Олег

internet-lab.ru/fio_linux

25 мая 2021

Для замера производительности дисковой системы в Linux есть замечательная утилита **fio** (Flexible I/O tester). Утилита позволяет использовать для тестирования заранее написанные файл-тесты, в которых указывается что именно мы хотим измерить.

К примеру, можно провести тест на производительность при последовательном и случайном чтении/записи на диск.

Установка fio

Ubuntu:

```
apt-get install fio
```

CentOS:

```
yum install fio
```

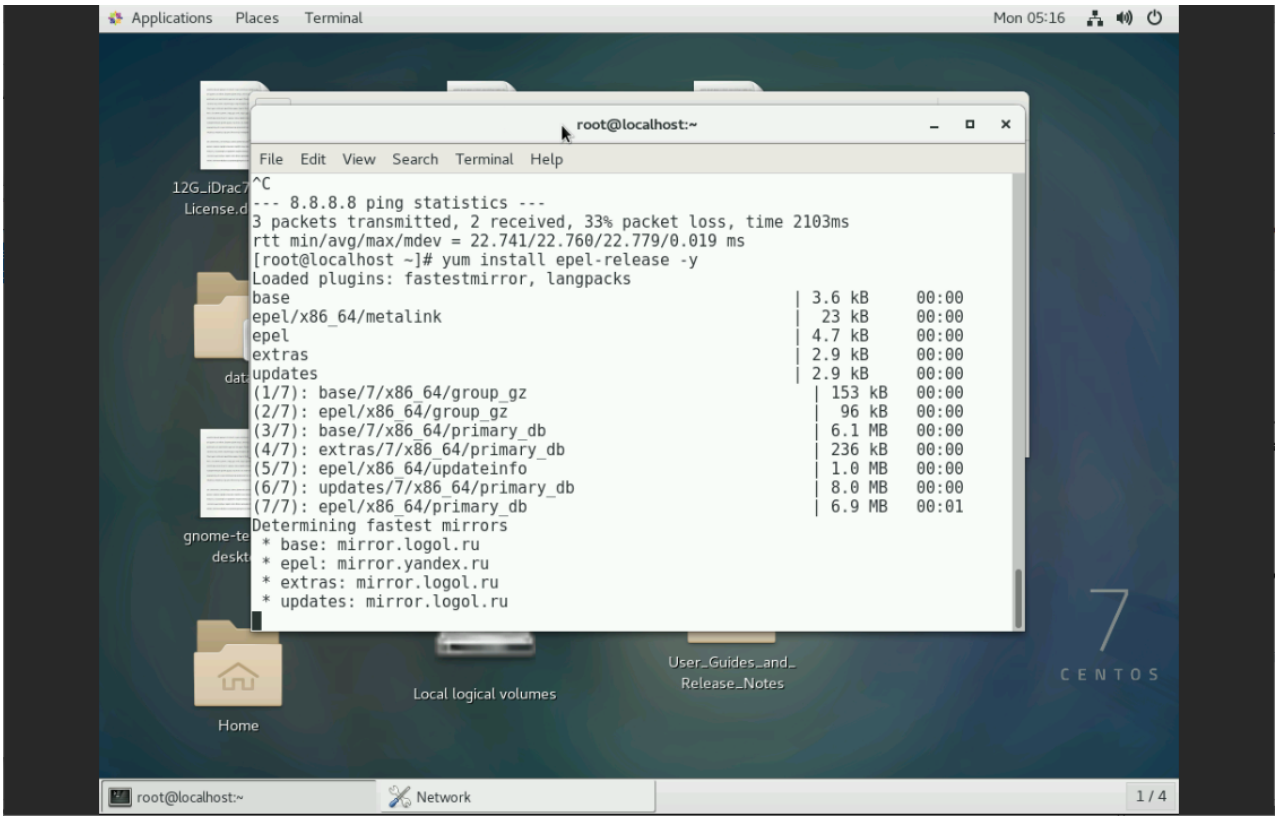
Измеряем производительность NVMe диска с помощью fio

