DUNE DAQ R&D integration in ProtoDUNE Single-Phase at CERN

Roland Sipos[1,*] for the DUNE Collaboration

[1]CERN, 1211 Geneva, Switzerland

Abstract. The DAQ system of ProtoDUNE-SP successfully proved its design principles and met the requirements of the beam run of 2018. The technical design of the DAQ system for the DUNE experiment has major differences compared to the prototype due to different requirements placed on the detector, as well as a radically different location of operation. The single-phase prototype at CERN is a major integration facility for R&D aspects of the DUNE DAQ system. The facility allows for the exploration of additional data processing capabilities and optimization of the FELIX system, which is the chosen TPC readout solution for the DUNE single-phase detectors. One of the fundamental differences from the prototype is that the DUNE DAQ relies on self-triggering. Therefore, real-time processing of the data stream for hit and trigger primitive finding is essential for the requirement of continuous readout. The supernova burst trigger requires a large and fast buffering technique, where 3D XPoint persistent memory solutions are evaluated and integrated. In order to maximize resource utilization of the FELIX hosting servers, the elimination of the 100 Gb network communication stack is desired. This implies the design and development of a single-host application layer, which is a fundamental element of the self-triggering chain. This paper discusses the evaluation and integration of these developments for the DUNE DAQ, in the ProtoDUNE environment.

1 Introduction

The Deep Underground Neutrino Experiment (DUNE)[1] is a leading-edge international experiment with a varied physics program including neutrino oscillation, proton decay and supernova studies. The DUNE Far Detector consists of four liquid-argon Time Projection Chamber (LAr-TPC) super-modules located 1.5 km underground in shielded caverns, excavated in the former Homestake gold mine. Each super-module has internal dimensions of 14 m x 14 m x 62 m, holding 17,000 tons of liquid argon at -186 °C. The ProtoDUNE-SP[2] experiment demonstrated the design, construction, and operation of the single-phase DUNE TPC technology during the beam run of Q4 2018. The active volume is 6 m high, 7 m wide, and 7.2 m deep (along the drift direction). It was filled with 750 tons of LAr and received a charged particle beam from the Super Proton Synchrotron (SPS). Ionization tracks are collected by the wires of the detector's six Anode Plane Assemblies (APAs), which amount to 4% of the 150 APAs of a DUNE super-module. The front-end cold electronics, mounted onto the APA frames, continuously amplify and digitize the induced signals on the sense wires at 2

[*] e-mail: roland.sipos@cern.ch

MHz, and transmit these waveforms to the data acquisition (DAQ) system. The DAQ system design differs fundamentally between ProtoDUNE and DUNE due to the environment of the detectors, the triggering principles, and the constraints of the readout units of the TPC. This paper discusses R&D on the TPC readout solution and the integration of these features into the ProtoDUNE-SP environment.

2 The DUNE DAQ system

The ProtoDUNE-SP DAQ system[3] had strict requirements and design considerations in order to cope with the detector's environment. The detector is located on the CERN SPS beam line which allowed for the exposure to charged particle beams. Additionally, the detector is located on the surface, leading to a large cosmic ray flux. The collaboration decided not to employ zero suppression for the prototypes, aiming instead to acquire an unbiased set of data. The DAQ system is externally triggered in order to aggregate the different trigger inputs (beam instrumentation, cosmic-ray tagger, photon-detectors). This triggering principle is highlighted in figure 1.

