FELIX based readout of the Single-Phase ProtoDUNE detector

Andrea Borga[1], *Eric* Church[2], *Frank* Filthaut[1,3], *Enrico* Gamberini[4,*], *Paul* de Jong[1,5], *René* Habraken[1], *Giovanna* Lehmann Miotto[4], *Thijs* Miedema[3], *Frans* Schreuder[1], *Jörn* Schumacher[4], *Roland* Sipos[4], *Milo* Vermeulen[1], and *Lynn* Wood[2]

[1]Nikhef, Amsterdam, The Netherlands
[2]Pacific Northwest National Laboratory, Richland, Washington, USA
[3]Radboud University, Nijmegen, The Netherlands
[4]CERN, Geneva, Switzerland
[5]University of Amsterdam, Amsterdam, The Netherlands

Abstract. The liquid argon Time Projection Chamber technique has matured and is now in use by several short-baseline neutrino experiments. This technology will be used in the long-baseline DUNE experiment; however, this experiment represents a large increase in scale, for which the technology needs to be validated explicitly. To this end, both the single-phase and dual-phase implementations of the technology are being tested at CERN in two full-scale ($10 \times 10 \times 10$ m^3) ProtoDUNE setups. Besides the detector technology, these setups also allow for extensive tests of readout strategies.

The Front-End LInk eXchange (FELIX) system was initially developed within the ATLAS collaboration and is based on custom FPGA-based PCIe I/O cards in combination with commodity servers. FELIX will be used in the single-phase ProtoDUNE setup to read the data coming from 2560 anode wires organized in a single Anode Plane Assembly structure. With a sampling rate of 2 MHz, the system must buffer and process an input rate of 74 Gb/s. Event building requests will arrive at a target rate of 25 Hz, and loss-less compression must reduce the data within the requested time windows before it is sent to the experiment's event building farm.

This paper discusses the design of the system as well as first operational experiences.

1 Introduction

ProtoDUNE-SP [1] is the single-phase DUNE Far Detector prototype currently in operation at the CERN Neutrino Platform (NP) starting from September 2018 after two years of construction and commissioning. ProtoDUNE-SP represents a crucial milestone for the DUNE collaboration's effort towards the construction of the first DUNE 10 kton fiducial (17 kton total) liquid argon (LAr) mass Far Detector module. The detector design, construction and data acquisition solutions for DUNE are being validated with this prototype. With a total LAr mass of 0.77 kton contained in a $10 \times 10 \times 10$ m^3 cryostat, it represents the largest monolithic single-phase LAr Time Projection Chamber (TPC) built to date. It is housed in an extension

[*] e-mail: enrico.gamberini@cern.ch

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

to the EHN1 hall in the North Area, where CERN has provided for a new dedicated charged-particle test beamline. ProtoDUNE-SP has collected its first beam data in the first half of September 2018.

The ProtoDUNE-SP TPC, illustrated in Figure 1, comprises two drift volumes defined by a central cathode plane that is flanked by two anode planes, each at a distance of 3.6 m, and a field cage (FC) that surrounds the entire active volume. The active volume is 6 m high, 7 m wide and 7.2 m deep (along the drift direction). Each anode plane is constructed of three adjacent Anode Plane Assemblies (APAs) that are each 6 m high by 2.3 m wide. Each APA consists of a frame that holds three parallel planes of induction and collection wires on each of its two faces for a total of 2560 channels; the wires of each plane are oriented at different angles with respect to those making up the other planes of the same face to enable hit position pinpointing.

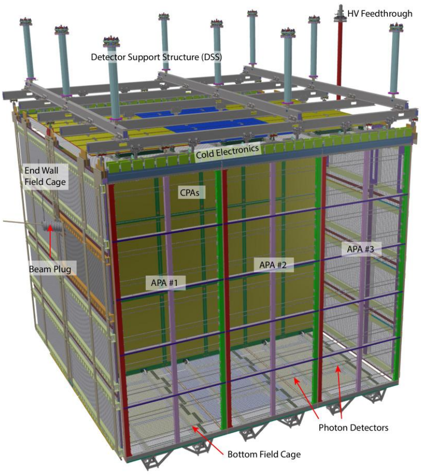

Figure 1. The major components of the ProtoDUNE-SP TPC.