Есть сервер Dell PowerEdge R640 с NVMe дисками SSD Dell EMC NVMe 3.84 TB — KCD5XLUG3T84. В спецификации указаны следующие параметры производительности для данной модели NVMe диска:

- Последовательное чтение: 3140 MBPS
- Последовательная запись: 1520 MBPS
- Случайное чтение: 465000 IOPS
- Случайная запись: 40000 IOPS

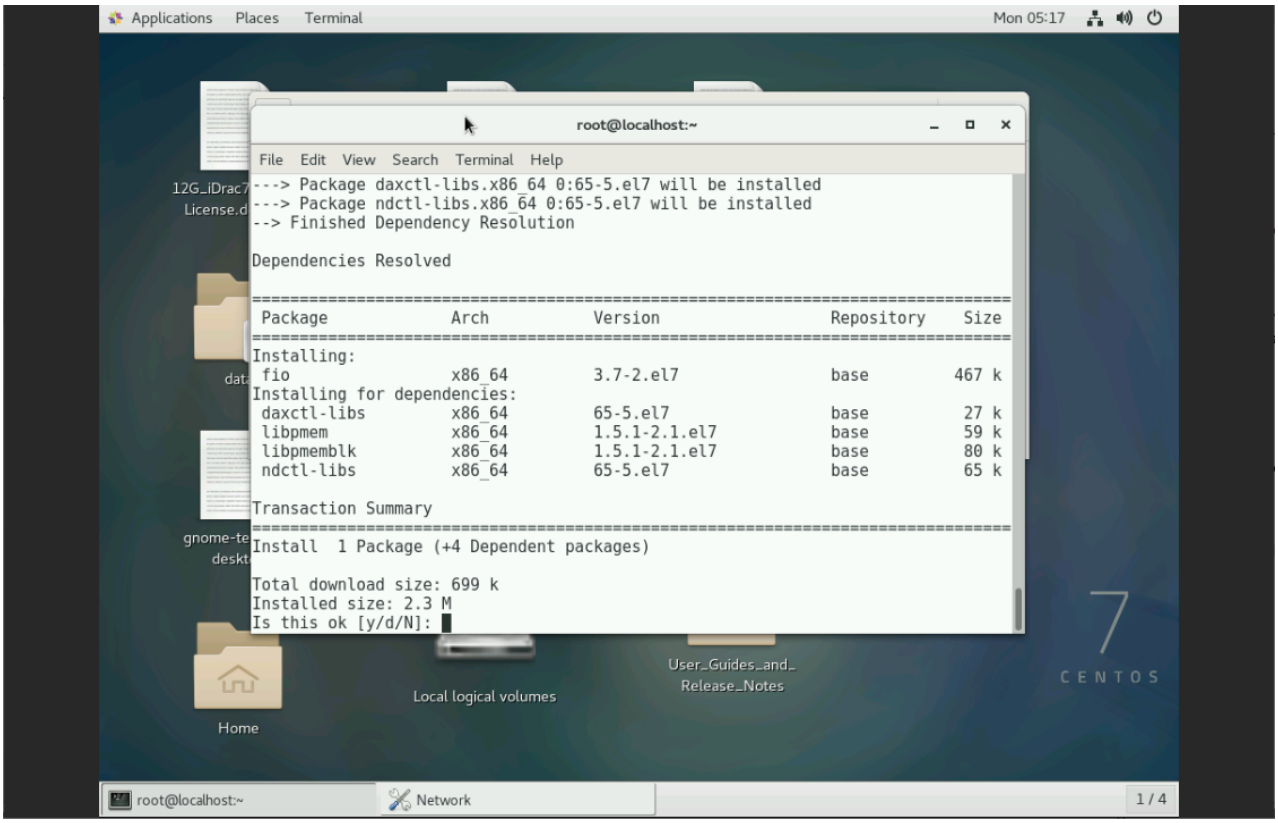
Загружаю на сервере Dell Support Live Image с операционной системой CentOS. Работаю в терминале под рутом. Устанавливаю репозиторий EPEL:

```
yum install epel-release -y
```



