Micron P320h 700 GB PCIe SSD evaluation under Linux

Hussein N. El-Harake and Thomas Schoenemeyer
Swiss National Supercomputing Centre (CSCS), Manno, Switzerland
hussein@cscs.ch, schoenemeyer@cscs.ch

Abstract – We evaluated the P320h drive built on Micron SLC NAND based on standard benchmark tools. This PCIe attached SSD device manufactured by Micron exhibits promising performance numbers for IOPS, read and write throughput combined with very low latency within a power envelope at only 25W.

The device with PCIe Gen 2.0 x8 interface is available in two configurations, 350GB and 700GB capacity. We used the latter in our evaluation study.

We carried out the study using two dual-socket servers using two different boards with Intel Xeon processors running RedHat 6.2.

As against other available studies, we used Linux instead of Windows. Some test have been done on raw device, however the results presented in this report were obtained using ext3 filesystem. At the end it is our goal to integrate such device in the HPC world like for Lustre or GPFS and to improve the parallel I/O performance.

Overall the device appeared to be a very reliable device with good performance. The test results are consistent with the specifications indicated in product briefs distributed by Micron.

1 Introduction

In our papers from June 2011 [1], September 2011 [2] and February 2012 [3] we presented performance numbers for several PCIe SLC Flash devices built by TMS, FusionIO and Virident.
Since October a new flash device manufactured by Micron is available at CSCS and we repeated the FIO benchmarks and added mdtest in this study. The P320h PCI flash device with 700 GB capacity were deployed to two different type of servers,

a) a dual-socket server (2U) manufactured by Supermicro. It deploys two Intel Xeon X5690 processors with 6 cores running at 3.46 GHz with 48GB Main Memory. The server offers four x16 PCIe Gen 2.0 slots. We used Red Hat 6.2. Hyperthreading and Turbo Boost was enabled. This allows for maximum core frequency of 3.73 GHz
b) a dual-socket server (2U) manufactured by ASUS based on the ESC4000/FDR G2 board with 32 GB RAM and two Intel Xeon E5-2670 CPUs. The board offers 8x PCIe 3.0 x8 slots. Hyperthreading was always enabled, and Turbo Boost was disabled for the first test series and enabled for the second series activated. This allows 3.3 GHz maximum core frequency.

The Micron device has a PCIe Gen 2.0 interface, therefore it is not expected to gain better performance on the server b). We should be able to get Gen2 speeds on the PCIe Gen3 link. The device will not run at PCIe Gen3 speeds.

2 P320h device

The Micron P320h device is available to the general market since mid 2012 and is build on SLC NAND. It is a very dense design with one mother card with 512 GB and two daughter cards with sixteen 16 GB NAND modules giving additional 512 GB of memory. The two daughter cards are mounted on top of the mainboard as shown in figure 1. In total the device provides 1 TB of raw NAND memory. The usable capacity is 700GB is obtained from the 1TB raw capacity due to the data protection solution and 22% Over Provisioning. The device features a single chip 32-channel controller on the card - a 1517-pin controller made by IDT [4] that connects all NAND packages.

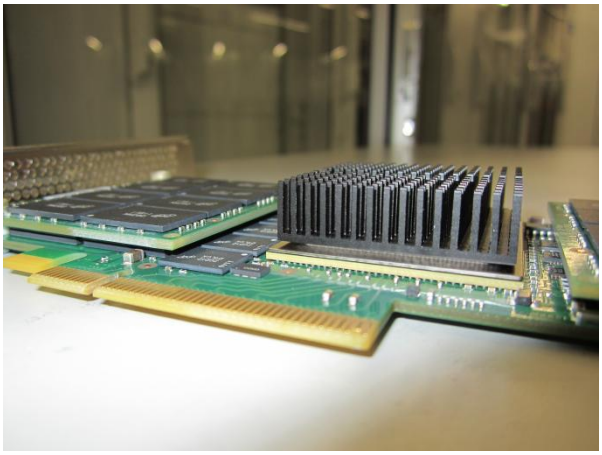

Figure 1: P320h SSD device with daughter cards on top.

As shown in figure 2, the form factor allows the card to plugged into the riser card of the ASUS board.

Reviews of this SSD device card had been published in [5], [6] and [7]. The numbers in [5] and [6] were measured with Windows, the results revealed good performance numbers of more than 2GB/s read performance and more than 217000 IOPS for 4K blocks (see figure 3).