The readout of the TPC wires consists of cold electronics (CE) mounted on the APAs inside the cryostat and the warm electronics outside the cryostat. CE data are received on the Warm Interface Boards (WIBs) which are situated on the top of flanges atop the cryostat. Each WIB multiplexes the data to four 4.8 Gb/s (or two 9.6 Gb/s) lines that are sent over optical fibers to the DAQ system. Two systems are used to receive data from the WIBs. The baseline system is based on Reconfigurable Computing Elements (RCE) [2] which are used to read out 5 of 6 APAs, while the alternative system described here is based on the Front-End LInk eXchange (FELIX) [3] technology and is used to receive the data from the remaining APA.

2 TPC data volume and structure

In the ProtoDUNE-SP context, the TPC readout systems must sustain the volume of data transferred from the WIBs. Each APA, outputting data over 2560 channels, is read out through 5 WIBs. The WIB sends data at a 2 MHz frame rate per optical link. Depending on the readout system receiving data, the number of links and line rate can have two combinations, based on different flavors of the WIB firmware: four 4.8 Gb/s links for the RCE readout or two 9.6 Gb/s ones for the FELIX readout. In the latter case, a single WIB frame represents a 500 ns time slice of 256 channels and its header contains a 20 ns granularity timestamp. Additionally, the frame is accompanied by a CRC-20 checksum generated by the WIB firmware. The checksum is evaluated by the FELIX firmware in order to verify the frame's data integrity; any CRC error is propagated with a flag in the data. Each frame contains 120 32-bit words, leading to a payload data rate of ~7.4 Gb/s; the additional overhead from 8b/10b encoding and protocol characters leads to a total line rate of 9.6 Gb/s. The FELIX readout system therefore receives data from 10 input links, corresponding to a total payload data rate of ~74 Gb/s.

The readout system must buffer the incoming data until a data request is received from the event building farm, and then transfer the data contained in a 3 ms time window, positioned at a programmable offset from the trigger timestamp. The DAQ system is designed for a target trigger rate of 25 Hz. An overview of the described readout system is seen in Figure 2.

3 FELIX-based readout

The primary driver of the FELIX concept is the assertion that a thin interface managing the interaction with detector front-end links and injecting data into commodity servers at an early stage of the DAQ chain provides the flexibility that is required for the optimization and maintenance of long-term and long lifetime systems. The FELIX I/O card provides a simple point-to-point interface to the detector front-end, supporting an 8b/10b encoded serial protocol at 9.6 Gb/s. Using the PCIe standard allows all data to be transferred to the host memory. This solution leverages the fast evolution of multi-core server performance, the possibility of using the sizable available host memory and the optimal choice of high-performance networking for data dispatching.

3.1 Topology of the readout system

The FELIX I/O card interfaces with its host PC through 16-lane PCIe Gen3. It transfers the incoming WIB data directly into the host PC's memory using a continuous DMA transfer accomplished through the Wupper [4] engine. The host PC runs a software process called *felixcore* [5], which publishes selected data to any client subscribing to it based on logical link identifiers. In ProtoDUNE, the clients to the *felixcore* application are the BoardReader processes, which are part of the *artdaq* [6] framework used for the ProtoDUNE DAQ data flow system.

From a hardware point of view, the FELIX and BoardReader hosts in use are based on dual socket Intel© Xeon© Processors (E5-2620 v4 2.1 GHz), 4×16 GB DDR4 memory banks driven at 2133 MHz and equipped with a Mellanox Technologies MT28800 Family (ConnectX©-5 Ex) 2×100 Gb/s NIC. The output of data from the BoardReader processes towards the DAQ event builders is carried out over 10 Gb/s Ethernet.