Устанавливаю fio:

```
yum install fio -y
```

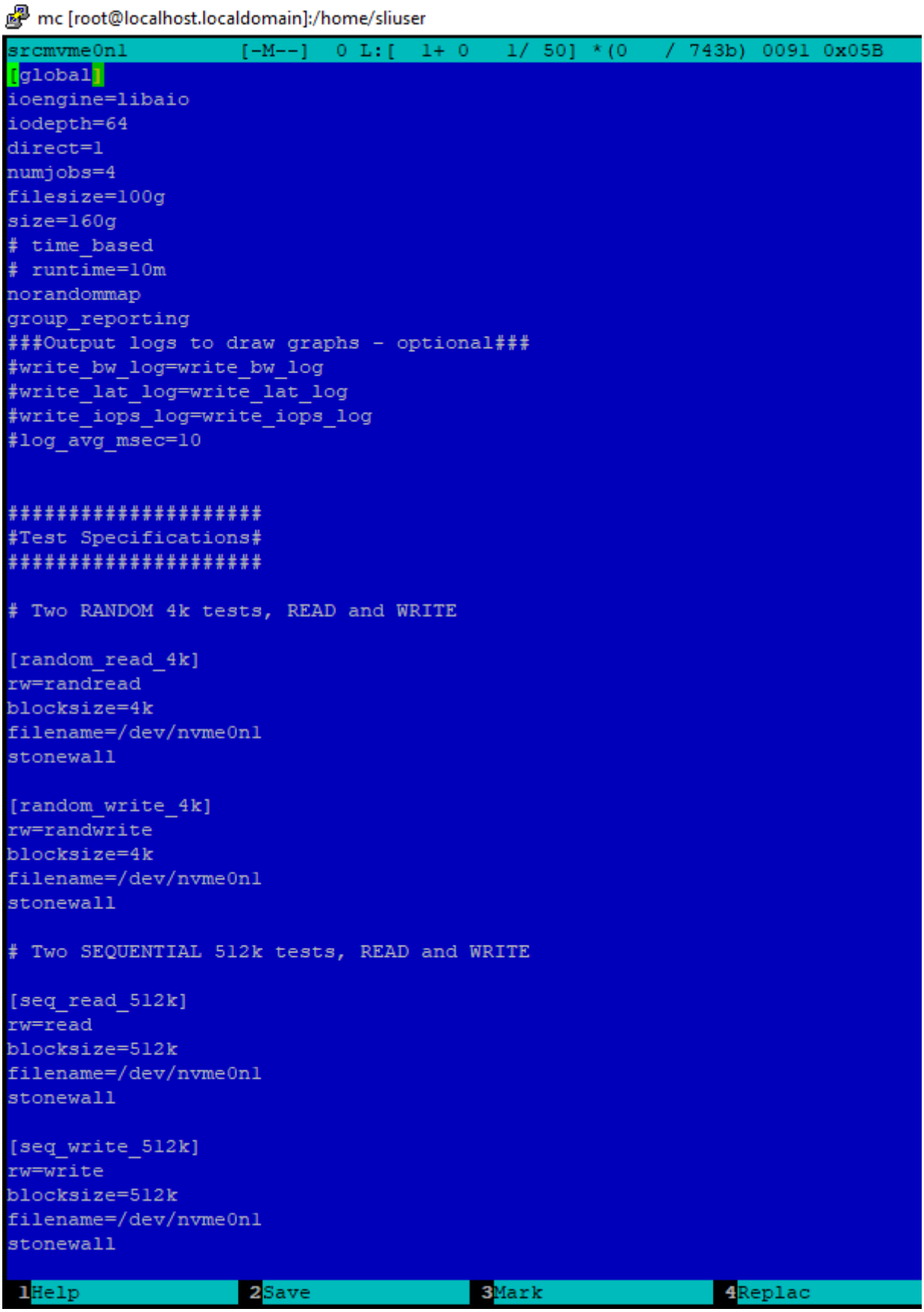


Создаю файл-тест с конфигурацией тестирования, назову его **srcmvme0n1**:

```
[global]
ioengine=libaio
iodepth=64
direct=1
numjobs=4
filesize=100g
size=160g
# time_based
# runtime=10m
norandommap
group_reporting
###Output logs to draw graphs - optional###
#write_bw_log=write_bw_log
#write_lat_log=write_lat_log
#write_iops_log=write_iops_log
#log_avg_msec=10

#####################
#Test Specifications#
#####################

# Two RANDOM 4k tests, READ and WRITE

[random_read_4k]
rw=randread
blocksize=4k
