EPJ Web of Conferences **251**, 04013 (2021) https://doi.org/10.1051/epjconf/202125104013
CHEP 2021
FERMILAB-CONF-21-761-SCD

Evaluation of a high-performance storage buffer with 3D XPoint devices for the DUNE data acquisition system

Adam Abed Abud[2,4,*], *Kurt* Biery[3], *Carlos* Chavez[4], *Pengfei* Ding[3], *Eric* Flumerfelt[3], *John* Freeman[3], *Giovanna* Lehmann Miotto[2], *Marco* Roda[4], *Philip* Rodrigues[5], *Roland* Sipos[2], *Alessandro* Thea[2,6], and *Brett* Viren[5]

[1]Brookhaven National Laboratory, Upton, NY, USA
[2]European Laboratory for Particle Physics (CERN), Geneva 23, CH-1211, Switzerland
[3]Fermi National Accelerator Laboratory, Batavia, IL, USA
[4]University of Liverpool, The Oliver Lodge, Oxford St, Liverpool L69 7ZE, United Kingdom
[5]University of Oxford, Oxford OX1 3RH, United Kingdom
[6]Rutherford Appleton Laboratory, Didcot, United Kingdom

Abstract. The DUNE detector is a neutrino physics experiment that is expected to take data starting from 2028. The data acquisition (DAQ) system of the experiment is designed to sustain several TB/s of incoming data which will be temporarily buffered while being processed by a software based data selection system. In DUNE, some rare physics processes (e.g. Supernovae Burst events) require storing the full complement of data produced over 1-2 minute window. These are recognised by the data selection system which fires a specific trigger decision. Upon reception of this decision data are moved from the temporary buffers to local, high performance, persistent storage devices. In this paper we characterize the performance of novel 3DXPoint SSD devices under different workloads suitable for high-performance storage applications. We then illustrate how such devices may be applied to the DUNE use-case: to store, upon a specific signal, 100 seconds of incoming data at 1.5 TB/s distributed among 150 identical units each operating at approximately 10 GB/s.

1 Introduction: the DUNE data acquisition

The Deep Underground Neutrino Experiment (DUNE) [1] is a liquid-argon (LAr) experiment that will be built at the Sanford Underground Research Facility (SURF) in South Dakota and will be dedicated to the study of neutrino physics. The role of the data acquisition (DAQ) system is to process, buffer and filter TB/s of incoming data from the the detector's front-end electronics. In the current baseline design for the first detector module, the plan is to use 160 PCIe-based devices for detector readout (FELIX system [2]), each operating at approximately 10 GB/s. The objective is to allocate one CPU for each board for both processing and storage. The data selection system is then responsible for issuing trigger decisions and to select the most interesting data fragments.

One of the physics goals of the DUNE detector is to store data originating from rare physics processes (e.g. supernova burst neutrino events or SNB [3]). This requires the data

*e-mail: adam.abed.abud@cern.ch

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

acquisition system to persist on storage media incoming data for over 100 seconds. Therefore, the research objective is to identify a suitable storage solution that is capable of sustaining the data bandwidth of at least one readout board for approximately 2 minutes.

The custom PCIe devices of the readout system are designed to produce, from 10 links, a user payload with a size of 464 bytes and with a rate of 2 MHz. However, as described in [4] in order to reduce the amount of memory I/O operations on the hosting server the user payloads are aggregated into *superchunks* of 5568 bytes and with a rate of 166 kHz. If an SNB candidate is detected by the data selection system (with a fake rate of typically once a month) the *superchunks* of data coming at 166 kHz from ten links for over 100 seconds will be saved in a local storage. Later on, those data will be transferred at a lower pace to the data filtering and storage system.

In this paper we will characterize the performance of novel, high-throughput SSDs based on the 3D XPoint technology as a possible candidate for the DUNE supernova local storage buffer. After a brief overview of the 3D XPoint technology (section 2), the results of the synthetic evaluation are illustrated in section 4. Finally, section 5 illustrates the results of the integration of the Micron X100 device in prototype test application emulating DUNE data acquisition system.

2 A high level overview of the 3D XPoint technology

3D XPoint is a non-volatile memory (NVM) technology that has been jointly developed by both Intel® and Micron Technology Inc.®. The innovation of the 3D XPoint technology is the use of a phase-change material that modifies the bulk resistance for the bit storage in the memory cell. Data is stored in a stack of memory cells and the data access is controlled by a selector switch that is responsible for the selection of the correct memory cells. This is different from typical memory/storage devices which instead use a transistor as a selector element [5]. For the evaluation we used the Micron X100 as an example of 3D XPoint, PCIe based NVMe SSD.