A single FELIX I/O card can receive data from an entire APA over ten links. Nevertheless, memory throughput tests showed that the used host architecture is not able to sustain the I/O from FELIX and to the network, due to memory bandwidth limitation. Therefore, the setup employed in the current data-taking is comprised of two FELIX cards hosted in two servers.

3.2 FELIX hardware and firmware

The interface card used in the ProtoDUNE setup is the latest version of the ATLAS FELIX hardware platform, named FLX-712. It is a standard height PCIe Gen3 card based on a Xilinx UltraScale FPGA (XCKU115) capable of supporting 48 bi-directional high-speed optical links via on-board MiniPOD transceivers.

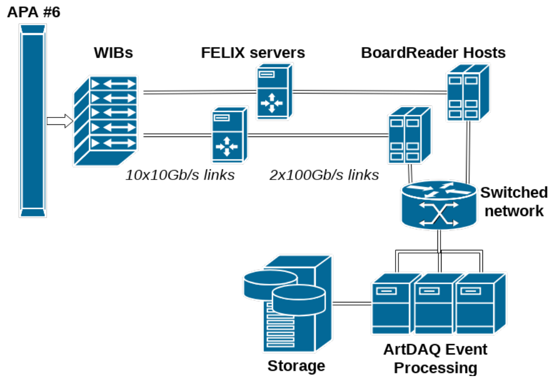

Figure 2. Overview of the FELIX data acquisition chain for ProtoDUNE-SP.

The FELIX firmware supports two modes: GBT mode (described in detail in Ref. [7]) and FULL mode. The latter uses a customized light-weight protocol for the *ToHost* path, providing maximum payload at a line rate of 9.6 Gb/s. As FULL mode uses 8b/10b encoding, a maximum user payload of 7.68 Gb/s can be achieved.

In the ProtoDUNE use case, the high rate of incoming frames (2 MHz) and the high throughput (74 Gb/s) represent a challenging load for the host concerning processing power and speed. Therefore, modest modifications to the original FULL mode firmware design were introduced in the *Central Router* module in order to sustain the flow of data (Figure 3):

- The incoming frames are grouped together in order to minimize the rate or processing operations and the number of network calls. These aggregated frames are transmitted as a single message via the networking library. The packing factor can be set to an arbitrary value.
- The size for each DMA transfer block between the FPGA and the host memory has been modified, allowing the optimization of frame parsing (minimizing the potential split between blocks) and serialization of data for network transmission.

The listed changes allow considerable simplification of the *felixcore* software running on the FELIX host, at the same time reducing CPU intensive operations on the host. The reduction can be appreciated directly from CPU usage statistics, as the interested cores are not saturated.

Increasing the packing factor and DMA payload size lowers the rate of intensive operations but, on the other hand, it increases the space and time requirements from the host memory access perspective, inducing backpressure on the firmware buffers. As a trade-off, a frame packing factor of 12 and a DMA payload size of 4 kB are currently used as the baseline, after having investigated different combinations.

4 DAQ software layer

The ProtoDUNE use-case has challenging requirements for the software layer, since it needs to operate at full data rate with trigger matching and loss-less data compression. In order to cope with the requirements, great care has been put into the implementation, with a main focus on avoiding dynamic memory allocation and memory copies of data fragments. Compression requirements introduce a challenge, as standard software implementations can't keep up, even with bare minimum trigger rates. Therefore, hardware features of the CPU for reordering the data to be byte aligned and dedicated accelerators for compression are utilized.

