Linux network storage tests

02.07.22

Pegasi Knowledge
https://ghost.pegasi.fi/wiki/

Table of Contents

Linux network storage tests

We have two 24 x NVME storage backends and multiple frontends. I am testing different ways to bring storage to a frontend server and might as well display the results here more publicly.

I used fio and hdparm and I have tested the following configurations so far:

- local drives for comparison
  - frontend server local SATA SSD RAID
  - storage server local NVME
- NVME-of, Infiniband 56/40 Gbps
  - NVME-of
  - striped LVM on top of RAID 1 MD from both storage backends, built on top of NVME-of
- iSER RDMA enabled iSCSI, Infiniband 56/40 Gbps
  - iSER
  - striped LVM on top of RAID 1 MD from both storage backends, built on top of iSER
- iSER RDMA enabled iSCSI, using drbd replicated device on the storage server, Infiniband 56/40 Gbps
  - iSER
  - DRBD backend
  - striped LVM on top of iSCSI devices

I am using an RDMA Infiniband 56/40Gbps network with ConnectX-3 cards. Each server has a single active connection. I could possibly speed things up by using a different connection per storage server but that may not be required.

For testing I am using two commands:

- hdparm -Tt <device>
- fio –name=random-write –ioengine=posixaio –rw=randwrite –bs=4k –numjobs=1 –size=4g –iodepth=1 –runtime=60 –time*based* –*end*fsync=1

Frontend server local SATA SSD

Supermicro MegaRAID based 2 SSD disk raid as an operating system disk.

```
hdparm -Tt /dev/sda:
 Timing cached reads:   21894 MB in  1.99 seconds = 11005.29 MB/sec
 Timing buffered disk reads: 3186 MB in  3.00 seconds = 1061.92 MB/sec

random-write: Laying out IO file (1 file / 4096MiB)
Jobs: 1 (f=1): [w(1)][100.0%][eta 00m:00s]
random-write: (groupid=0, jobs=1): err= 0: pid=3542: Mon Jun 28 18:39:00
2021
```

```
  write: IOPS=46.5k, BW=182MiB/s (191MB/s)(12.0GiB/67589msec); 0 zone resets
    slat (nsec): min=879, max=127715, avg=1779.53, stdev=650.57
    clat (nsec): min=202, max=1674.4k, avg=8290.97, stdev=2533.87
     lat (usec): min=7, max=1676, avg=10.07, stdev= 2.73
    clat percentiles (nsec):
     |  1.00th=[ 6688],  5.00th=[ 6944], 10.00th=[ 7136], 20.00th=[ 7328],
     | 30.00th=[ 7456], 40.00th=[ 7520], 50.00th=[ 7648], 60.00th=[ 7712],
     | 70.00th=[ 7904], 80.00th=[ 8512], 90.00th=[10816], 95.00th=[11456],
     | 99.00th=[16192], 99.50th=[24192], 99.90th=[28032], 99.95th=[36096],
     | 99.99th=[50432]
   bw (  KiB/s): min=78840, max=418144, per=100.00%, avg=363084.28,
stdev=77272.87, samples=69
   iops        : min=19710, max=104536, avg=90771.12, stdev=19318.21,
samples=69
  lat (nsec)   : 250=0.01%
  lat (usec)   : 10=85.06%, 20=13.99%, 50=0.94%, 100=0.01%, 250=0.01%
  lat (usec)   : 750=0.01%, 1000=0.01%
  lat (msec)   : 2=0.01%
  cpu          : usr=15.36%, sys=14.12%, ctx=3195292, majf=0, minf=701
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%,
>=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     issued rwts: total=0,3145729,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=182MiB/s (191MB/s), 182MiB/s-182MiB/s (191MB/s-191MB/s),
io=12.0GiB (12.9GB), run=67589-67589msec

Disk stats (read/write):
    dm-0: ios=0/127661, merge=0/0, ticks=0/1201685, in_queue=1201685,
util=50.69%, aggrios=0/161924, aggrmerge=0/12, aggrticks=0/2302922,
aggrin_queue=2302922, aggrutil=56.08%
  sda: ios=0/161924, merge=0/12, ticks=0/2302922, in_queue=2302922,
util=56.08%
```