Figure 2: Device mounted to the riser card inside the ASUS board

The study in [7] is focused on Linux and confirms the excellent performance specs, too. The test method is not clearly explained, but the author mentions the Storage Networking Industry Association (SNIA) industry accepted performance test. We assume all results were derived using the card as raw device similar to the SNIA standards.

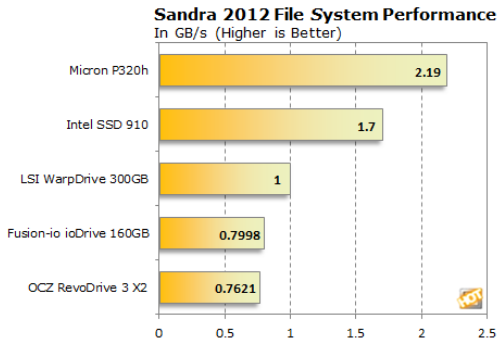

Figure 3: Taken from [5].

For our purposes at CSCS we are mainly interested in the performance when using a common Linux file system, however this kind of evaluation is not yet available.

Micron offers on their website a variety of excellent material such as whitepapers and technical documents [8], [9], [10], [11] and [12].

The level and the quality of these papers is excellent, e.g. the technical note [11] describes in detail how to optimize and test

the performance of the solid state drive. In our case the document was very helpful, since we were initially faced with unexpected low IOPs numbers. We checked the system settings and noticed that power limiting was enabled. This parameter causes the device to stay within 25W PCIe slot power limit. Enabling this limit can lower the performance by 40%. After we disabled power limiting, we observed the expected performance numbers.

The public list price of this card is roughly $US 7000 according to [5], which is from our perspective a very competitive price/performance ratio compared to other products in the field.

The form factor of this card is very compact, it is a half height, half length design, that would allow us to put 4 of those cards in our server.
Other outstanding features are the indicated read and write latencies as well as the endurance which is estimated to 50 PB of write during the lifetime of the 700 GB device.

The main characteristics in the product brief [13] are summarized in table 1. The brief specifies a maximum read performance of 3.2 GB/s and up to 785K IOPs at 4 KB blocksize. The peak write bandwidth is estimated with 1.9GB/s and 205K IOPs at 128KB block size. The max read performance is close to the max bandwidth of what a PCIe Gen2.0 x8 interface could deliver.

The excellent specs regarding the write latency (9µs) are given for write caching enabled, which has been tested by Micron internally, however Micron does not allow the users to enable this feature until now.

	P320h
Capacity[1]	350GB, 700GB
Interface	PCIe (Gen2-compliant) x8
Sequential READ	Up to 3.2 GB/s[2]
Sequential WRITE	Up to 1.9 GB/s[2]
Random READ	Up to 785,000[3] IOPs
Random WRITE	Up to 205,000[3] IOPs
READ Latency	<42µs (512 bytes) posted
WRITE Latency	<9µs posted
Active Power Consumption	25W max
Endurance[4] (total bytes written)	350GB – 25PB; 700GB – 50PB
Operating Temp	(0°C to +50°C)
Form Factor	HHHL*
Dimensions	68.90mm x 167.65mm x 18.71mm

[1] Unformatted. 1GB = 1 billion bytes. Formatted capacity is less.
[2] 128KB transfer size, steady state. [3] 4KB transfer size, steady state.
[4] Lifetime endurance is measured not in years, but in the number of bytes that can be written to the device.
** Half height, half length: 68.90mm x 167.65mm x 18.71mm*

Table 1: Micron P320h datasheet [14]

Micron also deploys an unique technique to enhance data protection analog to RAID5. This data protection solution is called RAIN (redundant array of independent NAND). When data is stored in nonvolatile memory, it is essential to have mechanisms that can detect and correct a certain number of errors. Micron implemented ECC that can detect 2-bit errors and can correct 1-bit errors per 256 or 512 bytes [11]. Since development goes on and controller complexity increases, simple ECC is no longer sufficient.

Micron decided among the different RAID options for a 7+1(P) RAID 5 architecture which proved to be the most balanced solution for both performance and failure protection. In an event of a failure, the drive can seamlessly recover the data of the failed NAND from the parity data.

