Answers (/community/s/group/0F90L0000001NDBSA2/) — dimamah _ (/community/s/profile/0050L00000939jvQAA) (Customer) asked a question.
January 1, 2014 at 9:20 PM (/community/s/question/0D50L00006BIuS2SAL/investigating-rwspeedtest-results)

investigating RWSpeedTest results

Hi,

I ran the test on my cluster and got suspicious results. [root@cluster hadoop-0.20.2]# for i in `seq 5`; do bin/hadoop jar lib/maprfs-diagnostic-tools-1.0.3-mapr-3.0.1.jar com.mapr.fs.RWSpeedTest /t/y${i} 10000 maprfs:///; done; Write rate: 751.9908204245553 M/s Write rate: 818.0496177250741 M/s Write rate: 788.7588385737927 M/s Write rate: 797.5174931548524 M/s Write rate: 756.3300634737449 M/s [root@cluster hadoop-0.20.2]# for i in `seq 5`; do bin/hadoop jar lib/maprfs-diagnostic-tools-1.0.3-mapr-3.0.1.jar com.mapr.fs.RWSpeedTest /t/y${i} -10000 maprfs:///; done; Read rate: 183.70931006692598 M/s Read rate: 179.8339850449338 M/s Read rate: 171.02575394301186 M/s Read rate: 185.99357894176705 M/s Read rate: 179.8339850449338 M/s
While running dd (`dd if=/dev/sd$i bs=4M count=1000 of=/dev/null`) from all the disks simultaneously results in :
4194304000 bytes (4.2 GB) copied, 23.9953 s, 175 MB/s 4194304000 bytes (4.2 GB) copied, 21.8476 s, 192 MB/s 4194304000 bytes (4.2 GB) copied, 22.2197 s, 189 MB/s 4194304000 bytes (4.2 GB) copied, 22.6475 s, 185 MB/s 4194304000 bytes (4.2 GB) copied, 23.028 s, 182 MB/s 4194304000 bytes (4.2 GB) copied, 23.1582 s, 181 MB/s 4194304000 bytes (4.2 GB) copied, 23.6283 s, 178 MB/s 4194304000 bytes (4.2 GB) copied, 23.8413 s, 176 MB/s 4194304000 bytes (4.2 GB) copied, 23.9408 s, 175 MB/s 4194304000 bytes (4.2 GB) copied, 24.3096 s, 173 MB/s
Can you please elaborate on what the RWSpeedTest does exactly?

Does it do IO on 1 disk only?

Also, how is it that the Writes are much faster than the reads? Is it writing to the controller's cache only?

The configuration is :

- Controller : PERC H700 Integrated Firmware 12.10.1-0001
- 10 Sata3 6gb disks
- No raid.

Thanks,

Dima.

Upvote Answer Share 6 answers 30 views

Ted Dunning (/community/s/profile/005E0000000L3z6IAC) (MapR)
5 years ago

What happens when you do a larger test than fits in memory? Try 100G instead of 10G.

Upvote Reply

dimamah _ (/community/s/profile/0050L00000939jvQAA) (Customer)
5 years ago

Same thing.

for i in `seq 5`; do bin/hadoop jar lib/maprfs-diagnostic-tools-1.0.3-mapr-3.0.1.jar com.mapr.fs.RWSpeedTest /t/y${i} 100000 maprfs:///; done;

Write rate: 784.0875781413573 M/s

Write rate: 799.7254360609227 M/s

Write rate: 802.5102937349803 M/s

Write rate: 765.132620672042 M/s

Write rate: 736.4681589279318 M/s

for i in `seq 5`; do bin/hadoop jar lib/maprfs-diagnostic-tools-1.0.3-mapr-3.0.1.jar com.mapr.fs.RWSpeedTest /t/y${i} -100000 maprfs:///; done;

Read rate: 200.28000878608742 M/s

Read rate: 199.47988699812615 M/s

Read rate: 200.40977340142504 M/s

Read rate: 198.84475493709868 M/s

Read rate: 197.32775297899087 M/s

The parallel `dd`'s :

107374182400 bytes (107 GB) copied, 564.069 s, 190 MB/s

107374182400 bytes (107 GB) copied, 575.983 s, 186 MB/s

107374182400 bytes (107 GB) copied, 584.314 s, 184 MB/s

107374182400 bytes (107 GB) copied, 593.428 s, 181 MB/s

107374182400 bytes (107 GB) copied, 597.763 s, 180 MB/s

107374182400 bytes (107 GB) copied, 607.329 s, 177 MB/s

107374182400 bytes (107 GB) copied, 614.917 s, 175 MB/s

107374182400 bytes (107 GB) copied, 616.022 s, 174 MB/s

107374182400 bytes (107 GB) copied, 618.106 s, 174 MB/s

107374182400 bytes (107 GB) copied, 624.126 s, 172 MB/s

Upvote Reply

Aaron Eng (/community/s/profile/005E0000000I1AGIA0) (MapR Technologies)
5 years ago

Hi Dima,

Can you run multiple RWSpeedTest processes in parallel? That test is single threaded. If you want to push load, run multiple in parallel. I'd suggest running at least 8 RWSpeedTests concurrently.