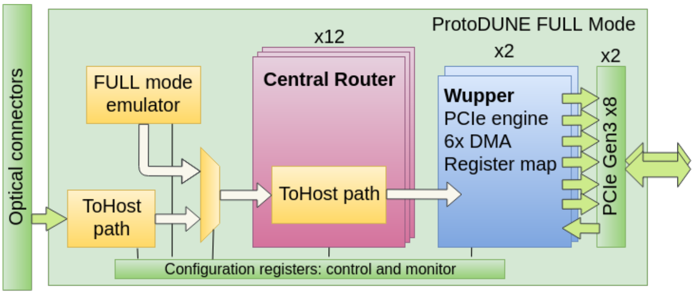

Figure 3. Main firmware logical blocks for 8b/10b data mode

4.1 The network layer

The *felixcore* application is in charge of routing data from the FELIX card to networked software clients and acts as the main publisher service of FELIX devices. It has been designed with the aim of being generic and unaware of the data that it routes.

The network layer is based on scatter-gather principles to collect fragments from the DMA buffer, which are serialized and published on 100 Gb/s Mellanox network interfaces, using the *Netio* [8] library. Two implementations exist:

1. POSIX sockets using *IOVEC* [9], supported by the Linux kernel,
2. RDMA over Infiniband solution, which relies on *libfabric* [10].

For additional performance gain, the latter was chosen for the final system.

4.2 BoardReader implementation

The processes receiving data from the publisher applications are implemented within the art-daq framework, which is used in ProtoDUNE-SP for the data flow system. In this framework, the applications connected to the detector readout are called BoardReaders. The FELIX-specific implementation relies on a network messaging layer that subscribes to the previously described FELIX publisher application. The solution focuses on flexibility, with the topology of the queues and links being scalable and configurable. In order to achieve the required performance, particular care has been put into the implementation, avoiding dynamic memory allocation as well as minimizing memory copies using unique pointers and move semantics. Every link has dedicated subscriber threads that populate Single-Producer Single-Consumer (SPSC) queues using the lock-free implementation from the Folly [11] library. Concurrency features from C++11 ensure the proper synchronization of the trigger matching threads and compliance with the internal state machine of artdaq.

The trigger matching mechanism is responsible for the extraction of data corresponding to the time window around a specific trigger timestamp. The BoardReader implements its own trigger receiver that directly processes the broadcasted trigger requests from the timing system. On each request, previously paused trigger-matcher threads are re-launched with a new task binding that carries the requested timestamp. The data in the circular buffer being continuous in time, the exact distance between the trigger window's position and the front pointer can be calculated. Until reaching the desired position, elements are popped at the queue's front pointer. The trigger window is then copied, reordered and compressed into an artdaq fragment and finally removed from the buffer. Lastly, the trigger-matcher threads are paused, waiting for the next trigger request. This thread-flow and the utilization of the circular buffer are shown in Figure 4.

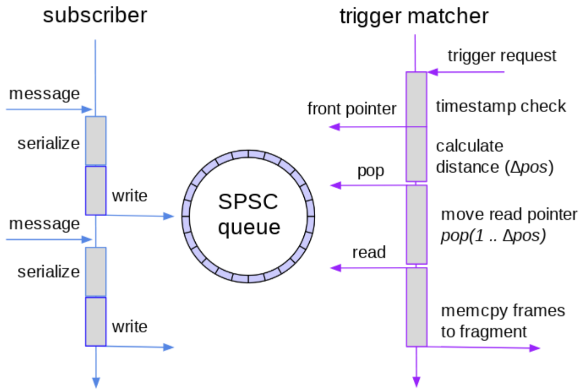

Figure 4. The logic of producer and consumer threads, and their utilization of the SPSC queue.

4.3 Reordering and compression

For unmodified frame-by-frame input data, the compression ratio of standard algorithms falls well below expectations. This is due in part to the non-contiguous storage of ADC data for individual channels in subsequent frames, but more importantly to the fact that all of the ADC values are cut up in the frame data. The BoardReader application therefore features a data reordering stage, where the ADC values are reconstructed into 16-bit words and reordered so that the ADC values for all of the subsequent digitization time slices are contiguous in memory. A solution that uses architecture-specific AVX2 and AVX512 [12] register instructions is evaluated and implemented, which provides additional performance gain.

