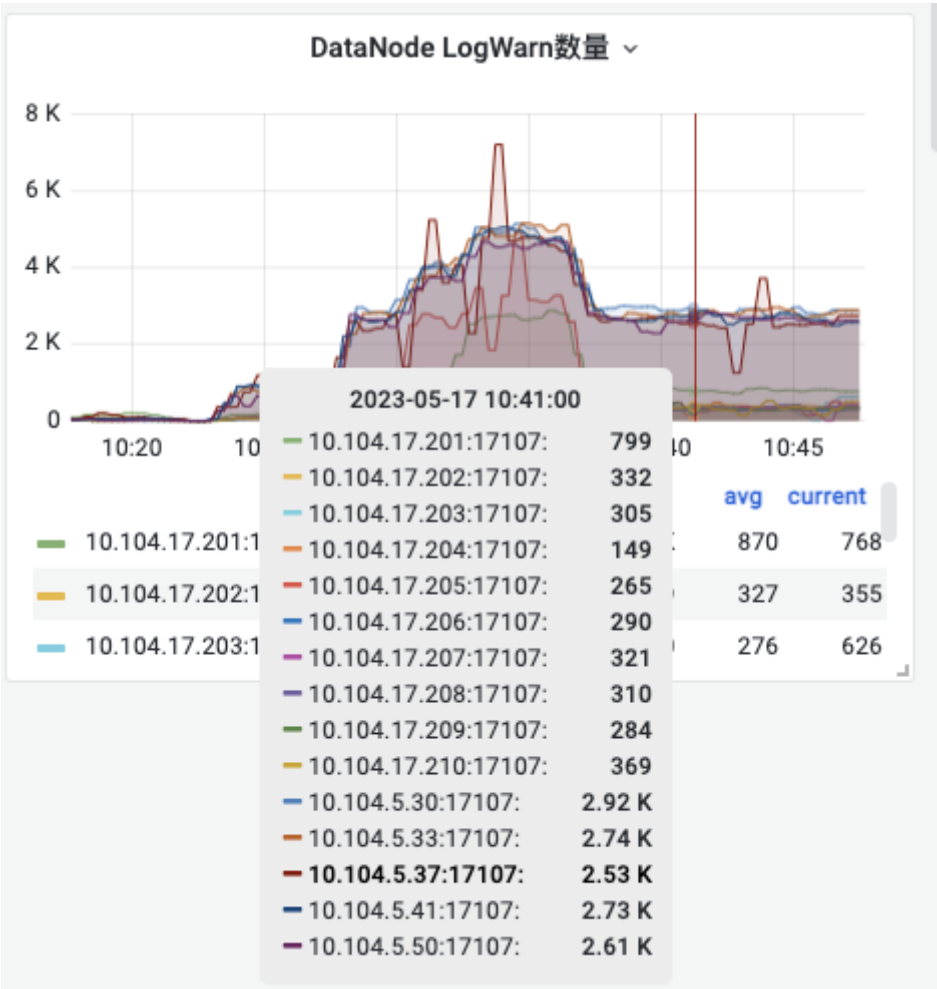

总Blocks数
1420 万
1400 万
1380 万
1360 万
1340 万
1320 万
10:20
10:25
10:30
10:35
10:40
10:45
2023-05-17 10:25:30
总 Blocks 数: 1410 万
current
总 Blocks 数
1410 万


DataNode All Slow数量
1.50 个/分
1 个/分
0.500 个/分
0 个/分
10:25
10:30
10:35
10:40
10:45
max
avg
dx-lt-yd-zhejiang-jinhua-5-10-104-5-33 1 个/分 0.0364 个/分
dx-lt-yd-zhejiang-jinhua-5-10-104-5-50 0 个/分 0 个/分
dx-lt-yd-zhejiang-jinhua-5-10-104-5-41 0 个/分 0 个/分