Blocks and pages are striped across logical units, with parity data calculated from every 7 blocks/page as shown in Figure 4.

Each physical block in a SLC SSD device can be programmed and erased reliably 100K

times. That means, a wear leveling algorithm needs to be implemented in order to spread the number of write cycles per block. This algorithm is implemented in the flash translation layer (FTL). Micron deploys both, static and dynamic wear leveling. The FTL is updated accordingly based on the results of running the wear levering algorithm across the NAND die array on the drive.

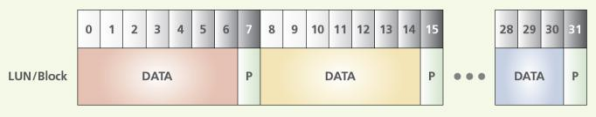

Figure 4: 7+1 RAID5 architecture per channel of the SSD

3 Evaluation Method

All experiments described below were performed with the ext3 file system. We used fio 2.0.10 [14] and mdtest [15] for our tests and used the latest driver provided by Micron. Our version number was 131.02.00 from November 2012. Following configuration was used for the tests:

```
Device Name                : rssda
Model No                   : Micron P320h-
MTFDGAR700SAH
FW-Rev                     : B1490300
Total Size                 : 700.15GB
Drive Status               : Drive in good health
SMARTSupport               : Yes
SMARTEnabled               : Yes
WriteCacheEnabled          : No
Interrupt Coalescing       : D801F
Power Limit                : Disabled
```

SMART is a self-monitoring, analysis and reporting technology feature set and helps to predict the likelihood of nearterm degradation or fault condition of the device. As mentioned before, power limiting was disabled.

We used in our setup an interrupt coalescing value of D801F which gives a higher IOPs value at a slightly higher latency. This is recommended in environments with significantly more than 256 outstanding commands such as we expect in our environment at CSCS.

Since the P320h has been designed to operate at peak performance at a queue depth of 256, we submitted the tests on the 6-core Intel X5690 CPU with 12 threads multiplied with an IOdepth of 24, and we used on the 8-core Intel E5-2670 CPU 16 threads multiplied with an IOdepth of 16 to have at least a queue depth of 256. Even exceeding this number did not result in any performance deviation by more than 5%.

All results for the Sandy Bridge server are shown for Turbo Boost disabled, since the deviation in the time series between Turbo Boost enabled and disabled is less than 3%.

All measurements with FIO have been carried out for block sizes between 512B and 1024KB.

4 Results

4.1 FIO

First, we compare the results for the two servers in the figures below. In all our experiments the device used in the Sandy Bridge Architecture with PCIe Gen3.0 slots delivered lower performance, especially for small block sizes. The throughput for block sizes larger than 8K, the performance was similar.

An interesting aspect is the excellent IOPS write result we measured at 4K block size. The data labels are shown in figure 5 and the peak of 401K exceeds the product brief specifications by about a factor 2. We

believe we did not reach the complete steady state condition in this test.

We verified that with another test using the SSD card as raw device. We were able to reach the complete steady state after 10-15 minutes with 4K random writes. The settling number was 211K which is consistent to other studies. However, when using a file system it is quite tricky to reach the steady state condition.

In all our studies we filled our drive up to nearly 100% of the capacity and all tests were running for 10 to 40 min. This usage pattern and the related performance results are more relevant for our environment at CSCS than steady state numbers.

The IOPs value for read at 4K is 704K (figure 6), which is 10% less than what is specified in the product brief (785K). We also measured the peak at 512B like reported in [7]. 1.3 Mio IOPs are achieved on the Westmere server and this is one of the best results we ever measured on a flash device.

The observed performance difference between the two server architectures is rather large in this case, especially at small block sizes below 16K (figure 6).

As shown in figure 7, the write bandwidth peaks at 1.9 GB/s between 8K and 1024K block size and confirms the product brief specs.

The read bandwidth shown in figure 8 is also consistent with the specs, we measured 3.2 GB/s for block sizes larger than 16K, at 8K the bandwidth is still 3.1 GB/s.

