Prepared for submission to JINST

Development of a High Throughput PCIe Card for DAQ System in the ATLAS and DUNE Experiments

K. Chen,[a,1,2] **S. Tang,**[a,2] **H. Chen**[a] **and F. Schreuder**[b]

[a] *Brookhaven National Laboratory, PO Box 5000, Upton, NY 11973, USA*

[b] *Nikhef National Institute for Subatomic Physics / University of Amsterdam, Science Park 105, 1098 XG Amsterdam, Netherlands*

E-mail: chenkai@ccnu.edu.cn

Abstract: In the Run 3 upgrade of ATLAS experiment and the Single-Phase ProtoDUNE (Prototype for the Deep Underground Neutrino Experiment) experiment, the FELIX (Front-End LInk eXchange) system has been deployed as the interface between front-end electronics and common Data Acquisition (DAQ) systems. Based on a PCIe card hosted in commodity servers, FELIX's flexibilty means it has also been adopted by other experiments, such as sPHENIX and CBM. The same PCIe based architecture is proposed for use in the ATLAS HL-LHC (High Luminosity Large Hadron Collider) upgrade and the DUNE experiment. To this end, the next generation of FELIX I/O card has been developed. It supports 25+ Gbps high speed fiber optical links and 16-lane Gen4 PCIe interface. There is also an on-board DDR4 module to buffer event data for DUNE experiment. This paper will report on the test results for the first demonstrator prototype of this next generation board, with which all major functions have been successfully evaluated.

Keywords: ATLAS; DUNE; FELIX; PCIe Gen4; High-speed links

ATL-DAQ-PROC-2021-004
19 June 2021

[1] Corresponding author.

[2] The first two authors contributed equally to this work.

Contents

1 Introduction

The data acquisition (DAQ) system plays a crucial role in nuclear and particle physics experiments. It reads out all data from the physics events occurring in the detector, and controls on-detector electronics. In recent years, new detector developments and data taking environments have significantly increased the functional and performance requirements which DAQ systems must satisfy, for instance the overall number of channels and density, link speed and throughput.

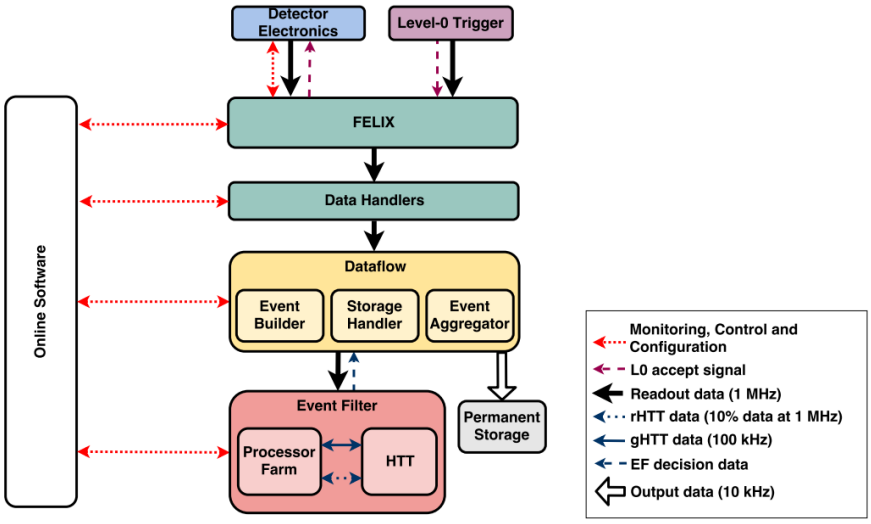


Figure 1: DAQ system in the ATLAS experiment for Run 4 [2].

ATLAS [1] is one of the two general-purpose particle physics experiments at the Large Hadron Collider (LHC). To boost the potential for new physics discoveries, the LHC will be upgraded to HL-LHC (High Luminosity LHC) around the year 2027, when the peak luminosity (collision intensity) will increase by a factor of 2.5 to 3.75. The luminosity increase will lead to more collisions in each bunch crossing (known as pile-up) from around 60 to 200, significantly increasing the size and complexity of events. To cope with the pile-up and read out all interesting event data, the