Typically, 3D XPoint devices are also characterized by a high value of *endurance* which is a parameter that represents the longevity of storage media. This value is usually expressed in PB written per TB of data, i.e. a device with 86 PBW/TB like the Micron X100 can be written 86K times with 1 TB blocks before a failure occurs. On the other hand, NAND based SSDs have a typical endurance value of 5 PBW/TB. In both cases, 3D XPoint or NAND based solution, the endurance is not a critical factor because in case a SNB event is detected the data will be written in average once per month for only 2 minutes and with a rate of 10 GB/s.

The Micron X100 drive was chosen for this evaluation because it provides a bandwidth closer to the one needed for the DUNE local storage. In the future developments of this work other storage technologies (e.g. NAND drives) will also be evaluated.

3 Description of the setup and the tools used

The goal of this section is to illustrate a sample of the performance characteristics of the Micron X100 device, keeping in mind a possible application as a storage device for the DUNE supernova buffer. The test machine used for the synthetic benchmarks is characterized by a dual socket CPU processor (16 cores) and with 32 DDR4 DRAM DIMMs each of 16 GB. The storage device used for testing is a Micron X100 SSD drive based on the PCIe Gen. 3 (16 lanes). It is worth noting that in order to achieve the highest performance it is critical to have proper thermal management (e.g. setting maximum fan speed from the BIOS) and setting all the CPU cores to *performance* mode. Table 1 summarizes the specification of the machine node used for evaluation.

Table 1. Overview of the test machine used for the evaluation.

CPU	AMD® Epyc® 7302 Dual socket, 16 cores
DRAM	DDR4 DRAM 16 GB, 3200 MT/s, 32 slots
OS	Centos 7, Linux Kernel 5.10
Storage	Micron X100 SSD (750 GB)
SW	FIO v3, AIO library
NOTE	All CPU cores have been set to *performance* mode

For the synthetic benchmarks a first evaluation was done with the *flexible-IO* (FIO) tool[6]. This is a standard tool used to benchmark storage technologies as it allows users to tune the tests to achieve an emulated workload as close as possible to the final application. In addition, *fio* also provides access to low-level parameters that make it easier to tune and achieve the highest performance from the storage hardware.

Another tool that was used in the evaluation is the *libaio* library [7]. One of the motivations for using this tool in addition to *fio* is to have a slightly more realistic, higher-level application that resembles the workload expected in the DUNE data acquisition system. The main reason to use libaio compared to other tools is motivated by the need to have a high-performance storage library that is capable of efficiently writing data to NVMe devices. In fact, the libaio library performs asynchronous file operations by relying on the native kernel AIO interface and thus it is closer to the hardware device rather than performing operations from user-space. The asynchronous nature is achieved by using the O_DIRECT kernel flag which allows to bypass the operating system page cache. In this way the write operation inserts the data directly into the storage device's queue which can then process the input.

4 Synthetic benchmarks

In this section, the performance evaluation of the Micron X100 SSD is illustrated in terms of its achieved sequential write throughput. The novelty of this section lies in evaluation of a recently announced device which was tested from a application perspective without relying only on low-level benchmarking tools.

Before illustrating the results it is also worth mentioning that we performed all the tests by carefully setting the CPU affinity of the executing thread to the corresponding NUMA node corresponding to the PCIe bus on which the Micron X100 SSD was inserted. This was done in order to achieve the highest performance and avoid unnecessary cross-NUMA access that could lead to a loss in performance or oscillations.

Figure 1 illustrates the sequential write throughput of the Micron X100 SSD as a function of the block size for a single thread. The scan was performed starting from a block size of 4 KiB up to a block size of 32 MiB. The maximum performance that the drive is capable of sustaining is more than 8.5 GiB/s. The block size at which the peak performance is reached (throughput stability) was achieved starting from a block size of 8 MiB. This suggests that, in case of a single thread writing to the Micron X100 SSD, it is necessary to use large block sizes to achieve the highest throughput.