Storage server local NVME

Intel Corporation SSDPE2KE016T8O PCIe NVME local device.

```
hdparm -Tt /dev/nvme2n1:
 Timing cached reads:   17856 MB in  2.00 seconds = 8943.15 MB/sec
```

```
 Timing buffered disk reads: 6910 MB in  3.00 seconds = 2303.19 MB/sec

random-write: Laying out IO file (1 file / 4096MiB)
Jobs: 1 (f=1): [w(1)][100.0%][w=349MiB/s][w=89.2k IOPS][eta 00m:00s]
random-write: (groupid=0, jobs=1): err= 0: pid=3782: Mon Jun 28 18:45:44
2021
  write: IOPS=88.6k, BW=346MiB/s (363MB/s)(20.3GiB/60194msec); 0 zone resets
    slat (nsec): min=580, max=81950, avg=1524.49, stdev=380.18
    clat (nsec): min=220, max=600600, avg=7236.48, stdev=1917.52
     lat (usec): min=4, max=602, avg= 8.76, stdev= 2.18
    clat percentiles (nsec):
     |  1.00th=[ 4320],  5.00th=[ 4576], 10.00th=[ 4768], 20.00th=[ 5344],
     | 30.00th=[ 6368], 40.00th=[ 7520], 50.00th=[ 7776], 60.00th=[ 7904],
     | 70.00th=[ 8096], 80.00th=[ 8256], 90.00th=[ 8768], 95.00th=[ 9152],
     | 99.00th=[11072], 99.50th=[12224], 99.90th=[16768], 99.95th=[19072],
     | 99.99th=[34560]
   bw (  KiB/s): min=20728, max=622960, per=100.00%, avg=411116.83,
stdev=106584.98, samples=103
   iops        : min= 5182, max=155740, avg=102779.18, stdev=26646.25,
samples=103
  lat (nsec)   : 250=0.01%, 750=0.01%
  lat (usec)   : 4=0.01%, 10=97.95%, 20=2.00%, 50=0.04%, 100=0.01%
  lat (usec)   : 250=0.01%, 500=0.01%, 750=0.01%
  cpu          : usr=18.80%, sys=25.26%, ctx=5379783, majf=0, minf=780
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%,
>=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     issued rwts: total=0,5333109,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=346MiB/s (363MB/s), 346MiB/s-346MiB/s (363MB/s-363MB/s),
io=20.3GiB (21.8GB), run=60194-60194msec

Disk stats (read/write):
  nvme0n1: ios=1/1405100, merge=0/23, ticks=0/2932174, in_queue=2932174,
util=24.13%
```

Frontend NVME-of bind using backend NVME

The previous Intel NVME exported with NVME-of, as seen in one front end server.

```
hdparm -Tt /dev/nvme0n3:
 Timing cached reads:   21334 MB in  1.99 seconds = 10722.82 MB/sec
 Timing buffered disk reads: 5940 MB in  3.00 seconds = 1979.68 MB/sec

random-write: Laying out IO file (1 file / 4096MiB)
Jobs: 1 (f=1): [F(1)][100.0%][w=4839KiB/s][w=1209 IOPS][eta 00m:00s]
random-write: (groupid=0, jobs=1): err= 0: pid=3456: Mon Jun 28 18:37:08
2021
  write: IOPS=77.8k, BW=304MiB/s (319MB/s)(18.2GiB/61171msec); 0 zone resets
    slat (nsec): min=860, max=110739, avg=1750.52, stdev=316.16
    clat (usec): min=4, max=576, avg= 8.12, stdev= 2.13
     lat (usec): min=7, max=578, avg= 9.87, stdev= 2.26
    clat percentiles (nsec):
     |  1.00th=[ 6624],  5.00th=[ 6944], 10.00th=[ 7072], 20.00th=[ 7200],
     | 30.00th=[ 7328], 40.00th=[ 7392], 50.00th=[ 7456], 60.00th=[ 7584],
     | 70.00th=[ 7712], 80.00th=[ 8384], 90.00th=[10688], 95.00th=[11328],
     | 99.00th=[13632], 99.50th=[17280], 99.90th=[29824], 99.95th=[44288],
     | 99.99th=[61696]
   bw (  KiB/s): min=15976, max=430968, per=100.00%, avg=370133.55,
stdev=90692.00, samples=102
   iops        : min= 3994, max=107742, avg=92533.38, stdev=22673.00,
samples=102
  lat (usec)   : 10=85.87%, 20=13.78%, 50=0.32%, 100=0.02%, 250=0.01%
  lat (usec)   : 500=0.01%, 750=0.01%
  cpu          : usr=23.38%, sys=20.65%, ctx=4824959, majf=0, minf=526
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%,
>=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     issued rwts: total=0,4762087,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=304MiB/s (319MB/s), 304MiB/s-304MiB/s (319MB/s-319MB/s),
io=18.2GiB (19.5GB), run=61171-61171msec

Disk stats (read/write):
  nvme0n1: ios=1/738871, merge=0/28, ticks=0/279727, in_queue=279727,
util=25.48%
```