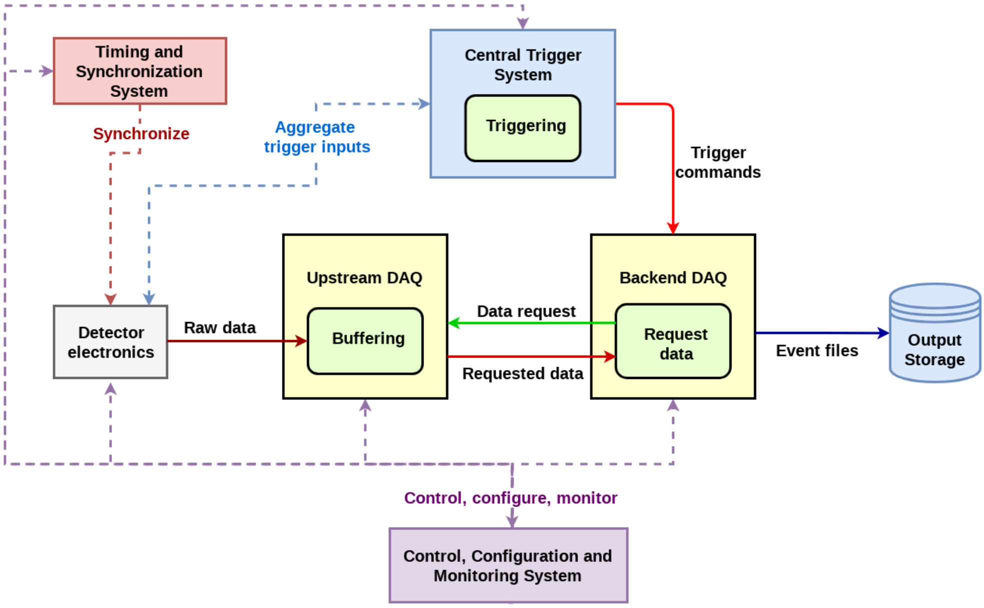


Figure 1. Overview of the ProtoDUNE-SP DAQ, highlighting the interfaces between different components and the flow of data in the system.

Unlike the prototypes, the DUNE Far Detector can not rely on external trigger information, therefore activity in the data needs to discovered in the form of so-called `trigger-primitives (TPs)`, described in section 3.1. A dedicated system, the data selection, is responsible for forming a trigger decision based on recognized patterns, specific signatures, and regions of interest based on the trigger primitives. After a trigger decision, a trigger command is sent to the back-end system that issues a data request for the readout units. This mechanism is referred to as the self-triggering chain, and an overview of it is seen in figure 2.

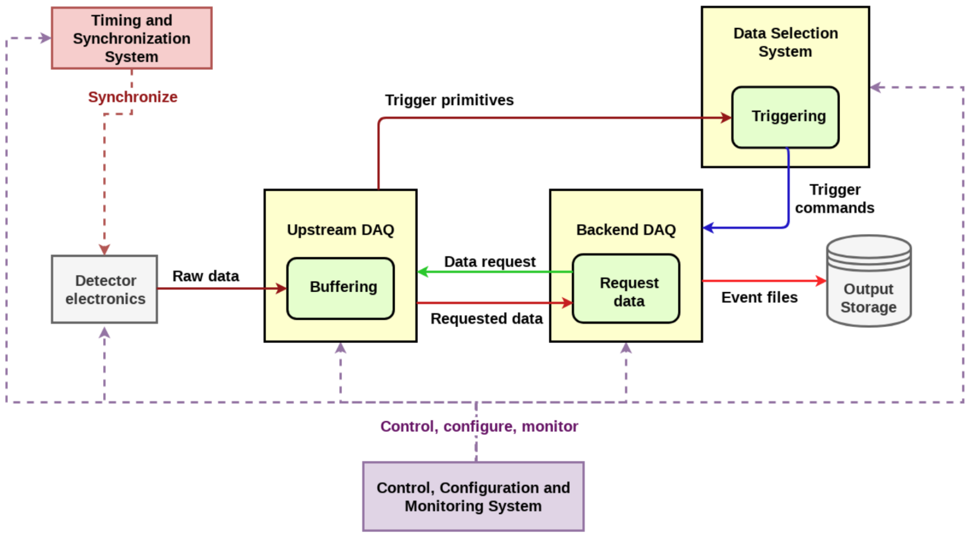


Figure 2. Overview of the DUNE DAQ system, highlighting the main differences from its prototype. The DUNE DAQ does not have a central trigger, but it is completely data driven and relies on the self-triggering approach.

2.1 TPC readout

Based on the experience of ProtoDUNE-SP, the baseline solution for the DUNE TPC readout system will be FELIX[4]. The application is described in previous contributions[5], and the topology of the setup is seen in figure 3.

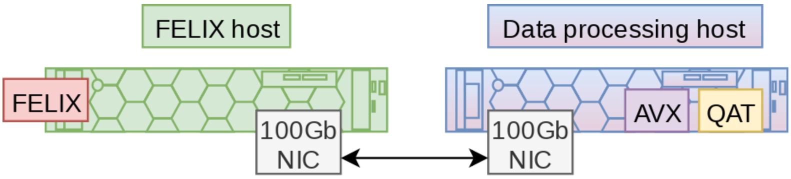


Figure 3. The publish-subscribe FELIX solution, initially applied in ProtoDUNE-SP's TPC readout.