Using simulated data with a noise RMS of 4 ADC counts [13], a compression factor of 3.7 is reached using standard compression algorithms. Compression of reordered fragments at high data acquisition rates requires hardware acceleration. Dedicated Intel® QuickAssist Technology (QAT) [14] adapter cards are used, which implement the DEFLATE algorithm, which is a combination of LZ77 lossless compression and Huffman coding. This allows for a reduction of the time required for the compression of one fragment's data to approximately 4 ms, compared to the software solution which takes more than 100 ms per fragment.

5 Performance tuning and integration results

The integration of the FELIX-based readout system in ProtoDUNE required an in-depth evaluation of the performance bottlenecks in the complete chain. The firmware and software optimizations result in better CPU utilization on the host side. Additionally, the software relies heavily on interrupt balancing and thread's CPU affinity in order to minimize context switching.

The allocation of the DMA buffer in memory also plays a crucial role in a multi-processor system. The allocation of the buffer to the same NUMA node where the FELIX I/O card is located allows copying data through the processor interconnect only once, upon transmission via the network. This is necessary as the PCIe Gen3 x16 slots required by the FELIX I/O and network cards are located on different NUMA nodes on the motherboard in use.

The FELIX-based readout system has been integrated as part of ProtoDUNE-SP DAQ system in the final topology described in Section 3.1. Ten BoardReader processes (one per WIB link from an APA) read out the requested events that are then merged in the ProtoDUNE-SP data stream and successfully reconstructed.

Tests have been carried out in order to understand the performances in terms of sustainable trigger rate and operation stability. The baseline trigger rate set for the ProtoDUNE-SP operation is 25 Hz. The FELIX-based readout system, and specifically the FELIX BoardReader, have been successfully tested for stability with reordering and compression up to 40 Hz random triggers (60 Hz without reordering and compression, limited by network bandwidth at the BoardReader hosts). Reordering is the most time expensive operation and therefore is the main limitation for trigger rate. It should be noted, nevertheless, that the FELIX system itself supports full readout on appropriate host architectures and that, if required, it would be possible to scale up the sustained trigger rate by distributing BoardReaders over more or more powerful hosts.

6 Summary

The ProtoDUNE-SP detector is a 770 ton LAr detector intended to validate the single-phase LAr Time Projection Chamber technology at the full scale of the DUNE experiment, and receiving beam from the CERN SPS accelerator since September 2018. One of its six Anode Plane Assemblies, representing 2560 anode wires, is read out using the FELIX system.

The FELIX readout system is based on the concept of having a thin interface between the front-end of a detector and commodity hardware. The current FELIX I/O cards receive 74 Gb/s of data over 10 links and use 16-lane PCIe Gen3 to copy it to the FELIX host PC's memory.

The input data rate can be sustained using a firmware and software modified from its original version used in the ATLAS experiment. Data frames are packed together in firmware and an optimized publisher software sends data to BoardReader processes running on separate hosts using Infiniband.

The BoardReader processes perform trigger matching and lossless data compression. The requirements on the compression, which is the most computationally demanding step, have been met by re-formatting the data in software and subsequently carrying out a hardware accelerated compression. Compressed data is subsequently forwarded to the artdaq-based event building framework.

The FELIX-based readout system is currently employed in ProtoDUNE-SP with great results in terms of sustained trigger rate and stability during data acquisition.

Acknowledgments

We thank our colleagues in the data acquisition group of the single-phase ProtoDUNE setup, within which this project is embedded, for their support in integrating the FELIX based readout.

We express our gratitude to the CERN *openlab* team, for providing support in the hardware accelerated compression achievements as well as to Intel, for having provided the hardware required for its development and benchmarking.