DAQ system [2] needs to be flexible, scalable, and closely integrated with the trigger system. As shown in Figure 1, Front-End LInk eXchange (FELIX) [3, 4, 6] distributes trigger, clock and control signals to the front-end on-detector electronics. These signals are embedded in the optical signal transmission and recovered by the front-ends. In the opposite direction, the data will be transmitted at a rate of 42 Tbps (about 5 MB per event) from front-end electronics and trigger system to the DAQ system through approximately 16,500 fiber optical links. The received data is then transmitted to the Data Handler system over a high performance network, where formatting, monitoring and other subdetector specific processing will take place. After this stage, data fragments are sent to the Dataflow system over a high performance network. The Dataflow system consists of a large distributed storage system known as the Storage Handler, upon which applications operate to further process the data. The Event Builder is responsible for logically assembling and managing complete event records. The Event Filter can then sample these data for the next stage of trigger selection. The Event Aggregator process collects, formats and transfers the final selected event data to CERN permanent storage [2]. The Data Handler and Dataflow systems consist of a farm of COTS (Commercial Off-The-Shelf) computers and network switches. The data can be processed by them soon after FELIX in the readout chain.

The Deep Underground Neutrino Experiment (DUNE) [8] is a neutrino beam experiment under construction at Fermilab and Sanford Underground Research Facility. The far detector will be 1.5 km under ground, and consist of four 10 kton modules. At least one of the modules will be instrumented with single phase Liquid Argon Time Projection Chamber (LArTPC) [9]. The data from the on-detector electronics will be streamed out via 1,500 fiber optical links, with a throughput of 14.4 Tbps. The DAQ system needs to buffer the compressed normal beam operations data, as well as a special big buffer for Supernova Neutrino Bust (SNB) data. Figure 2 shows the proposed deployment of FELIX for DUNE. The modules labeled 'Upstream DAQ' are commodity servers hosting FELIX I/O cards that interface the on-detector electronics to the event builder and control, configuration and monitoring systems. FELIX also performs low level data selection by analyzing data from front-ends, and outputs trigger candidates to the module level trigger generator. The module level trigger generator evaluates and merges these inputs and forms the global trigger decision and data request.

PCIe-based readout has been a paradigm shift adopted by many LHC experiments such as ALICE [10], ATLAS and LHCb [11]. FELIX [4] is a high-throughput Gen3 PCIe based readout system developed for the ATLAS Run 3 upgrade. As shown in Figure 3, it is a commercial server with PCIe input/output (I/O) cards (known as BNL-712) [6] and network interface cards. FELIX replaces previous custom and divergent detector readout technologies with a single common platform, based on one custom component integrated with commodity compute and network hardware. The only custom hardware in the DAQ system is the BNL-712. The architecture has been widely used by nuclear physics and particle physics experiments like sPHENIX [12], CBM [13] and ProtoDUNE-SP (Single-Phase Prototype for the Deep Underground Neutrino Experiment) [14], each time integrating with the DAQ frameworks of these experiments.

To meet the new requirements of the ATLAS HL-LHC upgrade and the DUNE experiment, a new generation of PCIe card is being developed. The main improvements include: doubling the bandwidth of the PCIe bus by upgrading to Gen4; increasing the supported maximum line rate of the fiber optical links to be higher than 25 Gbps; upgrading the FPGA from Kintex Ultrascale to Virtex