Upvote Reply

MC Srivas (/community/s/profile/0050L0000093EJHQA2) (Customer)
5 years ago

Looks like the file system is compressing the data. Can you turn off compression in the /t directory via "hadoop mfs -setcompression off /t" and rerun the tests?

Upvote Reply

dimamah _ (/community/s/profile/0050L00000939jvQAA) (Customer)
5 years ago

MC Srivas, the compression was turned of from the begging :

`vrwxr-xr-x U - hadoop hadoop 5 2014-01-02 14:27 268435456 /t`

and files inside :

`-rwxrwxrwx U 1 root root 104857648280 2014-01-03 11:47 268435456 /t/y1

-rwxrwxrwx U 1 root root 104857648280 2014-01-03 11:49 268435456 /t/y2

-rwxrwxrwx U 1 root root 104857648280 2014-01-03 11:51 268435456 /t/y3

-rwxrwxrwx U 1 root root 104857648280 2014-01-03 11:54 268435456 /t/y4

-rwxrwxrwx U 1 root root 104857648280 2014-01-03 11:56 268435456 /t/y5

`

Aaron, running in parallel (from 1 server) leads to high CPU usages of the mfs process (~180%).

The results for 8 parallel writes of 50gb each :

Write rate: 117.31268809505164 M/s

Write rate: 110.7155602952201 M/s

Write rate: 110.20010089120565 M/s

Write rate: 108.8603967899671 M/s

Write rate: 108.06830947448701 M/s

Write rate: 104.95941062830946 M/s

Write rate: 104.45440599286752 M/s

Write rate: 104.48875469592339 M/s

Reads :

Read rate: 95.4630784497592 M/s

Read rate: 90.55295042185894 M/s

Read rate: 88.51431658562939 M/s

Read rate: 85.44737442550279 M/s

Read rate: 84.11892117380911 M/s

Read rate: 83.054209042233 M/s

Read rate: 83.3347373775104 M/s

Read rate: 81.7532999118984 M/s

This is very strange to my understanding.

If previously a single process could throughput 800mbs in write it surely means it was writing simultaneously to more than 1 disk and this aligns with these results that the throughput stayed ~800mbs.

As for the reads, we are getting ~700mbs which is still less than the writes (why?) and also this is much faster than a single process.

Is it possible that reading is done from a single disk per execution? otherwise this makes no sense to me and hopefully you can help.

My goal is to get MFS IO benchmark and check my configuration across clusters.

Upvote Reply

Aaron Eng (/community/s/profile/005E0000000I1AGIA0) (MapR Technologies)
5 years ago

When I run RWSpeedTest for writing, I see that the client process ends up generating multiple concurrent outstanding write RPCs to the MFS service. When I run the test for read, I see for the most part that the client process has just a single outstanding read RPC to MFS. Parallelism achieved with multiple client processes with the read test seems to give better performance. I think this is mostly an artifact of the design of the RWSpeedTest code.

As for why you see higher throughput for write at 800MB/s vs. read at 700MB/s, the testing mechanism of launching multiple concurrent RWSpeedTest processes is not precise. Its not as if each process begins to read or write at precisely the same time and then stops reading/writing at precisely the same time, thus summing up the rates shown by each process is not precise. The RWSpeedTest program is not a precision performance analysis tool.

That aside, if we presume write throughput is significantly higher than read throughput, I would assume the culprit is either in some sort of read ahead code or in the actual nature of your disks responding more quickly to our write requests than read requests.

As a simple measure, while you have the RWSpeedTest write running enable MFS debug logging by running:

<code>

maprcli trace setlevel -level debug

maprcli trace setmode -mode continuous

sleep 30

maprcli trace setlevel -level default

maprcli trace setmode -mode default

</code>

Then copy /opt/mapr/logs/mfs* from the node to a backup location. Then repeat with the RWSpeedTest read running, then send all these mfs log files for analysis and I can review them.

Upvote Reply

Write an answer...