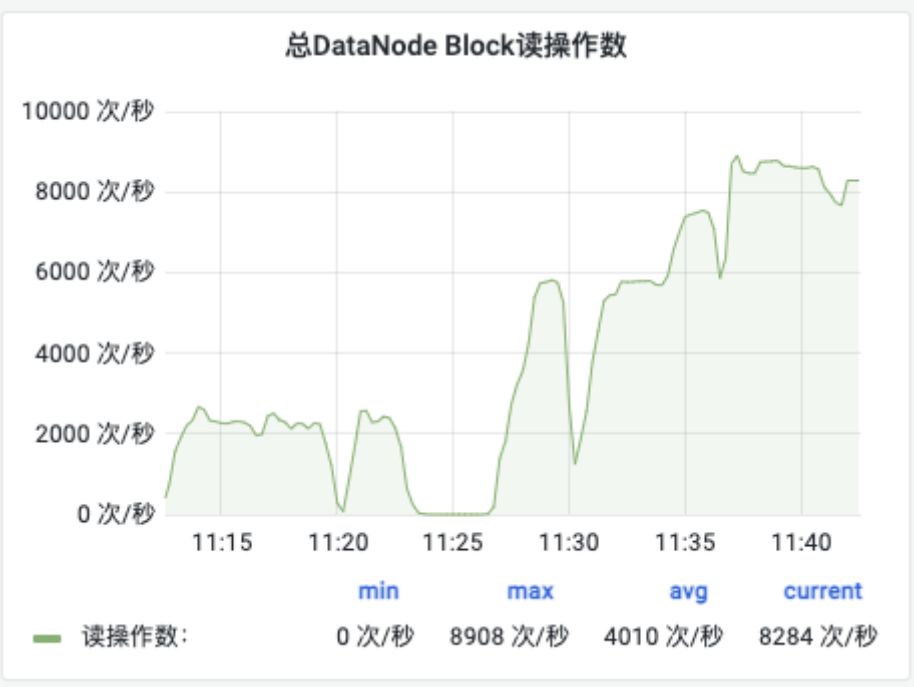

1分钟负载 dx-lt-yd-zhejiang-jinhua-5-10-104-4-41 (agent采集)
6
5
4
3
2
1
0
10:24
10:26
10:28
10:30
10:32
10:34
10:36
2023-05-17 10:35:00
load 1min: 5.09
10:44
10:46
10:48
10:50
10:52
load 1min


磁盘IO负载 dx-lt-yd-zhejiang-jinhua-5-10-104-5-50
100%
75%
50%
25%
0%
10:20
10:25
10:30
10:35
10:40
10:45
sda: 3.1%
sdb: 9.4%
sdc: 10.4%
sdd: 10.8%
sde: 11.6%
sdf: 9.7%
sdg: 10.3%
sdh: 10.0%
sdi: 10.1%
sdj: 10.0%
sdk: 10.1%
sdl: 11.2%
sdm: 10.2%
avg
sdg 9.4%
sde 9.3%
sdf 8.7%
sdl 8.7%
sdd 8.6%
sdk 8.5%
sdi 8.4%


CPU dx-lt-yd-zhejiang-jinhua-5-10-104-4-41 (agent采集)
100%
75%
50%
25%
0%
10:25
10:30
10:35
10:40
10:45
10:50
avg
all (right-y) 4.9%




```
        at java.lang.Thread.run(Thread.java:748)
2023-05-17 10:43:28,451 WARN org.apache.hadoop.ipc.Client: interrupted waiting to send rpc request to server
java.lang.InterruptedException
        at java.util.concurrent.FutureTask.awaitDone(FutureTask.java:404)
        at java.util.concurrent.FutureTask.get(FutureTask.java:191)
        at org.apache.hadoop.ipc.Client$Connection.sendRpcRequest(Client.java:1148)
        at org.apache.hadoop.ipc.Client.call(Client.java:1409)
        at org.apache.hadoop.ipc.Client.call(Client.java:1367)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:228)
        at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:116)
        at com.sun.proxy.$Proxy26.getBlockLocations(Unknown Source)
        at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.getBlockLocations(ClientNamenodeProtocolTranslatorPB.java:317)
        at sun.reflect.GeneratedMethodAccessor150.invoke(Unknown Source)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.hdfs.server.namenode.ha.RequestHedgingProxyProvider$RequestHedgingInvocationHandler$1.call(RequestHedgingProxyProvider.java:135)
        at java.util.concurrent.FutureTask.run(FutureTask.java:266)
        at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
        at java.util.concurrent.FutureTask.run(FutureTask.java:266)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:748)
```