filename=/dev/md127
stonewall

[random_write_4k]
rw=randwrite
blocksize=4k
filename=/dev/md127
stonewall

# Two SEQUENTIAL 512k tests, READ and WRITE

[seq_read_512k]
rw=read
blocksize=512k
filename=/dev/md127
stonewall

[seq_write_512k]
rw=write
blocksize=512k
filename=/dev/md127
stonewall
```

Будем проводить четыре теста:

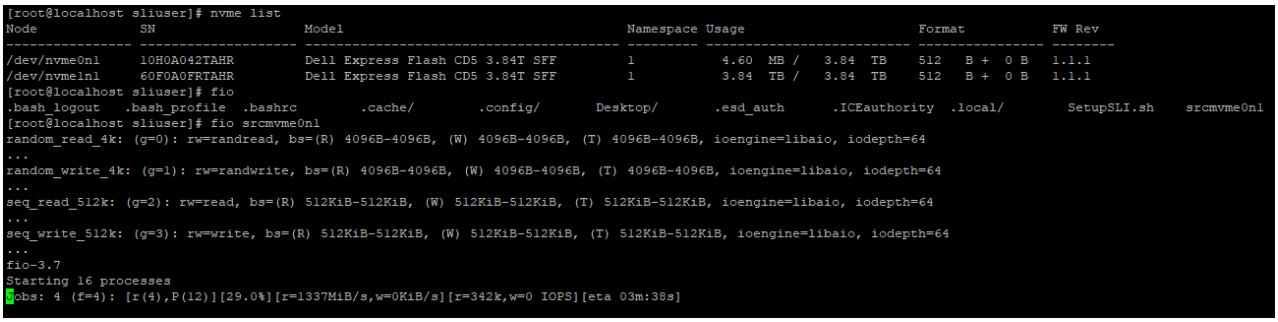
- Случайное чтение блоками 4k
- Случайная запись блоками 4k
- Последовательное чтение блоками 512k
- Последовательная запись блоками 512k



Запуск теста:

```
fio srcmvme0n1
```

Процедура тестирования может занять продолжительное время.



