DUNE-doc-1846
version 3
October 27, 2016

TPC Readout using FELIX

Frank Filthaut

Abstract

One out of six ProtoDUNE TPC Anode Plane Assemblies will be read out using the FELIX system. This note presents a preliminary design of this FELIX-based readout.

1 The FELIX Project

The Front-End Link EXchange (FELIX) [1] is a project initially developed within the ATLAS Collaboration at CERN. Its purpose is to facilitate the development of high-bandwidth readout, needed for example for the experiment's new muon detectors to be installed in the upcoming LHC shutdown (starting at the end of 2018). The aim is for FELIX to be used in the majority of the experiment's readout after the subsequent shutdown preparing for the High-Luminosity LHC, presently planned to start in 2026. The motivation for FELIX is the desire to move away from custom hardware at as early a stage as possible, and instead employ commercial off-the-shelf PC-based hardware and networking.

The FELIX design is based on a shared firmware/software solution. A PCIe card is used to stream input data arriving from the detector front-ends to a circular memory buffer in a host PC using a continuous DMA transfer (with fixed 1 kB block size); software running on the host PC routes the data to multiple output destinations using network interface cards, as shown in Fig. 1 (taken from Ref. [1]).

1.1 Hardware

Three PCIe Gen-3 (8 or 16 lane) cards are presently being considered, all based on Xilinx FPGAs:

- The Xilinx VC-709 connectivity kit [2]. This 8-lane card features four SFP+ input links; external timing and trigger control (TTC) signals (as needed for LHC operation) can be provided through an FGPA Mezzanine Card (FMC) known as the TTCfx.

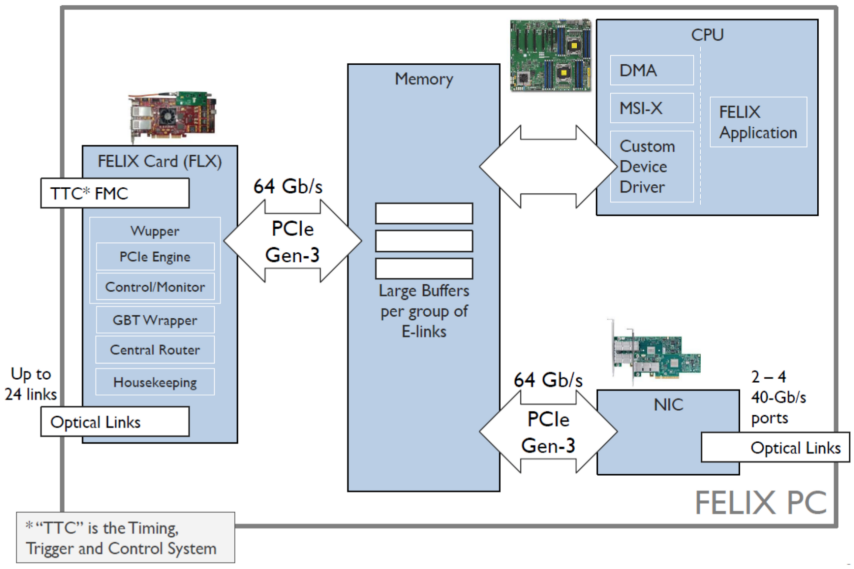

Figure 1: Block diagram of the FELIX system.

- The Hitech Global HTG-710 card [3]. It also has an 8-lane PCIe interface and features two CXP connectors, amounting to an equivalent 24 bidirectional links to the front-end electronics. TTC signals can again be handled using the same TTCfx FMC as used for the VC-709 card.

- The BNL-711 card [4] being developed at BNL. It is a 16-lane PCIe card with a total of 48 duplex optical links (MiniPODs). Handling of TTC signals will be integrated directly on the card.

The BNL-711 card is being developed as a baseline for LHC operation.

The host PCs being investigated are Supermicro server PCs (an example motherboard can be found as Ref. [5]), with either a single or two multi-core CPUs, and a total of 40 Gen-3 PCIe lanes per CPU. High-bandwidth network interface cards (NICs) can be obtained e.g. from Mellanox [6], with a NIC offering dual 40 Gb/s ethernet or 56 Gb/s Infiniband connectivity serving as a baseline.

1.2 Readout modes

Three readout modes have been defined for FELIX. Two of these make use of the bi-directional GigaBit Transceiver (GBT) protocol [7], developed for the readout of the radiation-hard GBTx ASIC [8]. The *GBT Normal mode* features a forward error correction mechanism intended to cope with radiation-induced data errors. This comes with a significant overhead: the 4.8 Gb/s link in these modes allows for a payload transfer rate of 3.2 Gb/s. In the *GBT Wide mode* (which does not use forward error correction) this is 4.48 Gb/s. The GBT protocol also allows for multiplexing many slower links ("e-links") on a single fast link.