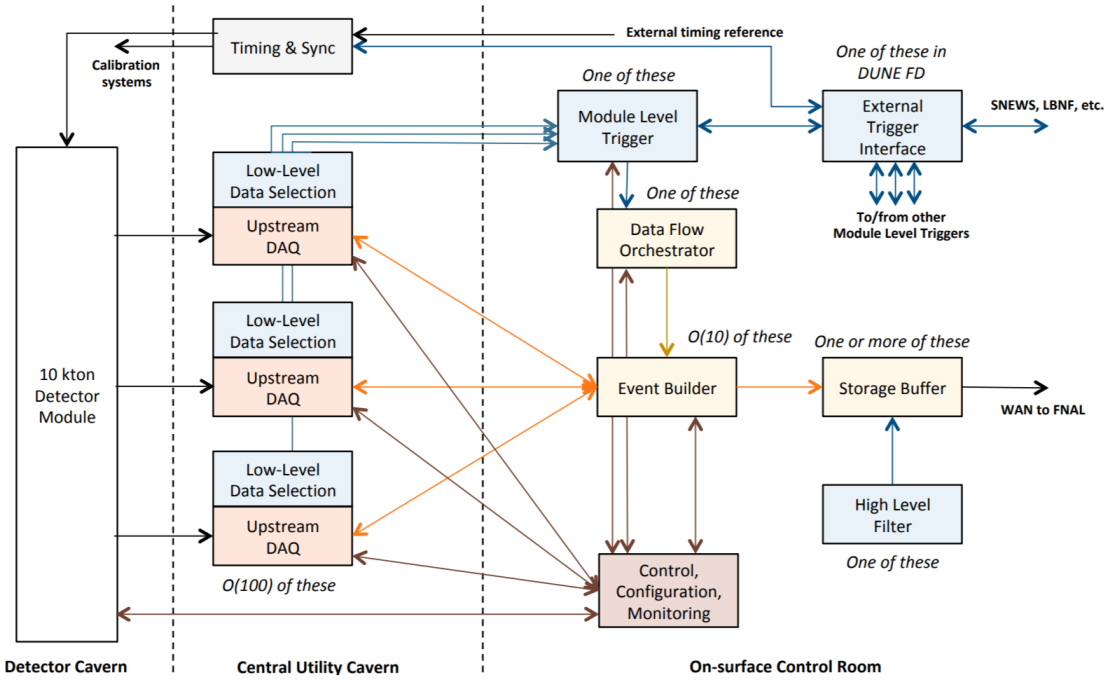


Figure 2: DAQ conceptual design overview for a single 10 kton single phase LArTPC module. Upstream DAQ modules hosting FELIX cards in the Central Utility Cavern (underground) interface with electronics on the detector, and route data to the rest of the DAQ system, located on the surface [9].

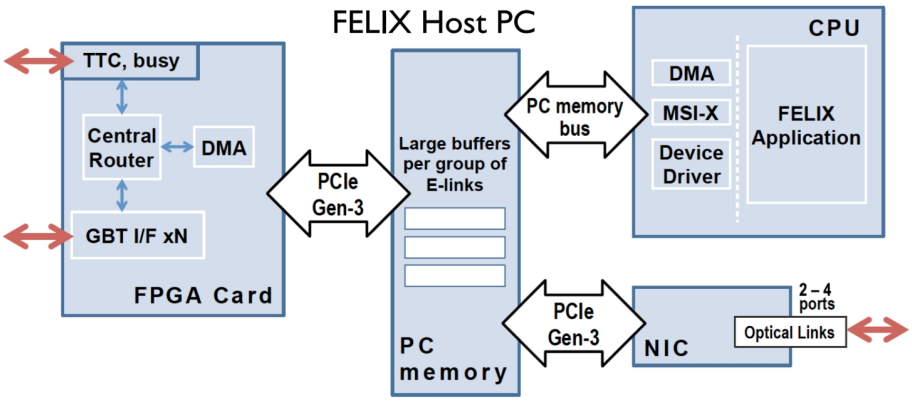


Figure 3: FELIX system with PCIe I/O card and NIC card [4].

Ultrascale+, to increase the FPGA logic capacity to support more sophisticated firmware designs; the DDR and/or SSD for possible on-board data buffering in DUNE experiment. A demonstrator card has been developed to evaluate these new functions. In the following sections, design and performance results from the testing of this new demonstrator will be discussed in detail.

2 Hardware Design

As shown in Figure 4, the demonstrator card was equipped with a Xilinx Virtex Ultrascale+ FPGA VU9P [5]. Compared to the FPGA on the previous FELIX card BNL-712 [6], it has 1.8 times more system logic cells, and 270 Mb extra UltraRAM [7] memory. The integrated 104 high-speed duplex GTY transceivers support a line rate up to 32.75 Gbps.

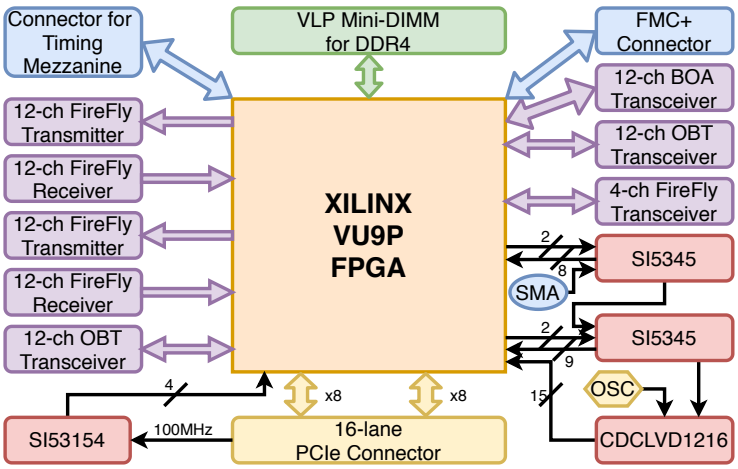


Figure 4: Connections between FPGA and other main components.