The throughput as a function of the number of threads is shown in figure 2. This measurement was done issuing up to 10 threads, which is the expected number of links that the

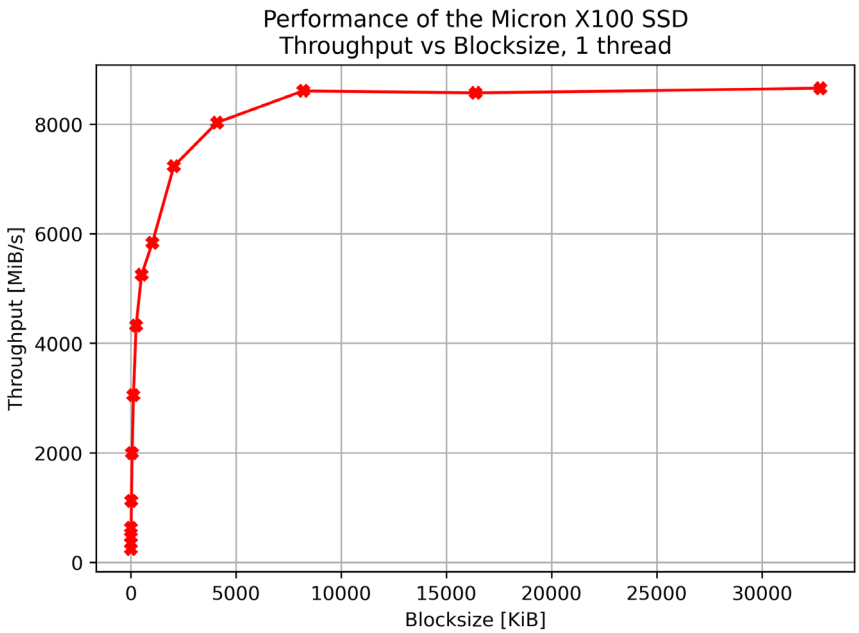

Figure 1. Write throughput as a function of the block size ranging from 4 KiB up to 32 MiB. The system reaches the maximum bandwidth with a block size of around 8 MiB.

readout board needs to sustain as described in section 1. A scan in block size was also executed ranging from the KiB regime up to the MiB regime. The results show that by increasing the number of writing threads it is possible to achieve the maximum throughput with smaller block sizes. As an example, with a block size of 32 KiB the throughput achieved with 10 threads is 4 times higher than the achieved throughput with only one thread. Thus, this confirms that in order to fully exploit the performance of the drive with block sizes smaller than 4 MiB a multi-threaded or multi-process approach is necessary. It is also worth noting that the measurement was performed by using direct I/O which is a feature of the linux file system in order to bypass the operating system cache. Therefore, in this way it is possible to write directly from the application to the storage device and measure the raw performance of the storage technology. In addition, when performing the test with a block size of 4 MiB the system CPU utilization (per thread) was approximately 15% for the whole testing time. This originates from the *libaio* library which relies on kernel calls to perform the asynchronous operations and it indicates that the throughput of the application is not bound by the CPU load but by the physical performance of the storage drive.

The synthetic evaluation has shown that the throughput achieved with the Micron X100 device is approximately 85% of the target throughput needed by the DUNE readout. This does not represent a major obstacle because the objective is to store incoming data for approximately 100 seconds and by increasing the size of the transient buffer it is possible to compensate for the missing throughput. Therefore, the Micron X100 still represents an interesting device for the local storage system of the DUNE experiment.

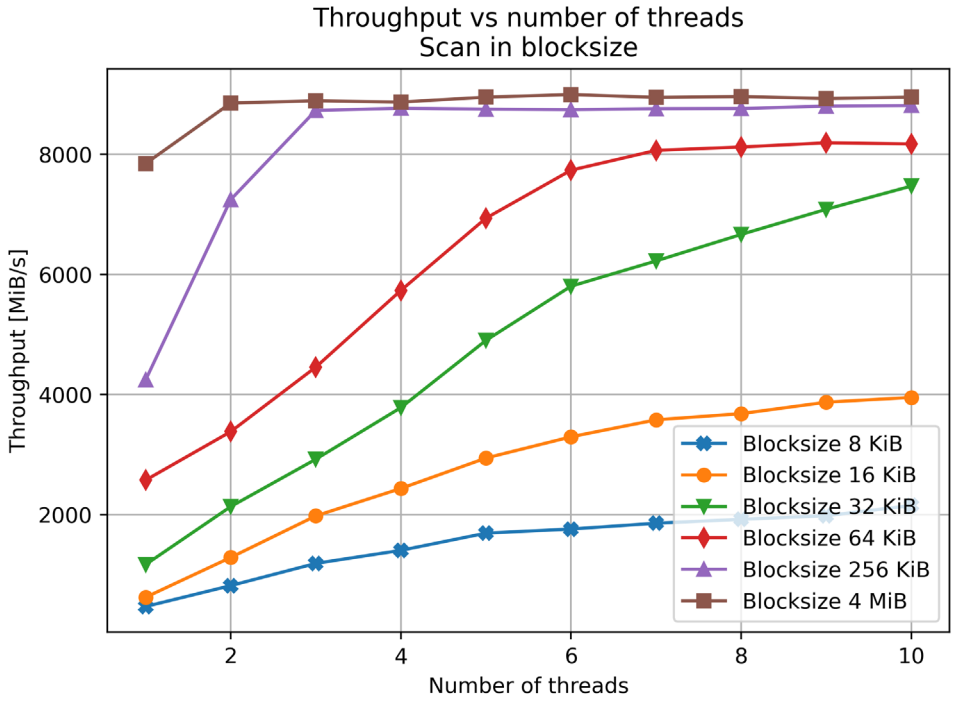

Figure 2. Write throughput as a function of the number of threads ranging from 1 to 10. A scan in block size has also been performed from 8 KiB up to 4 MiB.