The third mode is called *Full mode*. This is a mode featuring a simple 8b/10b encoding, without multiplexing of e-links, and is to be used only in environments where radiation induced data errors are not a concern. The input links in this mode are unidirectional and can run at 9.6 Gb/s, corresponding to a payload rate of 7.68 Gb/s.

The different readout modes affect the operation of the software. In the GBT modes, one thread is created for each input e-link; it streams the data to "subscribed" clients. In *Full mode*, one thread is created for each physical input link.

1.3 Status

The FELIX project is a collaboration between Argonne and Brookhaven national laboratories, CERN, Nikhef, UC Irvine, Radboud University, the Weizmann Institute, and Royal Holloway University of London. Hardware development is ongoing mainly on the BNL-711 card, a first version of which has been tested successfully in early 2016; a second version (with minor improvements and bug fixes) has been produced and is being tested [9], while a third version is to be produced shortly. Test stands exist at some of the collaborating institutes. In particular, at Nikhef equipment exists to test the full chain. The *Full mode* is fully specified; its firmware implementation is presently being tested in separate stand-alone firmware tests; an emulation is foreseen to allow for more integrated tests, on the timescale of the next months.

Data throughput limitations are being investigated. Recent tests of the BNL-711 card have established that its optical links can operate at speeds up to 12.8 Gb/s; a total PCIe throughput (i.e., for 16 lanes) of 101.7 Gb/s has been measured [9].

An extensive suite of software tools exists for board communication and control, data flow control and verification, and performance testing; all three of the cards mentioned above are supported. The FELIX development team has also started to support front-end test users.

2 Implementation in ProtoDUNE

The ProtoDUNE environment [10] differs from the LHC in a number of respects. First, the TPC signal is much slower than that in LHC experiments. On the other hand, the baseline is for the experiment to read out the TPC (at 2 MHz digitisation rate) without any zero suppression. With six anode plane assemblies (APAs), each containing 2560 channels, and using 12-bit ADCs, the total data flow from the detector is nevertheless very high, approximately 430 Gb/s. A further difference is that ProtoDUNE will not be subject to high radiation doses.

As a consequence mainly of the high costs associated with data storage, the following decisions have been taken:

- The data should be compressed, but in a lossless way. Previous experience [11] indicates that compression by a factor of 4–5 may be achievable.
- The experiment will not read out all its data but instead will use a trigger. For each trigger a window of 5 ms (the exact window size should be a configurable parameter) will be read out, and a trigger rate of 25 Hz (during 4.8 s long SPS spills) is foreseen.

The ProtoDUNE trigger and timing system is described in more detail in Ref. [10]. Timing is based on the concept of 64-bit *time-stamps* incrementing at the experiment's system frequency of 50 MHz. Synchronisation is relevant especially to the front-end boards inside the cryostat, through the Warm Interface Boards (WIBs) mounted on the outside of the cryostat, and will use direct links from the timing system to the WIBs. The trigger information will consist primarily of a time-stamp, which is communicated to the readout system either using a custom hardware interface or through network messages, depending on the specifics of the readout system.

2.1 Input data handling

Of the six ProtoDUNE APAs, the baseline solution adopted at present is to read out five using the RCE system [12], and one using FELIX. This one APA corresponds to 2560 channels, and a total data flow of about 75 Gb/s (not counting any overhead from 8b/10 encoding and ignoring idle data). These five WIBs will send out data in a slightly different format for FELIX than for RCE. In particular, each WIB (which combines the data from four front-end boards, corresponding to eight 64-channel COLDATA [13] ASICs) will have two 9.6 Gb/s links to FELIX using 8b/10b encoding, compatible with the FELIX *Full mode*. In *Full mode*, data must be transferred in