The PCIE4 block in the VU9P FPGA supports version 0.5 of the PCIe Gen4 standard [15]. The bandwidth of the PCIe bus is doubled compared to previous BNL-712 card using a 16-lane PCIe Gen3 bus, which helps to improve the integration density and reduce the quantity of cards and servers. As shown in Figure 4, the PCIe edge connector is connected to two PCIe hardcores in the FPGA, with 8 lanes each. When integrated with motherboards supporting PCIe bifurcation, the server will identify two 8-lane endpoints. The data transmission between FELIX and the new ATLAS trigger electronics may use high speed fiber optic links of line rates up to 25.78125 Gb/s, using either the Aurora or Interlaken protocols. To evaluate and compare the performance of optical modules from different vendors, the board has been designed to host several modules. There are two pairs of 12-channel 25G FireFly transmitter and receiver modules from Samtec, which support copper module ECUE or optical module ECUO [16]. There is also one 4-channel duplex 25/28G FireFly module. Besides these FireFly modules, there are one 12-channel duplex 25G Board-mount Optical Assembly (BOA) module [17] from II-VI and two 12-channel duplex 25G OBT module [18] from Amphenol-FCI. There are two different footprints for the BOA modules, one of them is compatible with the OBT module, as shown in Figure 7. The SI5345 jitter cleaner and CDCLVD1216 clock buffer chip are used to clean the system clock and provide good quality reference clocks for all the high speed transceivers. The design includes a VLP (Very Low Profile) Mini-DIMM (Dual In-line Memory Module) connector to support DDR4 memory, which can be used for data buffering with a capacity up to 16 GB. The board is designed with 2 high-density connectors, which are reserved for two mezzanine cards. One mezzanine will provide an interface to the timing and trigger system of the ATLAS experiment. The other one is a FMC+ mezzanine card with SSD modules for SNB data buffering in DUNE experiment. The mezzanine architecture

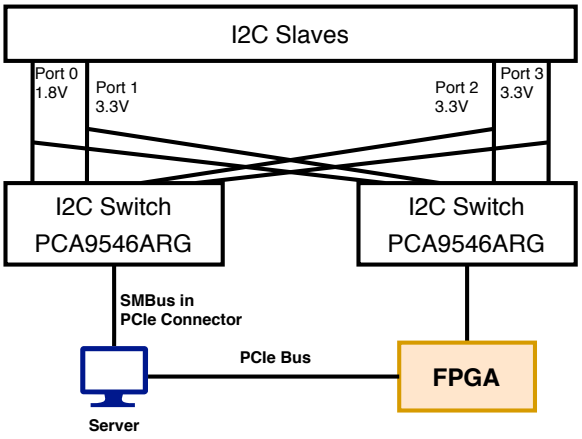


Figure 5: The I2C buses for device configuration.

will make the demonstrator board suitable for different experiments with minor changes to the mezzanine design.

There are many devices with I2C interfaces on the board, for instance all optical modules, the jitter cleaners and the GPIO expander. As shown in Figure 5, to make the communication with I2C devices more flexible, all are connected to the ports of two I2C switch chips (PCA9546ARG). The devices can be controlled by software via the PCIe interface through the I2C master implemented in the FPGA firmware. Software running on the server can also communicate with these devices via SMBus (System Management Bus [21]) through the PCIe connector. This control path does not rely on operational status of the FPGA. To avoid contention, these two I2C switch chips can not be configured to connect with the same port at the same time.

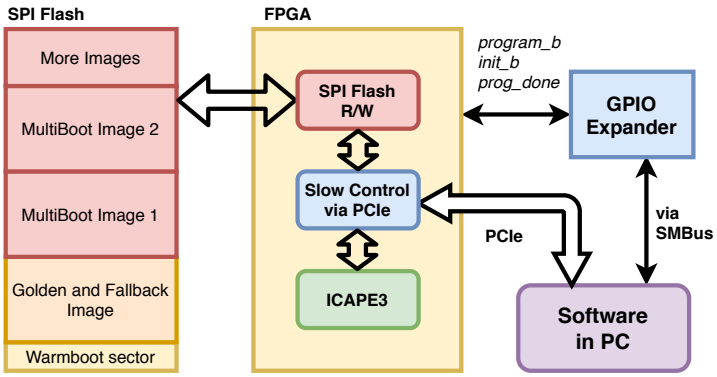


Figure 6: FPGA firmware version control with SPI flash and ICAPE3 core.