5 Integration with the DUNE data acquisition system

Although the results in section 4 show a very good match between the performance tests and the specification of the hardware drives, an evaluation with a more realistic workload was performed to assess the suitability of the Micron X100 drives for the DUNE experiment. This was done using a test application (*MiniDAQ*) that contains the most relevant features needed to emulate the final data acquisition system. Figure 3 illustrates the main components of the *MiniDAQ* application: Readout (grey), Dataflow (blue) and Trigger (green). The Readout Emulator emulates the data input from half a readout board. The Data Link Handlers manage the temporary buffering of raw data for each link. The Trigger Decision Emulator acts as the data selection system. Trigger decisions are translated into data requests by the Request Generator and the Data Link Handlers respond to any data requests by extracting the raw data, formatting them and forwarding them to the Fragment Receiver. The Fragment Receiver aggregates data corresponding to the same trigger decision and forwards data to the Data Writer, which implements the interface to the permanent storage. Each Data Link Handler receives a stream of 5568 bytes at rate of 166 kHz. Two instances of the *MiniDAQ* application were used in parallel to emulate the traffic of one DUNE readout board.

Algorithm 1 describes the mechanism of the Data Writer process. The most relevant elements in the initialization steps are the creation of a memory aligned buffer (by using the *posix_memalign* function), starting of the worker thread and setting the CPU affinity to the relevant physical core. This last instruction is particularly important to achieve a stable

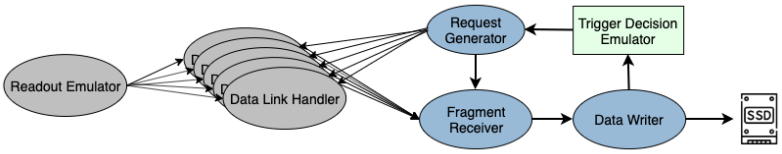

Figure 3. Schematic representation of the main elements that constitute the MiniDAQ test application.

system and to reduce the number of context switches during the execution and that can lead to a loss in the total bandwidth. In addition, ensuring that the memory buffer is aligned is necessary when using the O_DIRECT flag. Initial tests showed that a 512 bytes alignment is not compatible with Micron X100 drive and therefore a 4 KiB alignment was chosen.

Algorithm 1: Data Writer mechanism.

```
initialize data store;
allocate mem-aligned buffer;
start data writer thread;
set CPU affinity;
while trigger_flag do
    receive requested_data;
    for fragment in requested_data do
        get ptr to fragment location;
        get fragment size;
        memcpy(buffer, fragment, size);
        flush_to_disk(buffer);
    end
    check inhibit(fragment);
end
```

An inhibit mechanism is also in place in case the Data Writer is not able to sustain the rate of incoming data and, in this case, a warning message (e.g. *Dataflow is BUSY*) is issued. This means that, in order to avoid data loss, it is necessary to reduce the data extraction rate because the writing process is not able to keep up with the data production rate.

When a *trigger_flag* is enabled the requested data is sent to the Data Writer with a configurable data request size and trigger rate, thus determining the overall storage rate of the application. Therefore, the total storage rate to disk is given by:

$$\text{storage-rate} = (\text{data.request.size}) \times (\text{trigger.request.rate})$$

Figure 4 shows the resulting throughput obtained with the prototype test application. Two MiniDAQ applications were started, each emulating the handling of 5 data links from a readout unit. At present the writer is handled in one thread per application. Due to the very large data sizes expected in the DUNE use case, two instances are sufficient to saturate the performance of the Micron X100 storage drive. Tests were done by varying the data request size and measuring the writing throughput to the drive. The request rate was chosen in order to maximise the throughput in a stable system without any trigger inhibit. The maximum throughput is achieved starting from a data request size of 16 MiB. As a comparison, the synthetic performance test with two running threads has been included on the same plot (orange

entries). Similarly to the results obtained in figure 1 the maximum throughput measured is approximately 8.5 GiB/s which corresponds to more than 80% of the target throughput needed for the DUNE storage buffer.