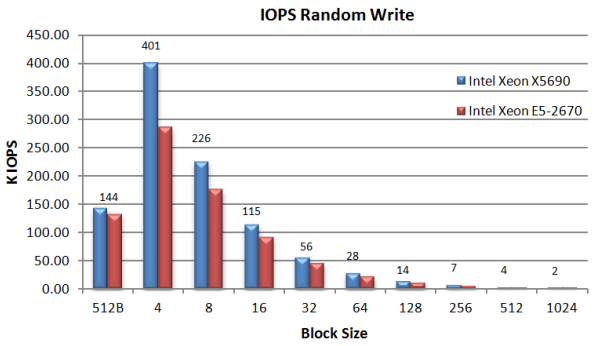

Figure 5: IOPS Random Write

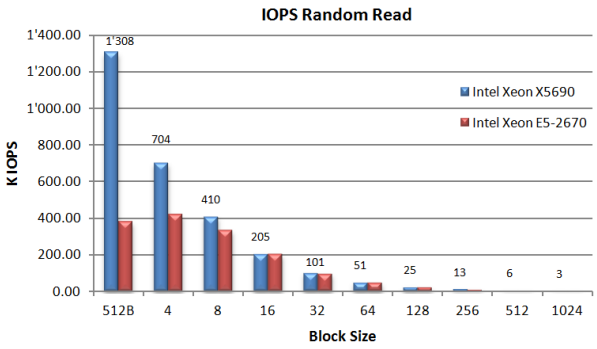

Figure 6: IOPS Random Read

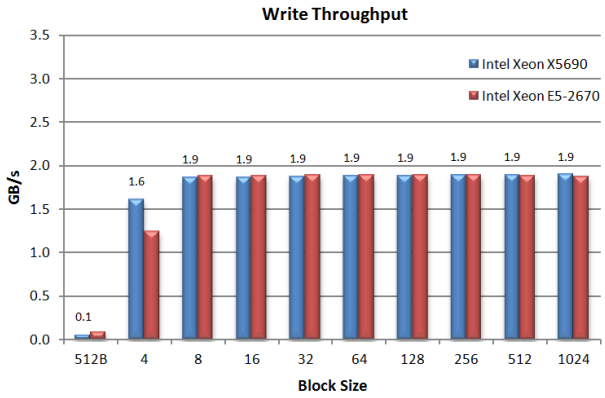

Figure 7: Write Throughput

As in the previous IOPs tests, the differences caused by Gen3.0 affect only the small block sizes. Because the PCIe Gen2.0 is limited to approximately 3.2 GB/s of throughput, hence the P320h saturates its interface.

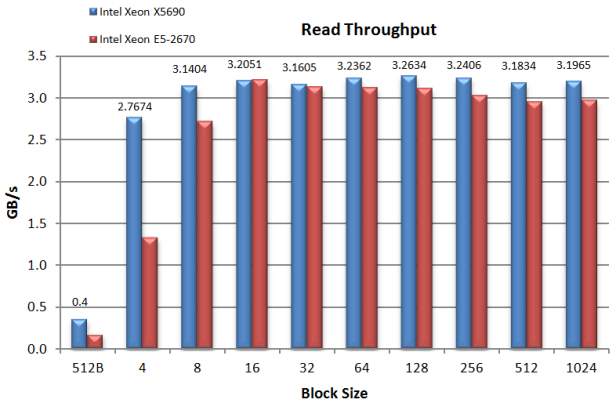

Figure 8: Read Throughput.

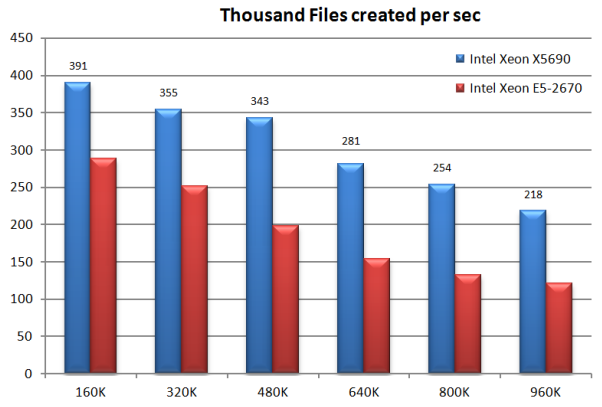

Figure 9: Mdtest result for file creates per second dependent on the number of files.

4.2 MDTEST

Mdtest measures the performance of multiple tasks creating, stating and deleting files and directories. It is a MPI code, so it can run processes in parallel. We used the "single process" version We run mdtest create, stat and remove files on the card on both servers up to a total number of 960K files with 24 MPI tasks on the Westmere server and with 32 MPI tasks on the Sandy Bridge system, the results are shown in figure 9.