The uncompressed configuration file for the VU9P FPGA is about 641 Mb [19]. Two 2 Gb SPI flash memories are used to store multiple firmware versions. As shown in Figure 6, the SPI flashes can be programmed with the Warmboot sector [20] and multiple bit file images, through JTAG or the SPI flash controller module in firmware. The default image is selected by value in the Warmboot sector. To reload the default firmware, one option is to use SMBus and the on-board

GPIO expander to control the pin related to FPGA configuration, for instance the 'program_b' to trigger the firmware loading process. This method does not rely on the operational status of FPGA. Another method is to control the Xilinx ICAPE3 core [19] via the PCIe interface, to load firmware from any selected flash address. The ICAPE3 core provides post-configuration access to the configuration functions of the FPGA from the FPGA logic.

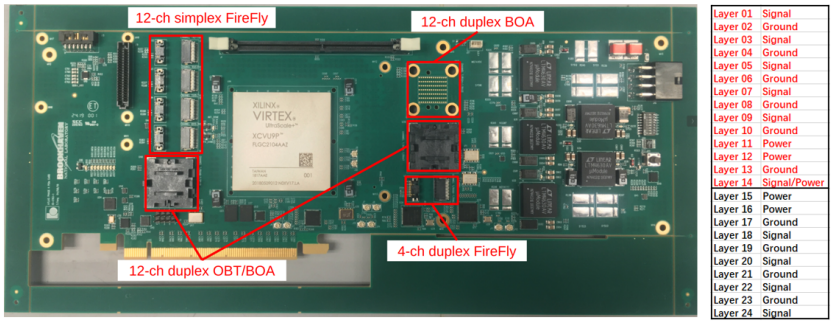


Figure 7: The demonstrator card with components assembled, and the stackup of the PCB.

The board is fabricated using Megtron 6(G) [22] as the base PCB material. As shown in Figure 7, there are two sub assemblies for the 24-layer 2.794 mm thick PCB. One is from layer 1 to layer 14 with thickness of 1.58 mm. The other one is from layer 15 to layer 24. The PCB around the PCIe edge connector uses only layer 1 to 14, in order to guarantee the thickness of this connector meets the PCIe mechanical specification. To guarantee good signal integrity for the high speed signals with GTY transceivers, 3 different kinds of backdrill are implemented to reduce the via stubs of theses traces [23]. Figure 7 shows one of the assembled boards with frame. The optical modules are placed close to the FPGA, to reduce the attenuation of high speed signals.

3 Performance Evaluation

A server with Gigabyte X570 Aorus Pro motherboard [24] and AMD Ryzen 3700X CPU was used to measure the throughput of data from board to host server. When bifurcation of the 16-lane slot is enabled and configured to be 8+8, two 8-lane Gen4 endpoints can be identified by the system. In firmware, data are continuously pushed into the two PCIe DMA cores. The software can see a consistent 12.6 GBps data stream from the two endpoints in parallel. The total throughput of 25.2 GBps is about 78% of the nominal speed for 16-lane PCIe Gen4 bus. The relative percentage is similar to that of the Gen3 FELIX card (BNL-712) [6].

The DDR4 module (A4K16Q28BNPBSE) from ATP Electronics Incorporation was installed on the board to measure the timing margin for data reading and writing. The Memory Interface Generator (MIG) core from Xilinx was implemented with a speed of 2.1 GTps. Figure 8 shows the margin of writing operations with complex pattern. It indicates that the margin is bigger than 65% Unit Interval (UI, 476 ps for 2.1 GTps) for all data bytes. The Byte 5 is with minimum margin of 68.6% UI, which is much bigger than the 30% UI mask required by Xilinx [25].

The PCIe bus and high speed optical modules are connected to GTY transceivers. The Xilinx IBERT (Integrated Bit Error Ratio Tester [26]) IP was used to measure the eye diagrams and Bit Error Rates (BER). For Gen4 PCIe, the transceivers operate at 16 Gbps. The BER test is carried

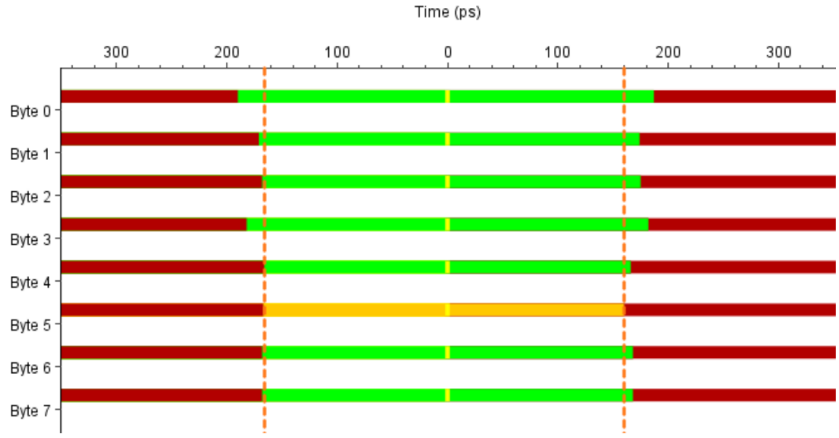


Figure 8: Timing margin for the DDR4 writing.