Frontend striped LVM - RAID1 - NVME-of

- Six NVME-of drives, 3 from each server

- Three RAID1 devices, each consisting of 2 drives from 2 different storage servers
- One striped LVM volume group consisting of the three RAID1 MD drives
- One logical volume from the volume group

```
hdparm -Tt /dev/datavault/testvolume:
 Timing cached reads:   21458 MB in  1.99 seconds = 10784.51 MB/sec
 Timing buffered disk reads: 7640 MB in  3.00 seconds = 2546.47 MB/sec

random-write: Laying out IO file (1 file / 4096MiB)
Jobs: 1 (f=1): [F(1)][100.0%][w=2029KiB/s][w=507 IOPS][eta 00m:00s]
random-write: (groupid=0, jobs=1): err= 0: pid=4526: Tue Jun 29 08:05:20
2021
  write: IOPS=76.2k, BW=298MiB/s (312MB/s)(17.8GiB/61302msec); 0 zone resets
    slat (nsec): min=802, max=142478, avg=1831.73, stdev=521.29
    clat (usec): min=5, max=3873, avg= 8.61, stdev= 6.04
     lat (usec): min=7, max=3874, avg=10.44, stdev= 6.13
    clat percentiles (nsec):
     |  1.00th=[ 6624],  5.00th=[ 6880], 10.00th=[ 7072], 20.00th=[ 7264],
     | 30.00th=[ 7392], 40.00th=[ 7520], 50.00th=[ 7648], 60.00th=[ 7840],
     | 70.00th=[ 8768], 80.00th=[10432], 90.00th=[11072], 95.00th=[11584],
     | 99.00th=[14016], 99.50th=[22656], 99.90th=[29312], 99.95th=[43264],
     | 99.99th=[52480]
   bw (  KiB/s): min=  838, max=438626, per=100.00%, avg=349216.78,
stdev=92392.10, samples=106
   iops        : min=  209, max=109656, avg=87303.93, stdev=23097.92,
samples=106
  lat (usec)   : 10=74.54%, 20=24.87%, 50=0.57%, 100=0.01%, 250=0.01%
  lat (usec)   : 500=0.01%, 750=0.01%, 1000=0.01%
  lat (msec)   : 2=0.01%, 4=0.01%
  cpu          : usr=23.60%, sys=22.23%, ctx=4868012, majf=0, minf=979
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%,
>=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     issued rwts: total=0,4674124,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=298MiB/s (312MB/s), 298MiB/s-298MiB/s (312MB/s-312MB/s),
io=17.8GiB (19.1GB), run=61302-61302msec

Disk stats (read/write):
    dm-3: ios=1/891284, merge=0/0, ticks=0/694354549, in_queue=694354549,
util=25.73%, aggrios=1/904069, aggrmerge=0/0, aggrticks=0/0, aggrin_queue=0,
aggrutil=0.00%
```

```
    md0: ios=1/904069, merge=0/0, ticks=0/0, in_queue=0, util=0.00%,
aggrios=0/903755, aggrmerge=0/188, aggrticks=0/333576, aggrin_queue=333577,
aggrutil=23.45%
  nvme0n1: ios=1/903756, merge=0/188, ticks=1/338804, in_queue=338805,
util=23.45%
  nvme1n1: ios=0/903755, merge=0/189, ticks=0/328349, in_queue=328349,
util=23.27%
```