3、读写文件

并发：400

文件大小：4KB

BlockSize：128MB

写：50%

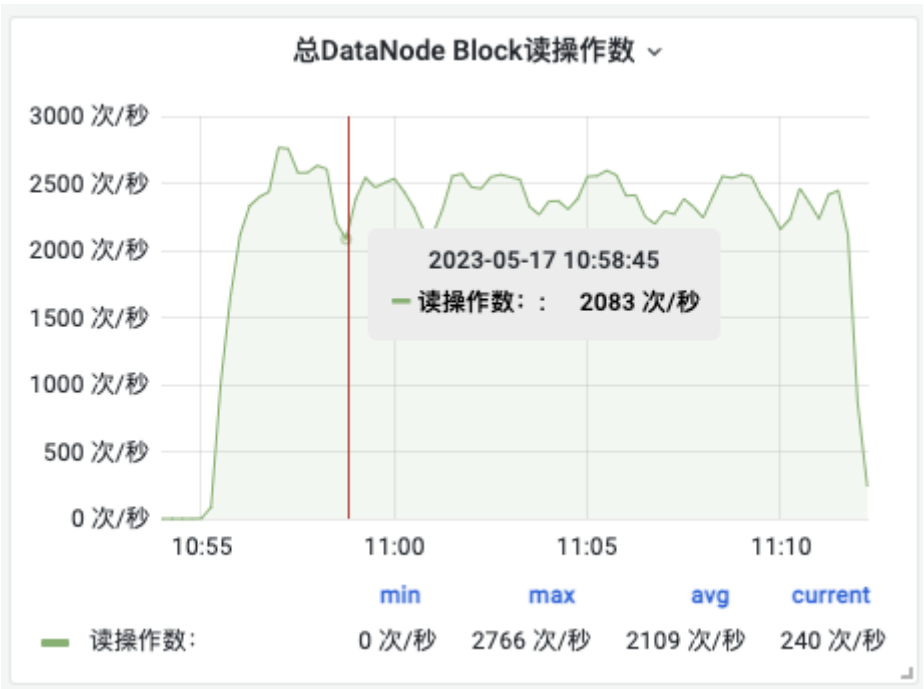
读：50%

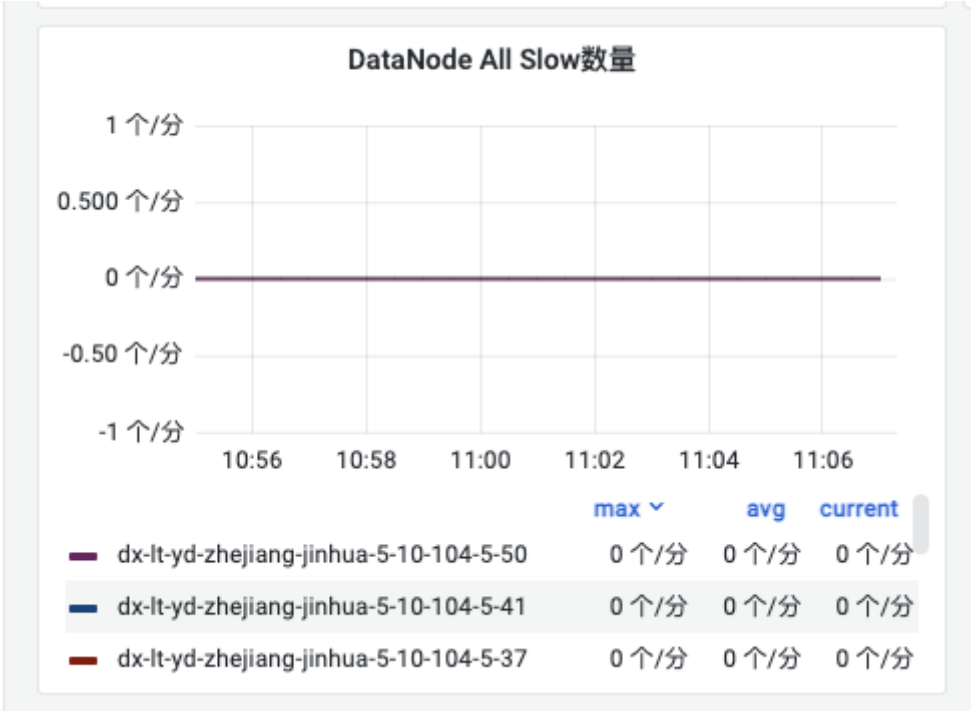