One FELIX card receives data from a single APA, supplying ten links (10x ~1 GB/s front-end links), where each link corresponds to 464 B at 2 MHz (~0.88 GB/s) of user payload. As the front-end payloads are fragmented in the DMA memory array, serialization of the user payloads is required, which imposes a demanding memory I/O workload on the hosting server. In order to reduce the memory I/O operations of the hosting server, modest modifications of the FPGA firmware were introduced:

- *Jumbo DMA Blocks*: Increased DMA payload size from 1 KB to 4 KB reduces block write and interrupt rate of the card imposed on the host server.
- *Super Chunks*: Chunks are user payloads from the FELIX, and their aggregation reduces the required memory I/O rate. With 12 aggregated front-end chunks, the chunk rate is 5568 B (464 B × 12) at ~166 kHz ($\frac{1}{12}$ of the original rate) per link.

The data routing software is highly customized for ProtoDUNE: there is a scatter-gather implementation for collecting chunk pointers for detector data fragments and there is a single copy pipeline for serialization to user buffers. Data is published on Infiniband over Ethernet to the 100 Gb peer-to-peer connected data processing host. The subscriber applications are the so-called `BoardReader` processes. Each of these subscribes to a single link and buffers data, then extracts data fragments for trigger requests. It also carries out data reordering using *Advanced Vector Extensions*[6] (AVX2 and AVX512) registers and instructions in order to achieve better byte alignment for further data processing steps. Finally, every event fragment is compressed via hardware accelerated compression algorithm using the *Intel® QuickAssist* (QAT)[7] technology.

3 DAQ research and development

Despite the environment of ProtoDUNE-SP differing from the DUNE Far Detector, it provides an excellent development and integration test stand for DAQ applications and mechanisms. In 2019, the DAQ consortium organized development and integration sprints to introduce new features and functional elements that fulfill certain requirements and follow DUNE DAQ oriented design principles. We describe the most notable advancements that are integrated in ProtoDUNE-SP.

3.1 Self-triggering chain

One of the most crucial aspects of the DUNE DAQ is that it does not have an external global trigger: it must rely on the recognition of detector activity in the data. This implies the need for a quasi real-time data processing element that identifies so-called hits or `trigger-primitives (TPs)` in the data stream. From clusters of hits, so-called `trigger-candidates (TCs)` are formed for each APA. One can join up TCs to form TPC global signatures called `module-level triggers`, which are treated as final trigger decisions. In order to provide a proof of concept, a simple prototype of the self-triggering chain was implemented in ProtoDUNE-SP with the following software components:

1. The `HitFinder` module subscribes to the published FELIX links and performs pedestal subtraction by running a median calculation over the waveform. The algorithm resembles a high pass filter that filters out the low frequency component of the signal. As a next step, the module implements a Finite Impulse Response (FIR) filter[8] that averages the hits to exclude high frequency noise. A hit is defined by the resulting waveform reaching a configured threshold. The implementation of this module relies heavily on AVX2 instructions for optimized performance.

2. The `TriggerCandidate` finder module identifies contiguous groups of TPs in 50 μs time-windows in a single APA. The module generates a TC, if the channel range of the largest group is bigger than a configured parameter.

3. The `ModuleLevelTrigger` evaluates and merges TCs on an APA basis, and makes a sliding window of one drift time, and looks for a consistent direction between the TCs. A global trigger decision and data request is formed, in case a specified minimum number of hits are found in the merged TCs of the APA.

The data-flow software also required modifications to allow for the implementation of a software trigger module responsible for forming trigger candidates based on clusters of the found hits. The working solution was demonstrated with triggered cathode crossing muons as seen in figure 4.

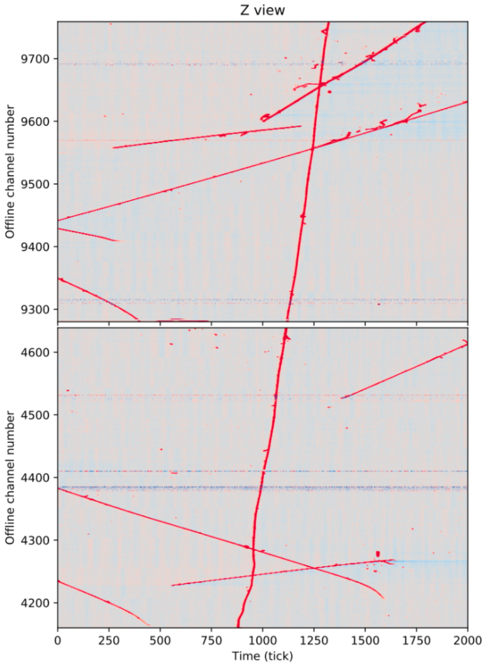


Figure 4. An example of a self-triggered event: a muon parallel to the face of the APA. The extracted window is centered around the timestamp that is contained in the trigger request.