Frontend iSER single drive

I set up iSER target to the storage backend, iSER initiator to the frontend and created an XFS filesystem. Exactly as with NVME-of cases. Looks slightly less in writing speed but very solid latency.

```
hdparm -Tt /dev/disk/by-path/ip-xxx-lun-1:

 Timing cached reads:   21130 MB in  1.99 seconds = 10618.36 MB/sec
 Timing buffered disk reads: 3704 MB in  3.00 seconds = 1234.42 MB/sec


random-write: Laying out IO file (1 file / 4096MiB)
Jobs: 1 (f=1): [F(1)][100.0%][eta 00m:00s]
random-write: (groupid=0, jobs=1): err= 0: pid=3116: Tue Jun 29 15:52:07
2021
  write: IOPS=70.8k, BW=277MiB/s (290MB/s)(17.3GiB/63904msec); 0 zone resets
    slat (nsec): min=836, max=324487, avg=1694.97, stdev=519.52
    clat (nsec): min=267, max=938988, avg=7613.02, stdev=1547.30
     lat (usec): min=6, max=940, avg= 9.31, stdev= 1.67
    clat percentiles (nsec):
     |  1.00th=[ 6688],  5.00th=[ 6880], 10.00th=[ 7008], 20.00th=[ 7136],
     | 30.00th=[ 7264], 40.00th=[ 7328], 50.00th=[ 7392], 60.00th=[ 7520],
     | 70.00th=[ 7584], 80.00th=[ 7776], 90.00th=[ 8256], 95.00th=[ 8512],
     | 99.00th=[10432], 99.50th=[22400], 99.90th=[25728], 99.95th=[26752],
     | 99.99th=[29568]
   bw (  KiB/s): min=33408, max=423456, per=100.00%, avg=394176.32,
stdev=72408.07, samples=91
   iops        : min= 8352, max=105864, avg=98544.10, stdev=18102.02,
samples=91
  lat (nsec)   : 500=0.01%
  lat (usec)   : 4=0.01%, 10=98.87%, 20=0.55%, 50=0.58%, 100=0.01%
  lat (usec)   : 250=0.01%, 750=0.01%, 1000=0.01%
  cpu          : usr=21.22%, sys=20.55%, ctx=4836908, majf=0, minf=893
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%,
>=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
```

```
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     issued rwts: total=0,4526608,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=277MiB/s (290MB/s), 277MiB/s-277MiB/s (290MB/s-290MB/s),
io=17.3GiB (18.5GB), run=63904-63904msec

Disk stats (read/write):
  sdd: ios=1/620934, merge=0/5878, ticks=1/224797, in_queue=224798,
util=34.91%
```

Frontend striped LVM - RAID1 - iSER

Similar setup to NVME-of but using iSER instead.

- Six iSER mounted drives, 3 from each server
- Three RAID1 devices, each consisting of 2 drives from 2 different storage servers
- One striped LVM volume group consisting of the three RAID1 MD drives
- One logical volume from the volume group

```
hdparm -Tt /dev/datavault/testvolume:
 Timing cached reads:   21368 MB in  1.99 seconds = 10739.75 MB/sec
 Timing buffered disk reads: 3898 MB in  3.00 seconds = 1298.67 MB/sec

random-write: Laying out IO file (1 file / 4096MiB)
Jobs: 1 (f=1): [w(1)][100.0%][w=74.6MiB/s][w=19.1k IOPS][eta 00m:00s]
random-write: (groupid=0, jobs=1): err= 0: pid=10059: Mon Jul  5 12:01:40
2021
  write: IOPS=69.7k, BW=272MiB/s (285MB/s)(16.1GiB/60525msec); 0 zone resets
    slat (nsec): min=839, max=261971, avg=1676.40, stdev=620.25
    clat (nsec): min=205, max=1344.2k, avg=8360.25, stdev=8926.87
     lat (usec): min=7, max=1345, avg=10.04, stdev= 8.97
    clat percentiles (usec):
     |  1.00th=[    7],  5.00th=[    7], 10.00th=[    8], 20.00th=[    8],
     | 30.00th=[    8], 40.00th=[    8], 50.00th=[    8], 60.00th=[    8],
     | 70.00th=[    8], 80.00th=[    8], 90.00th=[    9], 95.00th=[   10],
     | 99.00th=[   25], 99.50th=[   35], 99.90th=[  155], 99.95th=[  186],
     | 99.99th=[  273]
   bw (  KiB/s): min=21752, max=424928, per=100.00%, avg=367624.85,
stdev=96931.23, samples=91
   iops        : min= 5440, max=106232, avg=91906.21, stdev=24232.72,
samples=91
  lat (nsec)   : 250=0.01%, 500=0.01%
```

```
  lat (usec)   : 4=0.01%, 10=95.68%, 20=2.66%, 50=1.27%, 100=0.14%
  lat (usec)   : 250=0.23%, 500=0.01%, 750=0.01%, 1000=0.01%
  lat (msec)   : 2=0.01%
  cpu          : usr=20.85%, sys=21.57%, ctx=4476176, majf=0, minf=871
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%,
>=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     issued rwts: total=0,4215656,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=272MiB/s (285MB/s), 272MiB/s-272MiB/s (285MB/s-285MB/s),
io=16.1GiB (17.3GB), run=60525-60525msec

Disk stats (read/write):
    dm-3: ios=1/548969, merge=0/0, ticks=1/830242240, in_queue=830242241,
util=36.64%, aggrios=1/623957, aggrmerge=0/0, aggrticks=0/0, aggrin_queue=0,
aggrutil=0.00%
    md0: ios=1/623957, merge=0/0, ticks=0/0, in_queue=0, util=0.00%,
aggrios=0/623048, aggrmerge=0/526, aggrticks=0/213754, aggrin_queue=213754,
aggrutil=35.33%
  sdb: ios=1/623078, merge=0/497, ticks=1/211402, in_queue=211402,
util=35.33%
  sdc: ios=0/623019, merge=0/556, ticks=0/216106, in_queue=216106,
util=35.32%
```