压测结果如下

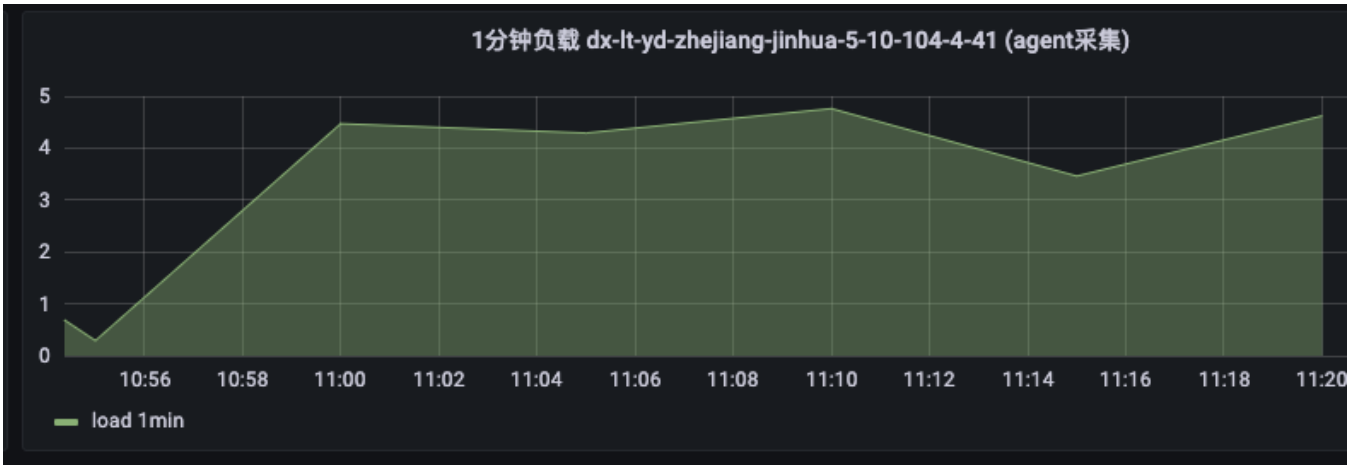
读写RPS都在 2000/s 左右，无Slow日志、NameNode资源使用不高、磁盘IO率不高