3.2 Merged data routing and processing

One particular constraint for the readout units of the DUNE DAQ system is high density, meaning that a single commodity server should host at least 2 FELIX cards, each responsible for the data routing and processing aspects of the connected APAs. This requires combining the distributed routing and data processing onto one server, eliminating the current 100 Gb network connection between two separate servers. We successfully eliminated the peer-to-peer connection between the two applications, essentially merging the data routing software with the data selection (trigger-matching) implementation in the so called `OnHost BoardReader` process, as seen on figure 5. Two processes handle two logical units of a single physical FELIX card, due to details of the firmware implementation.

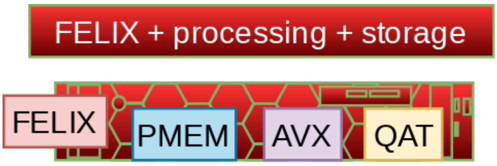


Figure 5. FELIX topology of DUNE favoring high density over routing features: Persistent Memory (PMEM) for buffering and storage, AVX for data processing, and hardware accelerated compression (QAT).

During this stage we performed extensive server evaluation on a handful of hosts with x86 architectures, also covering topology and PCIe riser configurations. BIOS and Linux kernel settings were heavily optimized for maximum memory throughput. This task was aided by performance profiling tools like *Intel® VTune™ Amplifier*[9] for optimizing NUMA allocation and thread affinity policies and the *Processor Counter Monitor (PCM)*[10] for verifying memory bandwidth utilization. As in this case data requests and transfers are handled by the same host, 10 Gb network optimization and interrupt moderation is required for optimal performance. Lastly, the hit-finding implementation was integrated into the `OnHost BoardReader`, resulting in an application that carries both high CPU utilization and memory I/O intensive threads. The total numbers of functional elements of a single readout unit are seen in table 1.

Table 1. Processes and threads on FELIX host server

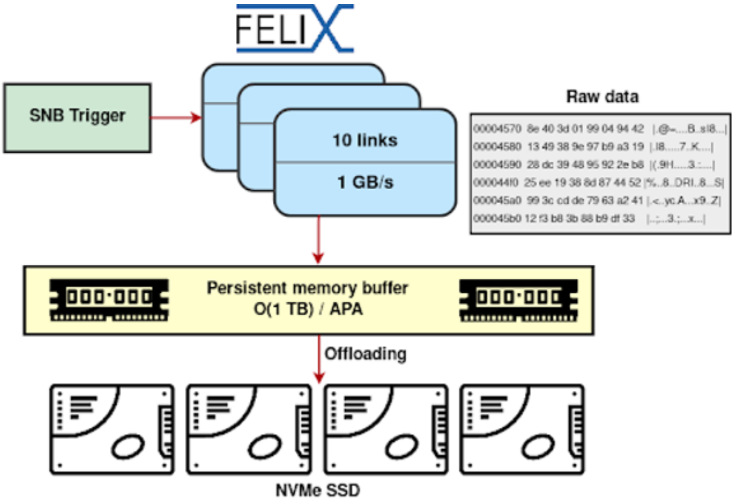
Element	Quantity	Detail
FELIX card and server	1 per APA	a single readout unit
OnHost readout process	2 per card	process per FPGA domain
DMA parser thread	2 (1 per process)	sensitive to NUMA locality
Link parser thread	10 (5 per process)	high memory I/O
HitFinder thread	10 (5 per process)	high CPU utilization
TriggerMatcher thread	10 (5 per process)	pinned to NIC interrupt cores

3.2.1 Firmware with hit-finding support

As the software implementation of the hit-finding procedure is CPU intensive due to the high rate of bitwise and byte operations, there is major interest in offloading this operation from the server's resources. The DUNE DAQ consortium dedicated substantial effort to extending the FELIX FPGA firmware with hit-finding capabilities. This includes similar algorithms and behavior as the software version, but fully implemented using FPGA resources in the FELIX system. Initial tests and integration attempts were made in ProtoDUNE-SP, which showed that individual building blocks are working correctly. The scaling requirements are not yet met, hence the solution is only working for a limited number of links. This is mainly due to the combination of very different logical firmware components, and time-domain requirements of individual blocks.