Single iSER - DRBD

- One iSER mounted drive from primary server

```
hdparm -Tt /dev/sdb:
 Timing cached reads:   21054 MB in  1.99 seconds = 10581.09 MB/sec
 Timing buffered disk reads: 3908 MB in  3.00 seconds = 1302.55 MB/sec

fio --name=random-write --ioengine=posixaio --rw=randwrite --bs=4k --
numjobs=1 --size=4g --iodepth=1 --runtime=60 --time_based --end_fsync=1
random-write: (g=0): rw=randwrite, bs=(R) 4096B-4096B, (W) 4096B-4096B, (T)
4096B-4096B, ioengine=posixaio, iodepth=1
fio-3.19
Starting 1 process
random-write: Laying out IO file (1 file / 4096MiB)
Jobs: 1 (f=1): [F(1)][100.0%][eta 00m:00s]
```

```
random-write: (groupid=0, jobs=1): err= 0: pid=11449: Tue Jul  6 14:14:06
2021
  write: IOPS=62.5k, BW=244MiB/s (256MB/s)(15.7GiB/65801msec); 0 zone resets
    slat (nsec): min=913, max=134212, avg=1668.50, stdev=498.55
    clat (usec): min=3, max=1588, avg= 7.58, stdev= 2.04
     lat (usec): min=7, max=1590, avg= 9.25, stdev= 2.13
    clat percentiles (nsec):
     |  1.00th=[ 6560],  5.00th=[ 6752], 10.00th=[ 6880], 20.00th=[ 7072],
     | 30.00th=[ 7200], 40.00th=[ 7328], 50.00th=[ 7456], 60.00th=[ 7520],
     | 70.00th=[ 7648], 80.00th=[ 7776], 90.00th=[ 8032], 95.00th=[ 8512],
     | 99.00th=[10176], 99.50th=[22656], 99.90th=[25984], 99.95th=[27008],
     | 99.99th=[36096]
   bw (  KiB/s): min=67536, max=451744, per=100.00%, avg=396963.38,
stdev=70142.99, samples=82
   iops        : min=16884, max=112936, avg=99240.79, stdev=17535.73,
samples=82
  lat (usec)   : 4=0.01%, 10=98.97%, 20=0.42%, 50=0.61%, 100=0.01%
  lat (usec)   : 250=0.01%, 500=0.01%
  lat (msec)   : 2=0.01%
  cpu          : usr=20.62%, sys=16.00%, ctx=4239733, majf=0, minf=807
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%,
>=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     issued rwts: total=0,4111551,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=244MiB/s (256MB/s), 244MiB/s-244MiB/s (256MB/s-256MB/s),
io=15.7GiB (16.8GB), run=65801-65801msec

Disk stats (read/write):
  sdb: ios=1/189398, merge=0/3781, ticks=1/229437, in_queue=229437,
util=38.13%
```