Table 1: Eye diagram scan results for optical modules.

Module	Line rate (Gbps)	Average Open Area	Average Open UI
Samtec FireFly	27.8125	8672	63.89%
Samtec FireFly	25.625	9600	66.67%
Amphenol-FCI OBT	25	9325	61.58%
Finisar BOA	25	7739	49.07%

out with the PCIe loopback card designed by WHIZZ SYSTEMS [27]. No errors occurred during the test with a PRBS-31 data stream, for transmission of more than 10^{15} bits. Similarly, the FireFly optical and copper modules, BOA and OBT optical modules are evaluated with speeds of 25 to 28 Gb/s. No errors occurred for the tested modules with transmission of more than 10^{15} bits. Table 1 shows the average results for eye diagram scans with modules in question. It should be noted that the 12-channel FireFly transmitter and receiver are still in the Alpha stage of development. The numbers listed in the table are for the 4-channel duplex optical transceiver module. Figure 9 shows the eye diagram for one channel in the FireFly module with line rate of 27.8125 Gb/s. The BER setting for the scan is set as 10^{-9}. The open area of the eye diagram is 8512, with an open UI of 66.67%, which is much bigger than the typical requirement of 35% UI.

The system clock of the FELIX card in the current ATLAS experiment is provided by the Timing, Trigger and Control (TTC) system [28]. The on-board jitter cleaner (SI5345) will clean the clock recovered from the TTC signal and generates reference clocks for the high speed transceivers. Figure 10a shows the phase noise of the 40 MHz clocks measured by a Signal Source Analyzer (SSA) for frequency range from 1 Hz to 5 MHz. One clock is from the TTC system, the other one is the output of SI5345. The digital loop bandwidth of the DSPLL [29] in the SI5345 is configured to be 640 Hz. As shown in the plot, noise for frequencies higher than 1 kHz is clearly suppressed. Figure 10b shows the phase noise of the 240 MHz output clock from the SI5345, which is a typical reference clock for the transceivers. The measured range is up to 100 MHz. Xilinx FPGAs have requirements for reference clock phase noise. The phase noise masks [30] at

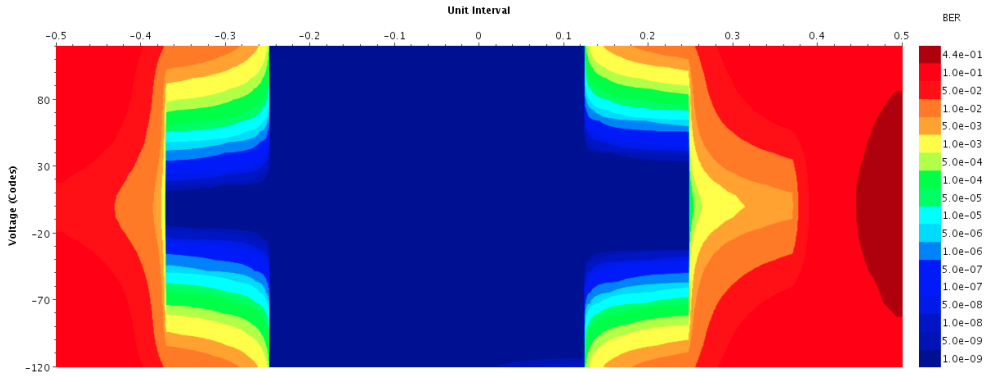
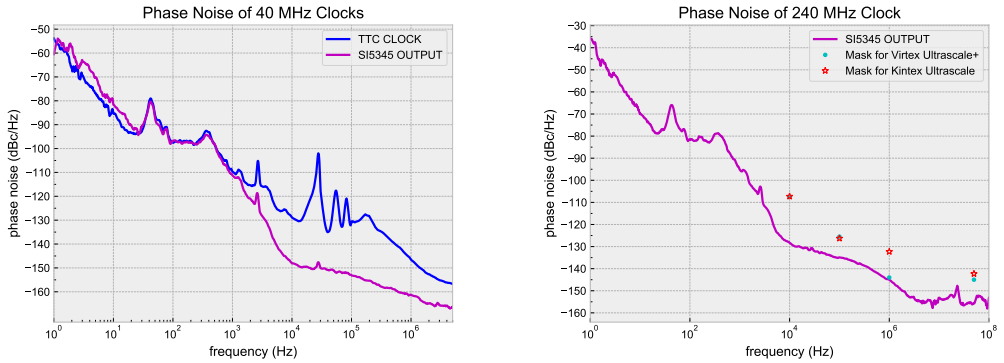