3.3 Supernova burst buffer

In order to fulfill the rich physics program of the DUNE experiment, the DAQ system will need to be carefully adapted for the study of SuperNova Burst (SNB) neutrino events. These are physics events expected to be produced with a rate of 2-3 per century. Detection of such rare signals is therefore extremely important. However, the signature of such events consists of distributed and low energy deposits which are hard to detect. In addition, considering the baseline DAQ architecture, the readout system is expected to generate 10 GB/s for each of the 150 readout units, resulting in a rate of 1.5 TB/s per detector module. It is foreseen that the SNB event will last for about 100 s and, thus, the storage system needs to buffer approximately 150 TB of data. The baseline solution is the extension of the FELIX readout cards with SSD buffers. Different solutions under evaluation include a dedicated ultra-fast distributed NAS, NVMe SSD RAIDs or buffering into persistent memory devices. The latter, using *Intel® Optane™ Persistent Memory* [11] technology has been tested as a possible candidate for the SNB buffer. After an initial characterization of the performance of the different storage media, the work for the SNB buffer with persistent memory has been focused on developing a realistic prototype with today's technologies. This is represented in figure 6.



```
00004570  8e 40 3d 01 99 04 94 42  |.@=....B..s18...|
00004580  13 49 38 9e 97 b9 a3 19  |.I8.....7..K....|
00004590  28 dc 39 48 95 92 2e b8  |(.9H.....3.:....|
000044f0  25 ee 19 38 8d 87 44 52 |%..8..DRI..8...S|
000045a0  99 3c cd de 79 63 a2 41 |.<..yc.A...x9..Z|
000045b0 12 f3 b8 3b 88 b9 df 33   |..;...3.;...x...|
```

Figure 6. One possible solution for storing the supernova burst events, based on persistent memory modules on the FELIX hosting server.

A basic application generating fragments at the required rate of 10 GB/s has been used to send the data to a testing platform of 6 TB of persistent memory. The results obtained are promising: it has been shown that it is possible to reach 80% of the target throughput with current-day technology. As a second stage, once the data has been buffered into memory, it can be offloaded to NVMe SSDs for longer term storage. In the future, more work will be dedicated to take advantage of all the features and optimizations provided by the persistent memory devices.

4 Summary

2019 was a successful year for DAQ R&D in ProtoDUNE-SP, with several test and integration periods along the year. Early on, the on-host data selection approach for the FELIX readout was implemented and used in operations for the rest of the year. After the readout software's capabilities were extended with the hit-finding components and features, the DAQ was able to collect and publish the found hits in this quasi-real-time environment. Extensions of the back-end data-flow software enabled the integration of the new self-triggering chain. Modules and their various configurations of the high-density readout approach were implemented and tested. Based on these results, the DAQ consortium will be able to draw conclusions on the direction of the various technologies under consideration, as well as continue the work on components that are identified as potentially advantageous to the readout system.

References

[1] B. Abi et al., *Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume IV: Far Detector Single-phase Technology* (2020), `2002.03010`
[2] B. Abi et al., *The single-phase protodune technical design report* (2017), `1706.07081`
[3] R. Sipos, IEEE Transactions on Nuclear Science **66**, 1210 (2019)
[4] G. Unel (ATLAS Tdaq), PoS **TWEPP2018**, 140 (2019)

[5] A. Borga et al., IEEE Transactions on Nuclear Science **66**, 993 (2019)

[6] *Introduction to Intel® Advanced Vector Extensions*, `https://software.intel.com/sites/default/files/m/d/4/1/d/8/Intro_to_Intel_AVX.pdf`

[7] *Intel® QuickAssist Technology (Intel® QAT)*, `https://www.intel.com/content/www/us/en/architecture-and-technology/intel-quick-assist-technology-overview.html`

[8] R. Oshana, in *DSP for Embedded and Real-Time Systems*, edited by R. Oshana (Newnes, Oxford, 2012), pp. 113 – 131, ISBN 978-0-12-386535-9, `http://www.sciencedirect.com/science/article/pii/B978012386535900007X`

[9] *Intel® VTune^{TM} Profiler*, `https://software.intel.com/content/www/us/en/develop/tools/vtune-profiler.html`

[10] *Processor Counter Monitor (PCM)*, `https://github.com/opcm/pcm`

[11] *Intel® Optane^{TM} Persistent Memory*, `https://www.intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory.html`