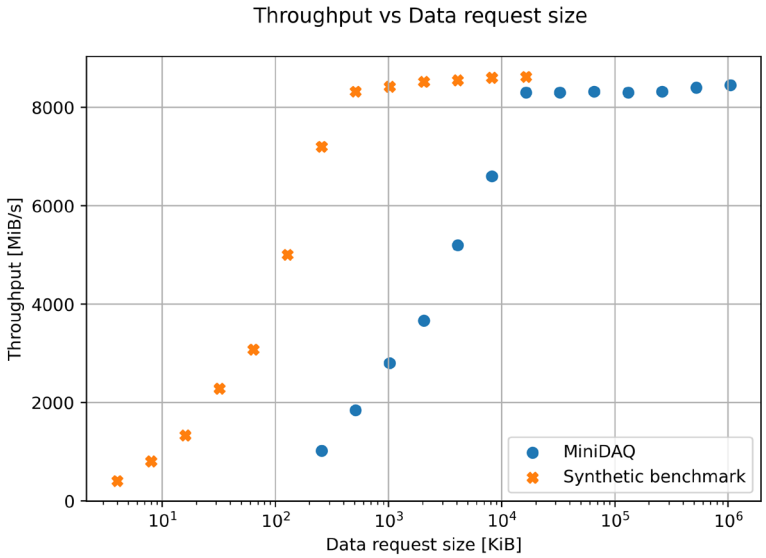

Figure 4. Comparison between the throughput achieved with a synthetic benchmark with 2 threads and with two MiniDAQ applications. The maximum bandwidth of the drive of approximately 8.5 GiB/s is achieved in both cases.

There is a notable difference in the measured throughput (figure 4) between the synthetic tests and the prototype of the more realistic data acquisition application: in the synthetic performance tests the buffers used are all memory aligned and multiples of the data request size. In the MiniDAQ application the data fragment size can vary and it is not guaranteed to be a multiple of the writing I/O block size, in addition of not being memory aligned. Therefore, it is necessary to copy the data to a previously allocated buffer before flushing them to disk. This operation has a clear impact on the performance. Preliminary code profiling tests indicate that the memory copy operation in the Data Writer is the most consuming task of the process and accounts for approximately 10% of the total CPU time when running the whole MiniDAQ application. Future developments of this work will thus focus on software optimizations to reduce the overheads.

6 Summary and outlook

This work presents an evaluation of the Micron X100 SSDs through a set of generic benchmark measurements and the use of those devices in the specific use-case of the local storage system for the SNB data of the DUNE experiment.

In the first part, the sequential writing throughput was measured as a function of both I/O block size and number of parallel writers. The synthetic performance evaluation showed that it is possible to saturate the bandwidth available and confirms the nominal throughput of approximately 8.5 GiB/s of the X100.

In the second part, the goal was to assess whether Micron X100 devices are suitable for the DUNE data acquisition system. A prototype application showcasing a realistic work flow was integrated with the Micron X100. Performance measurements showed that it is possible to achieve the maximum throughput for the very large data block sizes expected in the experiment.

This work demonstrated that a single Micron X100 device can sustain steadily 80% of the traffic generated by one readout unit. It is thus a suitable solution for the storage of a SNB event where the throughput needs to be sustained for only 100 seconds. We expect that the next generation of such devices will match or exceed the throughput produced by the DUNE readout units.

Future studies will continue along two paths: on one hand the results obtained with the Micron X100 SSD will be compared to arrays of NAND based PCIe Gen4 storage devices in terms of performance, usability, power consumption and space; on the other hand the software will be further developed and optimised to handle in parallel the special SNB handling and the other physics and calibration triggers.

Acknowledgment

We would like to thank Micron Technology for their technical advise and support.

References

[1] B. Abi et al. (DUNE), JINST **15**, T08008 (2020), 2002.02967

[2] A. Borga et al., EPJ Web Conf. **214**, 01013 (2019)

[3] C. Cuesta (DUNE), *Core-Collapse Supernove Burst Neutrinos in DUNE*, in *40th International Conference on High Energy Physics* (2020), 2011.06969

[4] Sipos, Roland, EPJ Web Conf. **245**, 01019 (2020)

[5] *Patent Search Supports View 3D XPoint Based on Phase-Change*, https://www.eetimes.com/patent-search-supports-view-3d-xpoint-based-on-phase-change (2020), accessed: 04/02/2021

[6] *Flexible I/O*, https://github.com/axboe/fio, accessed: 12/02/2021

[7] S. Bhattacharya, S. Pratt, B. Pulavarty, J. Morgan, *Asynchronous I/O Support in Linux 2.5* (2010)