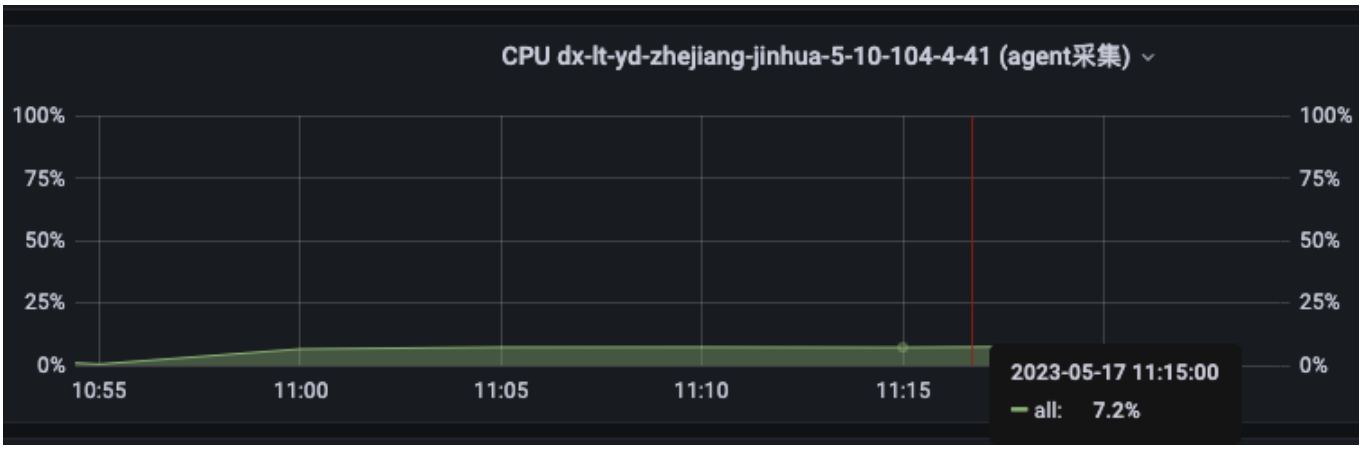
增加并发后RPS无升高，瓶颈和写一致

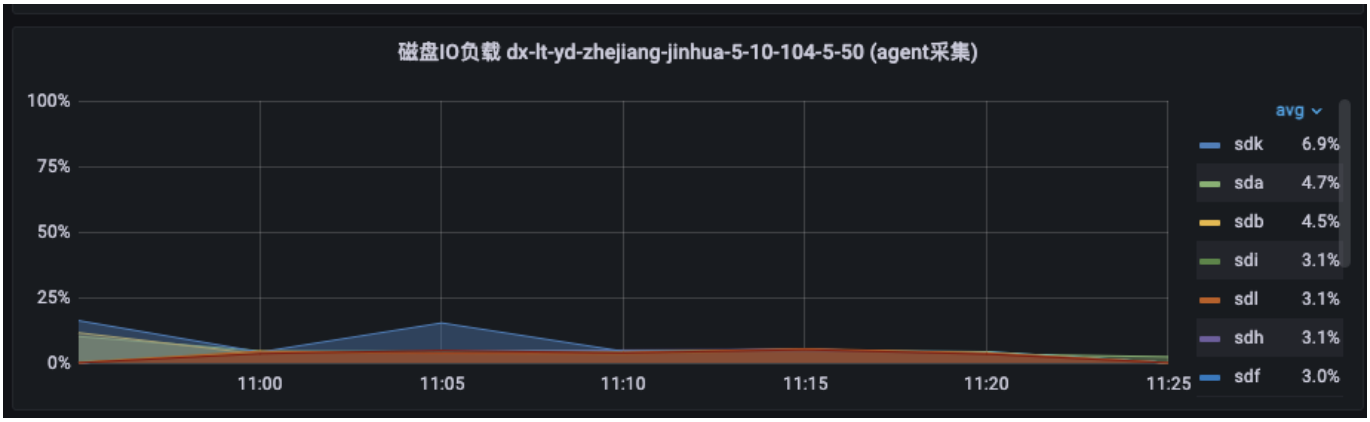

创建文件操作数
3000 个/秒
2500 个/秒
2000 个/秒
1500 个/秒
1000 个/秒
500 个/秒
0 个/秒
10:55
11:00
11:05
11:10
max
avg
current
10.104.4.41:17106
2705 个/秒
2094 个/秒
34.2 个/秒
10.104.5.12:17106
0 个/秒
0 个/秒
0 个/秒
10.104.7.155:17106
0 个/秒
0 个/秒
0 个/秒