Результаты тестирования:

```
random_read_4k: (g=0): rw=randread, bs=(R) 4096B-4096B, (W) 4096B-4096B, (T)
4096B-4096B, ioengine=libaio, iodepth=64
...
random_write_4k: (g=1): rw=randwrite, bs=(R) 4096B-4096B, (W) 4096B-4096B, (T)
4096B-4096B, ioengine=libaio, iodepth=64
...
seq_read_512k: (g=2): rw=read, bs=(R) 512KiB-512KiB, (W) 512KiB-512KiB, (T)
512KiB-512KiB, ioengine=libaio, iodepth=64
...
seq_write_512k: (g=3): rw=write, bs=(R) 512KiB-512KiB, (W) 512KiB-512KiB, (T)
512KiB-512KiB, ioengine=libaio, iodepth=64
...
fio-3.7
Starting 16 processes
Jobs: 4 (f=4): [_(12),W(4)][99.9%][r=0KiB/s,w=1515MiB/s][r=0,w=3030 IOPS][eta
00m:01s]
random_read_4k: (groupid=0, jobs=4): err= 0: pid=10156: Mon May 17 05:48:24 2021
   read: IOPS=342k, BW=1336MiB/s (1401MB/s)(400GiB/306556msec)
    slat (nsec): min=1327, max=638832, avg=3041.06, stdev=1293.64
    clat (usec): min=14, max=2272, avg=744.70, stdev= 7.07
     lat (usec): min=16, max=2274, avg=747.82, stdev= 6.99
    clat percentiles (usec):
     |  1.00th=[  734],  5.00th=[  742], 10.00th=[  742], 20.00th=[  742],
     | 30.00th=[  742], 40.00th=[  742], 50.00th=[  742], 60.00th=[  750],
     | 70.00th=[  750], 80.00th=[  750], 90.00th=[  750], 95.00th=[  750],
     | 99.00th=[  758], 99.50th=[  758], 99.90th=[  766], 99.95th=[  766],
     | 99.99th=[  783]
   bw (  KiB/s): min=270177, max=342465, per=21.38%, avg=292523.40,
stdev=13487.06, samples=2449
   iops        : min=67544, max=85616, avg=73130.48, stdev=3371.77, samples=2449
  lat (usec)   : 20=0.01%, 50=0.01%, 100=0.01%, 250=0.01%, 500=0.01%
  lat (usec)   : 750=91.49%, 1000=8.49%
  lat (msec)   : 2=0.01%, 4=0.01%
  cpu          : usr=10.97%, sys=31.78%, ctx=34171019, majf=0, minf=3902
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=100.0%
     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=0.1%, >=64=0.0%
     issued rwts: total=104857600,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=64
random_write_4k: (groupid=1, jobs=4): err= 0: pid=10939: Mon May 17 05:48:24 2021
  write: IOPS=278k, BW=1085MiB/s (1138MB/s)(400GiB/377376msec)
    slat (nsec): min=1368, max=7484.5k, avg=3137.78, stdev=2150.57
    clat (usec): min=11, max=9732, avg=917.35, stdev=547.32
     lat (usec): min=16, max=9736, avg=920.56, stdev=547.29
    clat percentiles (usec):
     |  1.00th=[  523],  5.00th=[  611], 10.00th=[  644], 20.00th=[  652],
     | 30.00th=[  652], 40.00th=[  652], 50.00th=[  652], 60.00th=[  660],
     | 70.00th=[  816], 80.00th=[ 1020], 90.00th=[ 1975], 95.00th=[ 2311],
     | 99.00th=[ 2606], 99.50th=[ 2704], 99.90th=[ 3621], 99.95th=[ 5800],
     | 99.99th=[ 6915]
   bw (  KiB/s): min=119038, max=324752, per=18.64%, avg=207144.76,
stdev=37126.12, samples=3016
   iops        : min=29759, max=81188, avg=51785.82, stdev=9281.53, samples=3016
  lat (usec)   : 20=0.01%, 50=0.01%, 100=0.01%, 250=0.01%, 500=0.43%
  lat (usec)   : 750=66.26%, 1000=12.83%
  lat (msec)   : 2=10.67%, 4=9.73%, 10=0.08%
```

```
  cpu          : usr=9.04%, sys=27.55%, ctx=42005653, majf=0, minf=4654
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=100.0%
     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=0.1%, >=64=0.0%
     issued rwts: total=0,104857600,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=64
seq_read_512k: (groupid=2, jobs=4): err= 0: pid=11970: Mon May 17 05:48:24 2021
   read: IOPS=4801, BW=2401MiB/s (2518MB/s)(400GiB/170602msec)
    slat (usec): min=14, max=534, avg=62.14, stdev=40.31
    clat (msec): min=7, max=107, avg=53.24, stdev= 2.04
     lat (msec): min=7, max=107, avg=53.30, stdev= 2.04
    clat percentiles (usec):
     |  1.00th=[49021],  5.00th=[50070], 10.00th=[51119], 20.00th=[51643],
     | 30.00th=[52167], 40.00th=[52691], 50.00th=[53216], 60.00th=[53740],
     | 70.00th=[54264], 80.00th=[54789], 90.00th=[55837], 95.00th=[56361],