Figure 9: The eye diagram for one channel in the 4-ch duplex FireFly module with 27.8125 Gpbs line rate.



(a) Comparison of the phase noise of 40 MHz clocks: TTC clock and SI5345 output.

(b) Phase noise of the 240 MHz SI5345 output clock.

Figure 10: Phase noise measured by signal source analyzer.

several frequency points (10 kHz, 100 kHz, 1 MHz and 50 MHz) for Kintex Ultrascale and Virtex Ultrascale+ series FPGAs are marked in the figure. Cleaning with the SI5345 brings the phase noise within the requirements of both series. Jitter can be measured by an SSA or can be calculated from the phase noise measurement result. Over the frequency range relevant for the FPGA, 1 kHz to 100 MHz, the jitter is 0.46 ps. If considering a more general frequency range from 100 Hz for most electronics [31], it is about 1.97 ps. When the loop bandwidth is adjusted to be less than 40 Hz, the jitter from 100 Hz to 100 MHz can be decreased to around 0.5 ps.

4 Summary and Outlook

A demonstrator for the PCIe Gen4 version of the FELIX I/O card for ATLAS and DUNE experiments has been developed and evaluated. The FPGA multiboot capability and the control of all on-board I2C devices have been verified. The throughput of the PCIe bus is measured to be about 25.2 GBps from card to host server. The timing margin of the DDR4 memory reading and writing operations is close to 70% of the UI for a speed of 2.1 GT/s. Several different 25+ Gbps optical modules were

evaluated with PRBS-31 data transmission. No errors occurred for all modules with transmission of more than 10^{15} bits. The eye diagrams and open UI are much bigger than the requirements from Xilinx. Based on the performance comparisons undertaken, the 12-channel FireFly modules will be promising candidates for future prototype and production FELIX boards, if their performance is comparable with the 4-channel duplex FireFly module. Xilinx has released the Versal Prime series FPGA, which is fully compliant to the 4.0 PCIe specification. The PCIe core can support 16 lanes directly. A new version prototype FELIX I/O card based on Versal Prime FPGA [32] is undergoing functional and performance tests. Xilinx has already shipped Versal Premium series FPGAs to some customers. This newer series has more transceivers and also supports PCIe Gen5 standard, which can double the bandwidth between the card and the server. It is also a possible candidate for future card development.

References

[1] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003.

[2] ATLAS Collaboration, *Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System*, https://cds.cern.ch/record/2285584

[3] J. Anderson et al., *FELIX: a High-Throughput Network Approach for Interfacing to Front End Electronics for ATLAS Upgrades*, *J. Phys.: Conf. Ser.* **664** (2015) 082050

[4] J. Anderson et al., *FELIX: a PCIe based high-throughput approach for interfacing front-end and trigger electronics in the ATLAS Upgrade framework*, *JINST* **11** (2016) C12023.

[5] Xilinx Inc., *UltraScale+ FPGAs Production Tables and Product Selection Guide*,[Online]. Available at: https://www.xilinx.com/support/documentation/selection-guides/ultrascale-plus-fpga-product-selection-guide.pdf

[6] K. Chen et al., *A Generic High Bandwidth Data Acquisition Card for Physics Experiments*, *IEEE Trans. on Instrum. and Meas.* , Vol. 69, no. 7, pp. 4569-4577, July 2020

[7] Xilinx Inc., *UltraRAM: Breakthrough Embedded Memory Integration on UltraScale+ Devices*,[Online]. Available at: https://www.xilinx.com/support/documentation/white_papers/wp477-ultraram.pdf

[8] DUNE Collaboration, *Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume I: Introduction to DUNE*, arXiv:2002.02967

[9] DUNE Collaboration, *Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume IV: Far Detector Single-phase Technology*, arXiv:2002.03010

[10] ALICE Collaboration, *The ALICE experiment at the CERN LHC*, *JINST* **3** (2008) S08002.

[11] LHCb Collaboration, *The LHCb Detector at the LHC*, *JINST* **3** (2008) S08005.