Frontend striped LVM - iSER - DRBD

- Three iSER mounted drives from primary server
- One striped LVM volume group consisting of the three iSER mounted devices
- One logical volume from the volume group

```
hdparm -Tt /dev/datavault/test:
 Timing cached reads:   20898 MB in  1.99 seconds = 10501.52 MB/sec
```

```
 Timing buffered disk reads: 3812 MB in  3.00 seconds = 1270.48 MB/sec

fio --name=random-write --ioengine=posixaio --rw=randwrite --bs=4k --
numjobs=1 --size=4g --iodepth=1 --runtime=60 --time_based --end_fsync=1
random-write: (g=0): rw=randwrite, bs=(R) 4096B-4096B, (W) 4096B-4096B, (T)
4096B-4096B, ioengine=posixaio, iodepth=1
fio-3.19
Starting 1 process
random-write: Laying out IO file (1 file / 4096MiB)
Jobs: 1 (f=1): [F(1)][100.0%][eta 00m:00s]
random-write: (groupid=0, jobs=1): err= 0: pid=11743: Tue Jul  6 14:25:35
2021
  write: IOPS=61.8k, BW=241MiB/s (253MB/s)(15.6GiB/66078msec); 0 zone resets
    slat (nsec): min=815, max=152844, avg=1670.57, stdev=552.98
    clat (usec): min=5, max=164, avg= 7.63, stdev= 1.64
     lat (usec): min=7, max=170, avg= 9.30, stdev= 1.76
    clat percentiles (nsec):
     |  1.00th=[ 6752],  5.00th=[ 6944], 10.00th=[ 7072], 20.00th=[ 7200],
     | 30.00th=[ 7264], 40.00th=[ 7392], 50.00th=[ 7456], 60.00th=[ 7520],
     | 70.00th=[ 7584], 80.00th=[ 7712], 90.00th=[ 8032], 95.00th=[ 8512],
     | 99.00th=[10944], 99.50th=[23168], 99.90th=[26240], 99.95th=[27264],
     | 99.99th=[40704]
   bw (  KiB/s): min= 1912, max=419584, per=100.00%, avg=399272.78,
stdev=66461.85, samples=81
   iops        : min=  478, max=104896, avg=99818.16, stdev=16615.56,
samples=81
  lat (usec)   : 10=98.81%, 20=0.42%, 50=0.76%, 100=0.01%, 250=0.01%
  cpu          : usr=18.56%, sys=16.97%, ctx=4211106, majf=0, minf=763
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%,
>=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%,
>=64=0.0%
     issued rwts: total=0,4084062,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=1

Run status group 0 (all jobs):
  WRITE: bw=241MiB/s (253MB/s), 241MiB/s-241MiB/s (253MB/s-253MB/s),
io=15.6GiB (16.7GB), run=66078-66078msec

Disk stats (read/write):
    dm-3: ios=1/103363, merge=0/0, ticks=0/116893, in_queue=116893,
util=35.07%, aggrios=1/204245, aggrmerge=0/1728, aggrticks=0/235478,
aggrin_queue=235478, aggrutil=38.83%
  sdb: ios=1/204245, merge=0/1728, ticks=0/235478, in_queue=235478,
util=38.83%
```

Conclusions

I still have not studied hdparm / fio analysis and possible alternate analysis methods well enough to make big conclusions. But quick study shows the Infiniband network seems to be quite transparent and performance seems good.

The latency seems to be unchanged on NVME-of compared to local NVME which is incredible. With RAID1 + LVM there is a slight increase but we are talking only a three nanosecond increase from 10 to 13 or so, unless I am reading that fio output wrong.

iSER surprises with it's solid latency but for some reason gains no advantage when using LVM striping.

My only worry would be the CPU usage of the MD RAID1. I may try DRBD backend replication and some solution on the front end to swap block devices on the fly.

So far the best performing options (rough roundings there) of networked storages would be:

- By lowest latency: iSER, 10us single drive, 10.05us striped raid1, beating NVME-of (11us / 12us) by ~10%
- By read speed: NVME-of, 1980MB/s single drive, 2546MB/s striped raid1, beating iSER (1234MB/s / 1299MB/s) by 40-50%
- By write speed: NVME-of, 319MB/s single drive, 312MB/s striped raid1, beating iSER (290MB/s / 285MB/s) by ~10%

Please leave a comment if you have any ideas.

Comments

All comments and corrections are welcome.