It took 0.41 seconds to create 160K files which corresponds to 391K file creates per sec, which corresponds very well to our result in the chapter before (figure 5). If we further increase the number of files, the performance degrades gradually and ends up at 4.4 sec for 960K files which is equivalent to 218K file creates per second. This amount of file creates seems to saturate the device and pushes this card obviously closer to the steady state that the FIO tests we run in the previous chapter.

Again the Sandy Bridge server performs significantly worse than the Westmere server.

A similar behaviour is observed for file removes as shown in figure 10, however this process is significantly slower by 40% compared to the file creates. We also see the saturation at the level of 960K files.

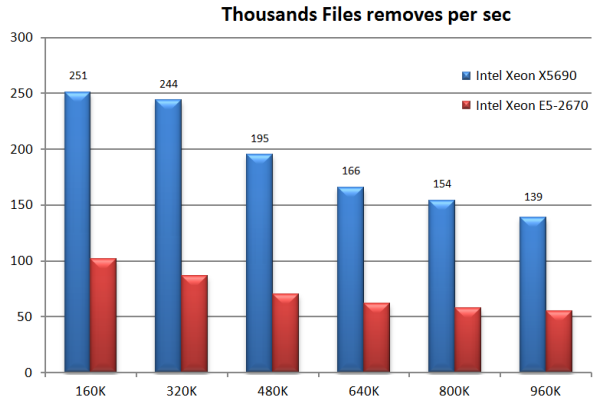

Figure 10: Mdtest result for file removes per second dependent on the number of files.

The number of operations per second for file stats achieves a remarkable high number of 4 Mio files per second as shown in figure 11.

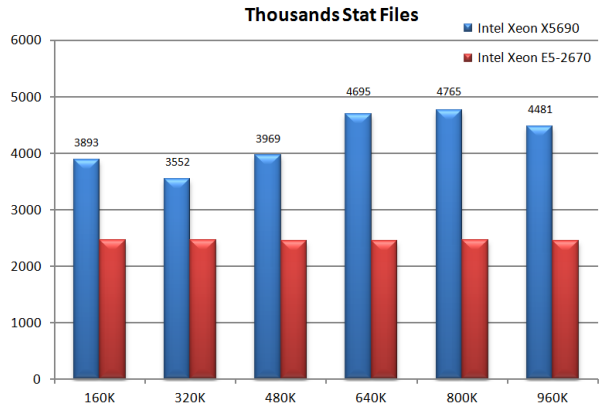

Figure 11: Mdtest result for file stats per second dependent on the number of files.

4.3 Latency

The SSD response time is critical for the overall system performance. Storage vendors usually do not report how they measure latencies of their devices.

The FIO benchmark provides two variables "clat" and "slat". The sum "lat" is the amount of time each IO request will take to complete. Since no rotating parts are involved in SSDs, it is the type of chip and the design of the chip controller that has an impact on the latency.

The product brief gives a latency of 42µs for read and 9µs for write. Micron explained the excellent number for write is measured with RAM write caching enabled. This is currently not available yet to users. There are no write latency indicated without write caching.

We run the latency measurements at 512B, 4K and 8K block sizes on the Westmere server, the results for lat (average) are shown in figure 12.

The average read latency is 47µs for 512B block size with a very low standard deviation of 1µs (see table 2), which is very close to the specification in the product brief. The average write latency is 336µs for the same block size with a standard deviation of 18µs. Thus, 95% of all results are below a read latency of 49µs and write latency of 372µs which are comparable with results measured on an older Micron PCIe SSDs reported in [16].

```
512  write

write: io=18975KB, bw=1482.4KB/s, iops=2964 , runt= 12803msec
   slat (usec): min=2 , max=34 , avg= 4.51, stdev= 2.67
   clat (usec): min=305 , max=1154 , avg=331.19, stdev=18.27
    lat (usec): min=324 , max=1159 , avg=335.91, stdev=19.29
   clat percentiles (usec):


512 read

w: (groupid=0, jobs=1): err= 0: pid=23423: Thu Dec  6 09:18:04 2012
  read : io=203360KB, bw=10577KB/s, iops=21153 , runt= 19227msec
    slat (usec): min=2 , max=28 , avg= 2.61, stdev= 0.59
    clat (usec): min=19 , max=276 , avg=43.81, stdev= 1.22
     lat (usec): min=22 , max=279 , avg=46.54, stdev= 1.27
    clat percentiles (usec):
```