     | 99.00th=[57934], 99.50th=[58983], 99.90th=[61080], 99.95th=[62129],
     | 99.99th=[79168]
   bw (  KiB/s): min=401147, max=636593, per=20.86%, avg=512853.36,
stdev=91600.08, samples=1360
   iops        : min=  783, max= 1243, avg=1001.18, stdev=178.93, samples=1360
  lat (msec)   : 10=0.01%, 20=0.01%, 50=3.30%, 100=96.69%, 250=0.01%
  cpu          : usr=0.40%, sys=8.01%, ctx=766631, majf=0, minf=10958
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=100.0%
     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=0.1%, >=64=0.0%
     issued rwts: total=819200,0,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=64
seq_write_512k: (groupid=3, jobs=4): err= 0: pid=12523: Mon May 17 05:48:24 2021
  write: IOPS=3020, BW=1510MiB/s (1583MB/s)(400GiB/271239msec)
    slat (usec): min=24, max=61814, avg=84.42, stdev=112.65
    clat (msec): min=3, max=165, avg=84.66, stdev= 1.84
     lat (msec): min=3, max=165, avg=84.74, stdev= 1.83
    clat percentiles (msec):
     |  1.00th=[   84],  5.00th=[   84], 10.00th=[   84], 20.00th=[   85],
     | 30.00th=[   85], 40.00th=[   85], 50.00th=[   85], 60.00th=[   85],
     | 70.00th=[   86], 80.00th=[   86], 90.00th=[   86], 95.00th=[   87],
     | 99.00th=[   89], 99.50th=[   89], 99.90th=[   93], 99.95th=[   95],
     | 99.99th=[  125]
   bw (  KiB/s): min=349184, max=418816, per=24.99%, avg=386355.32, stdev=3075.09,
samples=2168
   iops        : min=  682, max=  818, avg=754.45, stdev= 6.04, samples=2168
  lat (msec)   : 4=0.01%, 10=0.01%, 20=0.01%, 50=0.03%, 100=99.93%
  lat (msec)   : 250=0.02%
  cpu          : usr=2.42%, sys=4.55%, ctx=819242, majf=0, minf=8829
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=100.0%
     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=0.1%, >=64=0.0%
     issued rwts: total=0,819200,0,0 short=0,0,0,0 dropped=0,0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=64

Run status group 0 (all jobs):
   READ: bw=1336MiB/s (1401MB/s), 1336MiB/s-1336MiB/s (1401MB/s-1401MB/s),
io=400GiB (429GB), run=306556-306556msec

Run status group 1 (all jobs):
  WRITE: bw=1085MiB/s (1138MB/s), 1085MiB/s-1085MiB/s (1138MB/s-1138MB/s),
```

```
io=400GiB (429GB), run=377376-377376msec

Run status group 2 (all jobs):
   READ: bw=2401MiB/s (2518MB/s), 2401MiB/s-2401MiB/s (2518MB/s-2518MB/s),
io=400GiB (429GB), run=170602-170602msec

Run status group 3 (all jobs):
  WRITE: bw=1510MiB/s (1583MB/s), 1510MiB/s-1510MiB/s (1583MB/s-1583MB/s),
io=400GiB (429GB), run=271239-271239msec

Disk stats (read/write):
  nvme0n1: ios=105677000/105676081, merge=0/0, ticks=121498402/165226199,
in_queue=289648223, util=100.00%
Bus error
```

Сравню полученные результаты с заявленными в спецификации (в скобках).

- Последовательное чтение: 2518 (3140) MBPS
- Последовательная запись: 1583 (1520) MBPS
- Случайное чтение: 342000 (465000) IOPS
- Случайная запись: 278000 (40000) IOPS

Чтение чуть меньше заявленного, а запись быстрее. Здесь, скорее всего, сыграло свою роль кэширование и недостаточно большой срок тестирования.

Заключение

Результат моего тестирования не выявил явных проблем с производительностью выбранного NVMe диска. Результаты отличались от заявленных производителем, но следует учесть, что тестируемый диск не новый, тест недолгий, задействован кэш.

Я не самый большой специалист в тестировании дисков и не досконально знаю все возможные опции для настройки файл-теста. Однако, готовые шаблоны можно всегда найти в Интернете.

https://fio.readthedocs.io/en/latest/fio_doc.html#examples

Дотошные читатели могли заметить, что мой тест завершается строкой "Bus error". Скорее всего, данной ошибкой можно пренебречь, советуют воспользоваться более новой версией fio. Но я не стал заморачиваться, поскольку результат тестирования меня устроил.

Хотелось бы отметить, что утилитой **fio** кроме дисков можно проводить тестирование программных и аппаратных массивов.