We gratefully acknowledge the assistance and support from the FELIX developers team of the ATLAS experiment, without whom this project would not have been possible.

Last but not least, our gratitude goes to the Neutrino Platform for the unconditioned support to the introduction of a FELIX based readout system for (Proto)DUNE, in terms of resources and promotion of the idea.

References

[1] B. Abi, R. Acciarri, M.A. Acero, M. Adamowski, C. Adams, D.L. Adams, P. Adamson, M. Adinolfi, Z. Ahmad, C.H. Albright et al., *The Single-Phase ProtoDUNE Technical Design Report* (2017), 1706.07081

[2] R. Herbst, R. Claus, M. Freytag, G. Haller, M. Huffer, S. Maldonado, K. Nishimura, C. O'Grady, J. Panetta, A. Perazzo et al., *Design of the SLAC RCE Platform: A general purpose ATCA based data acquisition system*, in *2014 IEEE Nuclear Science Symposium and Medical Imaging Conference, NSS/MIC 2014* (IEEE, 2016), pp. 1–4, ISBN 9781479960972

[3] J. Anderson, A. Borga, H. Boterenbrood, H. Chen, K. Chen, G. Drake, M. Dönszelmann, D. Francis, B. Gorini, F. Lanni et al., *A new approach to front-end electronics interfacing in the ATLAS experiment* (2016), Vol. 11, pp. C01055–C01055, ISSN 1748-0221

[4] A. Borga, R. Blankers, F. Schreuder, O. Kharraz, *Wupper: PCIe DMA Engine for Xilinx FPGAs*, https://opencores.org/project/virtex7_pcie_dma/overview

[5] J. Schumacher, F.P. Schreuder, A. Borga, H. Boterenbrood, K. Chen, W. Vandelli, H. Chen, J.T. Anderson, J. Vermeulen, L. Levinson et al., *Improving packet processing performance in the ATLAS FELIX project* (ACM Press, New York, New York, USA, 2015), pp. 174–180, ISBN 9781450332866

[6] K. Biery, C. Green, J. Kowalkowski, M. Paterno, R. Rechenmacher, *Artdaq: An event-building, filtering, and processing framework* (2013), Vol. 60, pp. 3764–3771, ISSN 00189499

[7] J. Anderson, K. Bauer, A. Borga, H. Boterenbrood, H. Chen, K. Chen, G. Drake, M. Dönszelmann, D. Francis, D. Guest et al., *FELIX: A PCIe based high-throughput approach for interfacing front-end and trigger electronics in the ATLAS Upgrade framework*, in *Journal of Instrumentation* (2016), Vol. 11, pp. C12023–C12023, ISSN 17480221

[8] J. Schumacher, C. Plessl, W. Vandelli, *High-Throughput and Low-Latency Network Communication with NetIO*, in *Journal of Physics: Conference Series* (2017), Vol. 898, p. 082003, ISSN 17426596

[9] *Scatter-Gather (The GNU C Library)*, https://www.gnu.org/software/libc/manual/html_node/Scatter_002dGather.html

[10] OpenFabrics, *Libfabric*, https://ofiwg.github.io/libfabric/

[11] Facebook, *Facebook Open-source Library*, https://github.com/facebook/folly

[12] Intel, *Intel Architecture Instruction Set Extensions Programming Reference* (2014), https://software.intel.com/sites/default/files/managed/c5/15/architecture-instruction-set-extensions-programming-reference.pdf

[13] R. Acciarri, C. Adams, R. An, J. Anthony, J. Asaadi, M. Auger, L. Bagby, S. Balasubramanian, B. Baller, C. Barnes et al., *Noise Characterization and Filtering in the MicroBooNE Liquid Argon TPC* (2017), Vol. 12, pp. P08003–P08003, ISSN 17480221, 1705.07341

[14] Intel, *Intel QuickAssist Technology*, https://www.intel.com/content/www/us/en/architecture-and-technology/intel-quick-assist-technology-overview.html