总DataNode Block读操作数
3000 次/秒
2500 次/秒
2000 次/秒
1500 次/秒
1000 次/秒
500 次/秒
0 次/秒
2023-05-17 10:58:45
读操作数：： 2083 次/秒
10:55
11:00
11:05
11:10
min
max
avg
current
读操作数:
0 次/秒
2766 次/秒
2109 次/秒
240 次/秒


DataNode All Slow数量
1 个/分
0.500 个/分
0 个/分
-0.50 个/分
-1 个/分
10:56
10:58
11:00
11:02
11:04
11:06
max
avg
current
dx-lt-yd-zhejiang-jinhua-5-10-104-5-50 0 个/分 0 个/分 0 个/分
dx-lt-yd-zhejiang-jinhua-5-10-104-5-41 0 个/分 0 个/分 0 个/分
dx-lt-yd-zhejiang-jinhua-5-10-104-5-37 0 个/分 0 个/分 0 个/分


1分钟负载 dx-lt-yd-zhejiang-jinhua-5-10-104-4-41 (agent采集)
5
4
3
2
1
0
10:56
10:58
11:00
11:02
11:04
11:06
11:08
11:10
11:12
11:14
11:16
11:18
11:20
load 1min


CPU dx-lt-yd-zhejiang-jinhua-5-10-104-4-41 (agent采集)
100%
75%
50%
25%
0%
10:55
11:00
11:05
11:10
11:15
2023-05-17 11:15:00
all: 7.2%