[12] PHENIX Collaboration, *An Upgrade Proposal from the PHENIX Collaboration*, arXiv:1501.06197, 2015.

[13] V. Friesea, *The CBM experiment at GSI/FAIR*, *Nuclear Physics A*, Vol. 774, pp. 377-386, 2006.

[14] A. Borga et al., *FELIX-Based Readout of the Single-Phase ProtoDUNE Detector*, *IEEE Trans. on Nucl. and Sci.*, Vol. 66, no. 7, pp. 993-997, July 2019

[15] Xilinx Inc., *UltraScale+ Devices Integrated Block for PCI Express v1.3*,[Online]. Available at: https://www.xilinx.com/support/documentation/ip_documentation/pcie4_uscale_plus/v1_3/pg213-pcie4-ultrascale-plus.pdf

[16] Samtec Inc., *FIREFLY™ MICRO FLYOVER SYSTEM™*, [Online]. Available at: https://www.samtec.com/optics/optical-cable/mid-board/firefly

[17] II-VI Inc., *25G BOA (Board-Mount Optical Assembly)*, [Online]. Available at: https://ii-vi.com/product/25g-boa-board-mount-optical-assembly/

[18] Amphenol ICC Inc., *LEAP® ON-BOARD TRANSCEIVER 300Gb/s High Speed Optical Module*, [Online]. Available at: https://cdn.amphenol-icc.com/media/wysiwyg/files/documentation/datasheet/opticalinterconnect/aop_leap_obt.pdf

[19] Xilinx Inc., *UltraScale Architecture Configuration User Guide*,[Online]. Available at: https://www.xilinx.com/support/documentation/user_guides/ug570-ultrascale-configuration.pdf

[20] Xilinx Inc., *SPI Flash Programming Including Bitstream Revision Selection*,[Online]. Available at: https://www.xilinx.com/support/documentation/application_notes/xapp1191-spi-flash-programming.pdf

[21] Intel Inc., *Interfacing I2C* Devices to an Intel SMBus Controller*, [Online]. Available at: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/smbus-controller-i2c-devices-paper.pdf

[22] Panasonic Inc., *High Speed, Low Loss Multi-layer Materials MEGTRON6*, [Online]. Available at: https://b2b-api.panasonic.eu/file_stream/pids/fileversion/244768

[23] Intel Inc., *PCB Breakout Routing for High-Density Serial Channel Designs Beyond 10 Gbps*, [Online]. Available at: https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/an/an651.pdf

[24] Gigabyte Inc., *X570 AORUS MASTER (rev. 1.1/1.2)*, [Online]. Available at: https://www.gigabyte.com/us/Motherboard/X570-AORUS-MASTER-rev-11-12#kf

[25] Xilinx Inc., *UltraScale Architecture-Based FPGAs Memory IP v1.4*, [Online]. Available at: https://www.xilinx.com/support/documentation/ip_documentation/ultrascale_memory_ip/v1_4/pg150-ultrascale-memory-ip.pdf

[26] Xilinx Inc., *IBERT for UltraScale GTY Transceivers v1.3*, [Online]. Available at: https://www.xilinx.com/support/documentation/ip_documentation/ibert_ultrascale_gty/v1_3/pg196-ibert-ultrascale-gty.pdf

[27] WHIZZ SYSTEMS Inc., *LOOPBACK CARDS*, [Online]. Available at: https://www.whizzsystems.com/loopback-card/

[28] B. G. Taylor, *TTC Distribution for LHC detectors*, *IEEE Transactions on Nuclear Science*, Vol. 45, No. 4, pp. 821-828, 1998.

[29] SiliconLabs Inc., *Si5345, Si5344, Si5342 Rev. D Family Reference Manual*, [Online]. Available at: https://www.silabs.com/documents/public/reference-manuals/Si5345-44-42-D-RM.pdf

[30] Xilinx Inc., *Virtex UltraScale+ FPGA Data Sheet: DC and AC Switching Characteristics*, [Online]. Available at: https://www.xilinx.com/support/documentation/data_sheets/ds923-virtex-ultrascale-plus.pdf

[31] W. Kester, *Converting Oscillator Phase Noise to Time Jitter*, [Online]. Available at: https://www.analog.com/media/en/training-seminars/tutorials/MT-008.pdf

[32] Xilinx Inc., *Versal ACAP Prime Series Product Selection Guide*, [Online]. Available at: https://www.xilinx.com/support/documentation/selection-guides/versal-prime-product-selection-guide.pdf