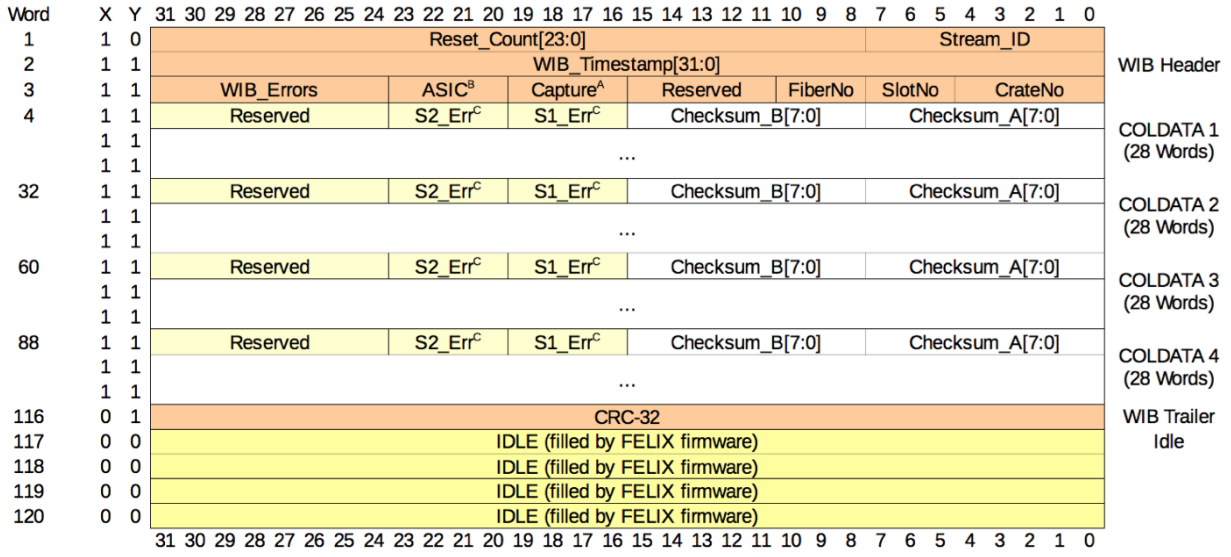

Figure 2: Format of the data sent from the WIB to FELIX. Two of the trailing four IDLE words will be used by the FELIX firmware.

multiples of 32-bit words. Figure 2 shows the draft specification (taken from Ref. [14]) of the 120-word frame corresponding to a single 500 ns time bin.

Figure 3 shows how FELIX could function in this context. Assuming that either the VC-709 or HTG-710 card is used, the 8-lane PCIe interface limits the throughput to 63 Gb/s, and therefore only six input links can be dealt with by a single FELIX system. With one APA corresponding to 10 input links, two or three systems will be required (two if one system uses a VC-709 card and one a HTG-710 card or if two VC-709 cards can be hosted by a single PC, three otherwise).

Figure 3 shows how one FELIX system could be incorporated in ProtoDUNE. More detail is provided below.

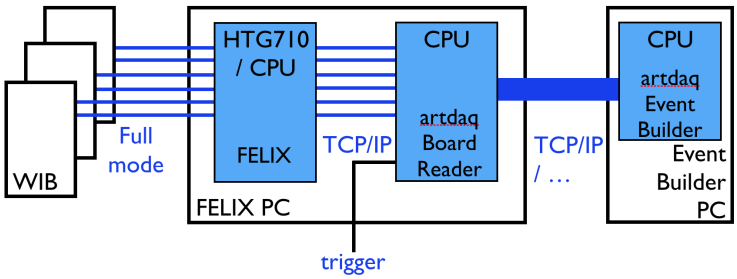

Figure 3: Block diagram showing envisaged FELIX functionality in ProtoDUNE. Only one FELIX system (with 6 input links) is shown.

2.2 Software backend

Communication with the event building system will use the *artdaq* [15] software, in the form of a *BoardReader* communicating with an *EventBuilder*. The BoardReader will package the data as *artdaq fragments*, where each fragment header will contain the trigger time-stamp as well as a *SequenceID*: an event counter to be reset for each run.

Since the *BoardReader* is a pure software process, it will be straightforward to configure it. In particular, the time window to be used for filtering, relative to the trigger timestamp, needs to be configurable. Any partitioning of the data (in which not all the ProtoDUNE data will be streamed to the same output) will be straightforward, as partitioning of the TPC data will be done by APA.

2.3 A possible software solution

An implementation that adapts to the existing FELIX interface is depicted in Fig. 3. Here, the BoardReader process is integrated with the required functionality: it receives data from the FELIX software back-end, filters the data according to the trigger time-stamps, and compresses the remaining data. The cheapest solution is for the BoardReader to run on the FELIX host PC; the TCP/IP connection between the FELIX software backend and the Board Reader can then make use of loopback communication, resulting in improved transfer speed. The trigger time-stamp information will be received in the form of network messages.

After filtering, sufficient CPU resources are likely to be available for the BoardReader (or perhaps yet another process) to compress the data. Unlike the case of the RCE system, where the data are compressed (wire by wire) in firmware, in this software environment it is not presently foreseen to compress the data on a wire by wire basis. Different algorithms will be investigated to determine what compression is achievable when acting on the data fragments arriving from the FELIX software backend.

A baseline approach is to implement the solution as described in Section 2.3 using two FELIX systems: one containing a VC-709 card and receiving data from two WIBs, and one containing a HTG-710 card and receiving data from three WIBs. In this setup, the network bandwidth to the EventBuilder will be well below 10 Gb/s, therefore allowing for the use of relatively cheap (dual) 10 Gb/s NICs.

2.4 Avenues under investigation

Owing to the versatility of the proposed setup, many of its parameters have not yet been determined. Several fall-back solutions and options under investigation are listed below:

Use of the BNL-711 card: This card, not being a commercial one, may end up not being (significantly) more expensive than the HTG-710 card. Due to its 16-lane Gen-3 PCIe interface, a single card may be able to deal with all input links. Consequences for the data throughput in the subsequent stages need to be worked out.