4、删除文件

并发：400

文件大小：4KB

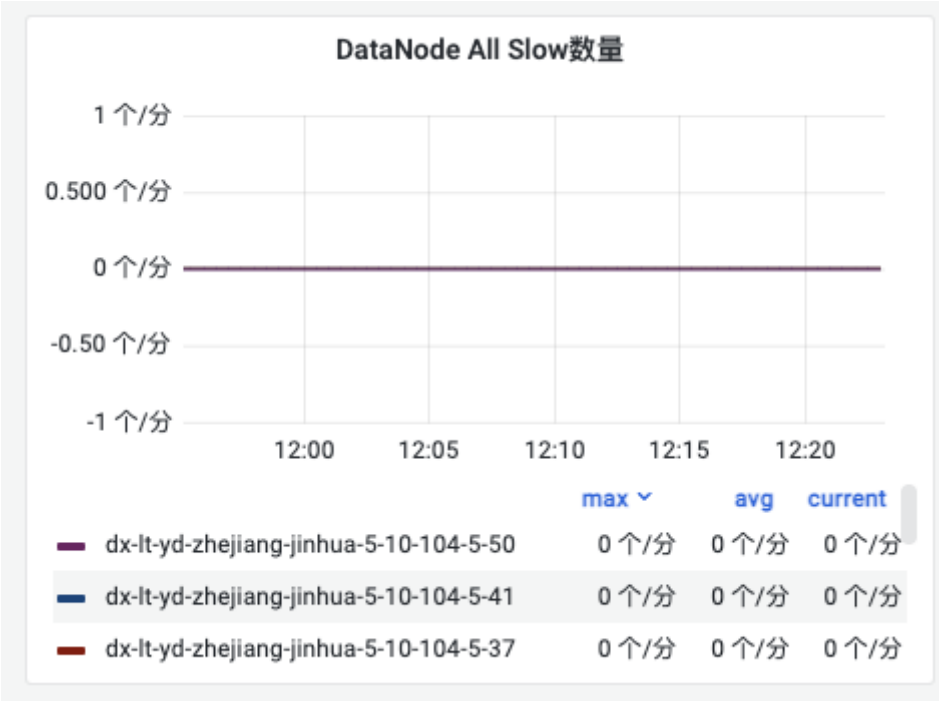
BlockSize：128MB

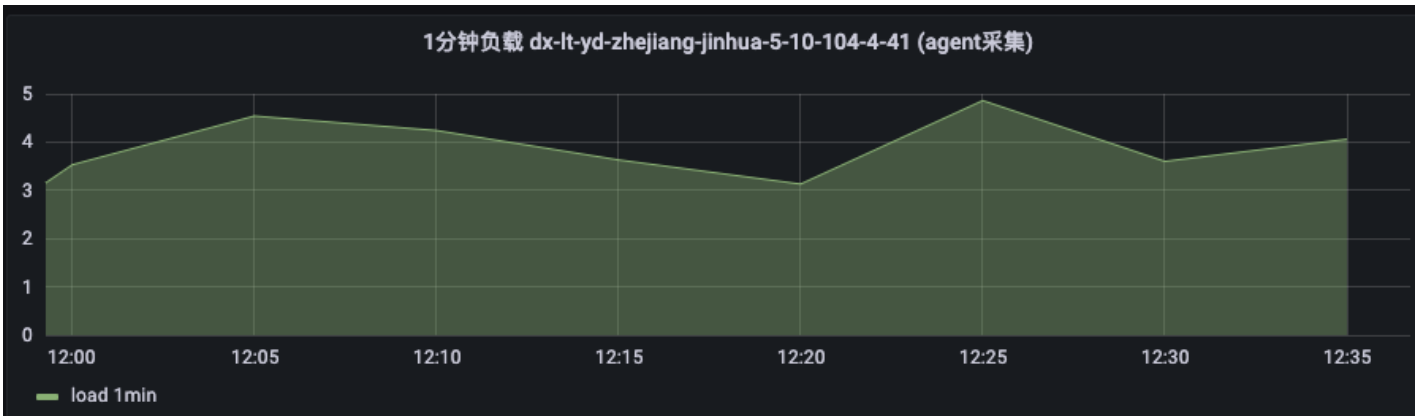
删除：100%

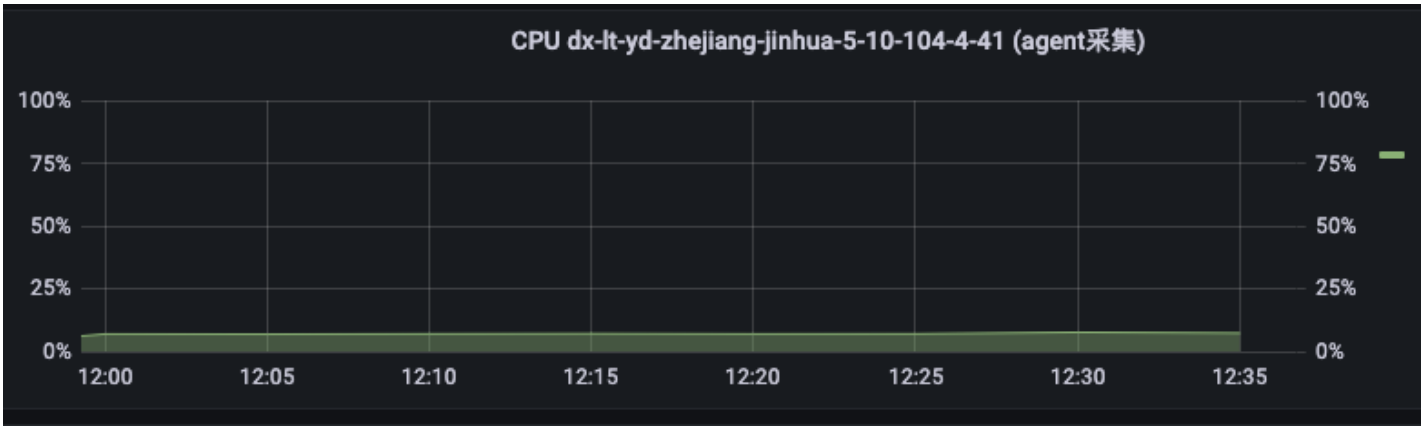
压测结果如下

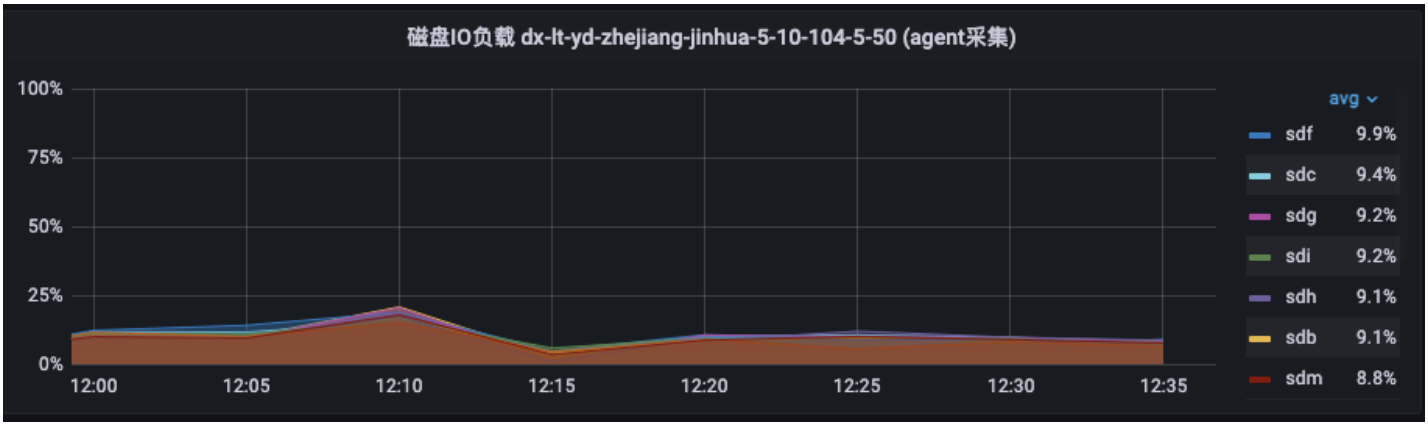
删除RPS在 3000/s 左右，无Slow日志、NameNode资源使用不高、磁盘IO率不高

增加并发量后RPS无法增加，数据与写入操作基本一致，瓶颈应该在与NameNode写文件时锁机制


DataNode All Slow数量
1 个/分
0.500 个/分
0 个/分
-0.50 个/分
-1 个/分
12:00
12:05
12:10
12:15
12:20
max
avg
current
dx-lt-yd-zhejiang-jinhua-5-10-104-5-50 0 个/分 0 个/分 0 个/分
dx-lt-yd-zhejiang-jinhua-5-10-104-5-41 0 个/分 0 个/分 0 个/分
dx-lt-yd-zhejiang-jinhua-5-10-104-5-37 0 个/分 0 个/分 0 个/分


1分钟负载 dx-lt-yd-zhejiang-jinhua-5-10-104-4-41 (agent采集)
5
4
3
2
1
0
12:00
12:05
12:10
12:15
12:20
12:25
12:30
12:35
load 1min


CPU dx-lt-yd-zhejiang-jinhua-5-10-104-4-41 (agent采集)
100%
75%
50%
25%
0%
12:00
12:05
12:10
12:15
12:20
12:25
12:30
12:35


磁盘IO负载 dx-lt-yd-zhejiang-jinhua-5-10-104-5-50 (agent采集)
100%
75%
50%
25%
0%
12:00
12:05
12:10
12:15
12:20
12:25
12:30
12:35
avg
sdf 9.9%
sdc 9.4%
sdg 9.2%
sdi 9.2%
sdh 9.1%
sdb 9.1%
sdm 8.8%