Table 2: logfile for FIO latency test for 512B block size

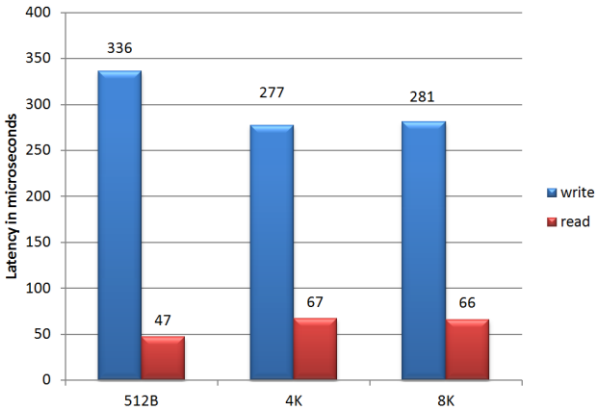

Figure 12: Average Latency for read and write depending on the block size.

5 Conclusion

We evaluated the new P320h SLC flash device with FIO and mdtest. A short summary of results is shown in table 2.

- Our measurements are consistent with most the specifications provided by Micron. In a few cases our observations peaked the datasheet specs
- The peak numbers for read and write throughput are identical or slightly better than indicated numbers in product briefs
- The highest number of IOPS was measured at 512B for random read and did exceed 1.3 Mio. IOPS. However we could not reproduce the IOPS number for 4KB, our result was lower by 81K IOPs
- The write IOPS for 4K block size are excellent, although we are aware that we did not reach the steady state
- Mdtest showed us excellent results for creating a huge number of files (up to 960K). This benchmark confirmed the IOPS write peak with nearly 400K and the number at the lower end (218K) which is close to the specified steady state number
- It is currently not useful to plug this device in a Intel Xeon E5-26xx server with Gen3.0 slots, since the performance is partially worse, especially for small block sizes
- All results were very well reproducible and device was stable and reliable through our whole testing period (1 month)
- Excellent material is available through the Micron website, some improvement on the man pages would be helpful
- Latency numbers are within our expectations and comparable to other PCIe SSDs device available in the market

MICRON	Read		Write	
P320h	4KB	128KB	4KB	128KB
GB/s datasheet	NA	3.2	NA	1.9
GB/s measured	2.8	3.3	1.6	1.9
IOPS datasheet	785K	NA	205K	NA
IOPS measured	704K	25K	401K	14K

Table 3: Device specs according to the datasheet compared to observations

6 Literature

[1] Hussein N. Harake and Thomas Schoenemeyer: Detailed Analysis of Solid State Disks, Technical Paper, CSCS, 7/2011.

[2] Hussein N. Harake and Thomas Schoenemeyer: Comparison of PCIe SLC Flash cards, Technical Paper, CSCS, 9/2011.

[3] Hussein N. Harake and Thomas Schoenemeyer: Comparison of PCIe SLC Flash cards – Virident FlashMAX and Texas Memory Systems RamSan-70, Technical Paper, CSCS, 2/2012.

[4] IDT 89HF32P08AG3 32-Channel PCIe® x8 G3 Enterprise Flash Controller; Product Brief, 2012

[5] Dave Altavilla : Review, 10/2012

[6] Anand Lal Shimpi: Review, 10/2012.

[7] Christopher Ryan, MICRON P320H PCIE Enterprise SSD Review, 10/2012

[8] Micron Whitepaper on how system settings impact PCIe SSD Performance, 7/2012

[9] Micron Technical Marketing Brief, NAND Flash Media Management through RAIN, 6/2012

[10] Micron Technical Marketing Brief, an overview of SSD Write Caching, 5/2012.

[11] P320h SSD Performance Optimization and Testing, Technical Note, 10/2012.

[12] Micron, ECC, Technical Note, 2011.

[13] Micron: P320h PCIe SSDs Best-in-class Performance Product Brief, 9/2012

[14] Fio I/O tool for benchmark and hardware verification, website, 2011

[15] Mdtest Benchmark.

[16] Seppanen et al., High Performance Solid State Storage under Linux, 2010.