Application of trigger filtering at an earlier stage: In the setup described above, all of the data are routed from the FELIX application to the BoardReader process. It may be possible to reduce this rate in two ways:

1. Use the FELIX subscription mechanism to have the BoardReader subscribe only for the relevant time frames. A translation between subscription time and time-stamp will need to be developed for this alternative.
2. Apply the trigger filter directly in the FELIX software application. Such modification of existing software is not foreseen at present, and the software would need to provide a "hook" for the relevant (user) code to be added.

Compression in firmware: Rather than carrying out the data compression in a separate software process, it may be possible to do this in the FELIX firmware itself (on a wire-by-wire basis, as also done in the RCE system), thereby allowing to reduce the subsequent data transfer rate. This represents a more substantial change to the FELIX firmware than the preceding item, and therefore it would require extensive negotiation with the core FELIX development team.

Offloading the BoardReader process: It is conceivable that the BoardReader process as described above may not be able to handle both the trigger filtering and the subsequent compression. If this turns out to be the case, the versatility of the FELIX solution will make it possible (albeit at the expense of having to purchase additional hardware and possibly the more expensive dual 40 Gb/s NICs) to offload part of this functionality to another PC.

2.5 Status

The ProtoDUNE FELIX effort presently is a collaboration between CERN, Pacific Northwest National Laboratory (PNNL) and Nikhef. The required hardware (VC-709 and HTG-710 cards,

as well as host PCs) is available. Test setups exist at PNNL and Nikhef (the test setup at Nikhef is mostly shared with that of the core FELIX development team), and team members are gaining experience with the system.

References

[1] J. Anderson et al., *A new approach to front-end electronics interfacing in the ATLAS experiment*, JINST **11** (2016) C01055.

[2] https://www.xilinx.com/products/boards-and-kits/dk-v7-vc709-g.html.

[3] http://www.hitechglobal.com/Boards/PCIE-CXP.htm.

[4] K. Chen, *FELIX: a PCIe based high-throughput approach for interfacing front-end and trigger electronics in the ATLAS upgrade framework*, 2016. http://indico.cern.ch/event/489996/contributions/2211051/attachments/1344397/2026274/TWEPP2016_FELIX.pdf. Topical Workshop on Electronics for Particle Physics.

[5] https://www.supermicro.nl/products/motherboard/Xeon/C600/X10SRA-F.cfm.

[6] http://www.mellanox.com/page/products_dyn?product_family=119&mtag=connectx_3_vpi.

[7] P. Moreira, A. Marchioro, and K. Kloukinas, *The GBT: A proposed architecure for multi-Gb/s data transmission in high energy physics*, in *Proceedings of the Topical Workshop on Electronics for Particle Physics*, S. Claude, ed., p. , 332. 2007. https://cds.cern.ch/record/1091474.

[8] P. Moreira, S. Baron, S. Bonacini, O. Cobanoglu, F. Faccio, S. Feger, R. Francisco, P. Gui, J. Li, A. Marchioro, C. Paillard, D. Porret, and K. Wyllie, *The GBT-SerDes ASIC prototype*, JINST **5** (2010) C11022.

[9] A. Borga. Private communication.

[10] The DUNE Collaboration, *ProtoDUNE-SP Preliminary Technical Design Report*, 2016. https://docs.dunescience.org:440/cgi-bin/RetrieveFile?docid=1794&filename=protodune-sp-prelim-tdr-7oct.pdf&version=3. DUNE-doc-1794.

[11] MicroBooNE Collaboration. Private communication.

[12] G. Barr, M. Graham, et al., *ProtoDUNE RCE-based TPC readout system proposal*, 2016. https://docs.dunescience.org:440/cgi-bin/RetrieveFile?docid=1207&filename=protoDUNE-RCE-Proposal-v1.1.pdf&version=2. DUNE-doc-1207.

[13] H. Chen, T. Shaw, et al., *COLDATA Draft Specification*, https://docs.dunescience.org:440/cgi-bin/RetrieveFile?docid=415&filename=Draft%20COLDATA%20Specification_v7.pdf&version=8. DUNE-doc-415.

[14] E. Hazen, *ProtoDUNE WIB Output Data Formats*, https://docs.dunescience.org:440/cgi-bin/RetrieveFile?docid=1701&filename=ProtoDUNE_to_FELIX.pdf&version=1. DUNE-doc-1701.

[15] K. Biery, C. Green, J. Kowalkowski, M. Paterno, and R. Rechenmacher, *artdaq: An Event-Building, Filtering, and Processing Framework*, IEEE Trans. Nucl. Sci. **60** (2013) 3764 – 3771.