


Answers (/community/s/group/0F90L0000001NDBSA2/)

— Tejas Rao (/community/s/profile/0050L0000093CPGQA2) (BROOKHAVEN NATIONAL LAB) asked a question.

August 23, 2013 at 5:13 AM (/community/s/question/0D50L00006BlugwSAD/performance-with-24-drives)



Performance with 24 drives

I am trying to test performance using NFS and hadoop API.

The server is a pretty beefy server with 24 cores, 96GB of memory. Attached storage is 24 drives (JBOD). Each drive can do about 150MB/sec , these are 4TB enterprise grade drives. I tried concurrent reads on all 24 drives at once using dd's and I get close to 3GB/sec in aggregate.

I am trying to read 1Ge files from the MaprFs NFS filesystem. The NFS filesystem is mounted on localhost.

All chunks are local to that server, so I am not using the network at all.

```
<code>

[root@~]# for i in {1..10} ; do dd if=/mapr/test/test$i.zip iflag=direct of=/dev/null bs=1024k count=1024& done

[root@ ~]# 1000+0 records in

1000+0 records out

1048576000 bytes (1.0 GB) copied, 17.9504 s, 58.4 MB/s

1000+0 records in

1000+0 records out

1048576000 bytes (1.0 GB) copied, 18.0461 s, 58.1 MB/s

1000+0 records in

1000+0 records out

1048576000 bytes (1.0 GB) copied, 18.1585 s, 57.7 MB/s

1000+0 records in

1000+0 records out

1048576000 bytes (1.0 GB) copied, 18.1727 s, 57.7 MB/s

1000+0 records in

1000+0 records out

1048576000 bytes (1.0 GB) copied, 18.1887 s, 57.6 MB/s

1000+0 records in

1000+0 records out

1048576000 bytes (1.0 GB) copied, 18.2644 s, 57.4 MB/s

1000+0 records in

1000+0 records out

1048576000 bytes (1.0 GB) copied, 18.3378 s, 57.2 MB/s

1000+0 records in

1000+0 records out

1048576000 bytes (1.0 GB) copied, 18.3362 s, 57.2 MB/s

1000+0 records in

1000+0 records out
```

1/15/2019

Performance with 24 drives

1048576000 bytes (1.0 GB) copied, 18.3689 s, 57.1 MB/s

1000+0 records in

1000+0 records out

1048576000 bytes (1.0 GB) copied, 18.3928 s, 57.0 MB/s

</code>

As you can see the performance is very disappointing. I am getting ~ 570 MB/sec in aggregate from 24 drives.

I can see 8 storage pools online with 3 drives in each storage pool.

I am using the default chunksize and trying to avoid compression by using the .zip suffix on each file. How can I improve this performance and get performance close to hardware speeds.

Thanks.

^ Upvote

Answer

Share

35 answers47 views

- Tejas Rao (/community/s/profile/0050L0000093CPGQA2) (BROOKHAVEN NATIONAL LAB)

5 years ago

I am trying to test performance using NFS and hadoop API.

The server is a pretty beefy server with 24 cores, 96GB of memory. Attached storage is 24 drives (JBOD). Each drive can do about 150MB/sec , these are 4TB enterprise grade drives. I tried concurrent reads on all 24 drives at once using dd's and I get close to 3GB/sec in aggregate.

Show More

UpvoteReply
- Tejas Rao (/community/s/profile/0050L0000093CPGQA2) (BROOKHAVEN NATIONAL LAB)

5 years ago

Using the hadoop API,I get ~920MB/sec. These number are pretty disappointing.

Even using the API, I see less than 1/3 of hardware speeds.

I am using RHEL6.4 x64 and the latest version of MapR(2.1.3.20987.GA (http://2.1.3.20987.GA)).

UpvoteReply
- Ted Dunning (/community/s/profile/005E0000000L3z6IAC) (MapR)

5 years ago

When you test with the Hadoop API or with NFS, are you using a single thread to write?

(btw... nice job documenting the hardware and speeds)

UpvoteReply
- Tejas Rao (/community/s/profile/0050L0000093CPGQA2) (BROOKHAVEN NATIONAL LAB)

5 years ago

The above tests are all reads. I am trying to read with 10 concurrent streams. i.e reading 10 files.

Reading using 20 or 30 streams seems to make little difference. Are there any tunables to improve performance? We are looking to use mapr as a fileservers/nfs filesystem and don't care about mapreduce.


UpvoteReply
- Ted Dunning (/community/s/profile/005E0000000L3z6IAC) (MapR)

5 years ago

You might be able to get better performance if you rearrange the drives into 3 storage pools of 8 drives. This gives the file system more degrees of freedom for scheduling.

But you are right that this is much lower than expected performance. Let me walk this around.


UpvoteReply

- 

[Tejas Rao \(/community/s/profile/0050L0000093CPGQA2\)](/community/s/profile/0050L0000093CPGQA2) [\(BROOKHAVEN NATIONAL LAB\)](#)

5 years ago


I am concerned about having more than 3-4 drives in a storage pool. With 4TB drives, when a drive fails, re-replication will take a long time and affect production traffic during that time.

Upvote   Reply
- 

[Tejas Rao \(/community/s/profile/0050L0000093CPGQA2\)](/community/s/profile/0050L0000093CPGQA2) [\(BROOKHAVEN NATIONAL LAB\)](#)

5 years ago

I tried playing with the Linux I/O scheduler, read ahead, queue depth etc and it seemed to make little difference.

Upvote   Reply
- 


[Ted Dunning \(/community/s/profile/005E0000000L3z6IAC\)](/community/s/profile/005E0000000L3z6IAC) [\(MapR\)](#)

5 years ago

Those will have little impact since MapR FS schedules the heads independently of the Linux I/O system. It is possible to adjust stripe size and such, but it is rare for this to matter except to make things worse.

There may also be some disk controller settings that are handled a bit different between Linux and MapR.

John B will be back on-line before long and he should have some good input.

Upvote   Reply
- 


[Aaron Eng \(/community/s/profile/005E0000000I1AGIA0\)](/community/s/profile/005E0000000I1AGIA0) [\(MapR Technologies\)](#)

5 years ago

First step, isolate whether this is related to the NFS service or the MFS service. When you mount NFS, the NFS gateway service just relays operations through to the MFS service which in turn does the disk access. So we need to repeat your test without going through the NFS gateway. If we get expected performance then the bottleneck is in the NFS service, if we still see it slow then the bottleneck is in the MFS service.

So instead of doing concurrent dd, launch concurrent "hadoop fs -cat file > /dev/null" commands, one for each file, and see how long it takes.

[Show More](#)

Upvote   Reply
- 

[Ted Dunning \(/community/s/profile/005E0000000L3z6IAC\)](/community/s/profile/005E0000000L3z6IAC) [\(MapR\)](#)

5 years ago

That is a concern. You can throttle replication traffic with the most recent release in order to moderate that impact.

Upvote   Reply

More answers

10 of 35



Write an answer...