FELIX: first operational experience with the new ATLAS readout system and perspectives for HL-LHC

Joaquin Hoya, on behalf of the ATLAS TDAQ Collaboration[1,*,**]

[1]Argonne National Laboratory
9700 S. Cass Avenue, Bldg. 362
Lemont, IL 60439

Abstract. The Front-End Link eXchange (FELIX) readout system is deployed as the new interface connecting detector front-end electronics with the data acquisition and Timing, Trigger and Control (TTC) systems. FELIX functions as a router between custom serial links from front-end ASICs and FPGAs to data collection and processing components via a commodity switched network. FELIX uses commodity server technology in combination with FPGA-based PCIe I/O cards. FELIX servers run a software routing platform serving data to network clients performing a number of data preparation, monitoring and control functions. This paper covers the design of FELIX as well as the first operational experience gained during the Run 3 start, including the challenges faced commissioning the system for each ATLAS sub-detector. Over the next decade, the ATLAS detector will be required to operate in an increasingly harsh collision environment and challenging data taking conditions. To maintain physics performance, the detector will undergo a series of upgrades. In particular, the readout system capabilities will be improved. The planned evolution of FELIX for High-Luminosity LHC will be described, including architectural changes and status of early integration with detector development projects.

ATL-DAQ-PROC-2023-008
21 September 2023

1 Introduction

The ATLAS experiment [1], located at the Large Hadron Collider (LHC) at CERN, investigates the physics at the energy frontier by analyzing proton-proton and heavy ions collisions. The LHC accelerates particle bunches and collide them at a center-of-mass energy up to $\sqrt{s}$ = 13.6 TeV every 25 ns. To filter the vast amount of collision data generated, ATLAS uses a two-level trigger system [2] that examines data from each bunch crossing and keeps only those events of interest. The hardware-based Level-1 (L1) trigger reduces the 40 MHz collision rate to a maximum of 100 kHz trigger rate, using coarse information from the calorimeter and the muon detectors. Then, the software-based High-Level Trigger (HLT) further reduces the L1 rate to 3 kHz, by using higher granularity calorimeter and muon detector information, together with inner detector tracking information.

FELIX, the Front-End Link eXchange readout system, was installed for the new detector and trigger systems taking part of the current LHC Run-3 (the third LHC data taking period

*e-mail: joaquin.hoya@cern.ch

scheduled from 2022-2025), and it is integrated alongside the existing legacy readout system (ROD and ROS [2, 3]), as illustrated in Figure 1. In Run-3, the readout of the new muon detector systems NSW (New Small Wheel) [4] and BIS7/8 (Barrel Inner Small)[5], the new trigger systems for the Liquid Argon (LAr) calorimeter digital readout [6], and the calorimeter hardware trigger (L1Calo) [7] are handled by FELIX.

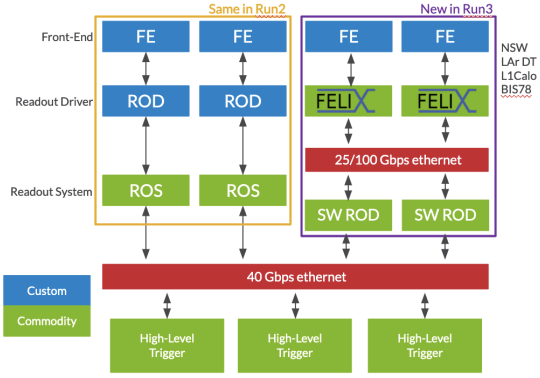


Figure 1: Diagram of the ATLAS TDAQ system in Run-3.

The FELIX system is the ATLAS-wide effort to harmonize detector readout systems. FELIX significantly reduces the number of custom electronic components and design effort (with respect to the fully-custom-made legacy readout architecture) by leveraging commercial computing systems, including network interface cards, servers, and network switches, and grants greater flexibility in maintenance, upgrades, and customization. The ATLAS FELIX readout system is comprised of commodity servers equipped with PCIe FELIX cards. Each FELIX I/O card houses an FPGA interfaced to the PCI express bus, along with fiber-optic transceivers. The data being routed include not only the detector readout (with a bidirectional communication), but also configuration, trigger decisions and LHC-related signals such as bunch and trigger counters reset. The TTC (Timing, Trigger, and Control) [8] system distributes the trigger and LHC information to FELIX with fixed deterministic latency. Both the TTC and to-detector links employ custom fully synchronous data transmission protocols. The FPGA role encompasses data transmission management and routing between the detector links and the PCIe interface. The high-throughput data routing between the FELIX cards and the switched network is enabled by the FELIX software.

The FELIX host interfaces through a switched internet with the SoftWare ReadOut Driver (SWROD) [9] that runs on commodity servers, and is tasked with event building and aggregation, and detector-specific data processing.

In Run-4 (2029-2031), LHC will commence the High Luminosity program (HL-LHC) [10], delivering collisions at higher instantaneous luminosity. During this time, the FELIX system will interface with all ATLAS detector and trigger systems.

A brief description of the FELIX hardware, firmware and software in use during Run-3 is presented in Section 2. The performance in the first operational experience with the readout of the new subdetectors is discussed in Section 3. Perspectives for the upcoming upgrades for the HL-LHC are mentioned in Section 4.

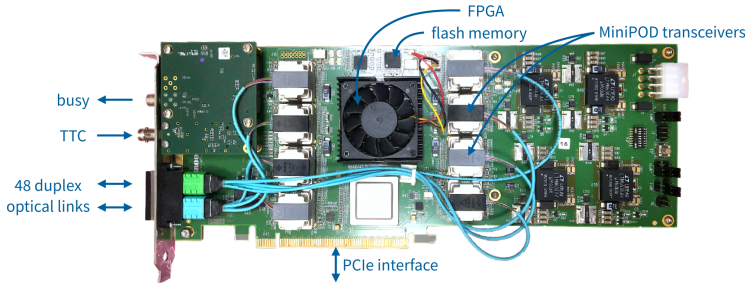


Figure 2: FELIX I/O card, FLX-712. This version of the card has 48 duplex optical links.

2 The Run-3 FELIX system

2.1 FELIX I/O card and server

FLX-712, shown in Figure 2, is a custom-designed PCIe card [11], equipped with a Xilinx Kintex Ultrascale (XCKU115) FPGA, 8 Avago Minipod transceivers to support up to 48 bidirectional optical links and a 16-lane PCIe Gen3 interface divided into two 8-lane endpoints with a switch (with throughput up to 128 Gb/s). The front-end optical links can connect via two optical multi-fiber couplers (MTP-24 or MTP-48 depending on the application). The card also incorporates a mezzanine card for the TTC link and a LEMO connector for busy propagation. Firmware images can be stored on the card thanks to a 2GB flash memory and a micro-controller. Approximately 300 FLX-712 boards were produced for ATLAS Run-3 DAQ, ProtoDune, NA62 and ATLAS tracker upgrade.

A FELIX PC consists of a commodity server equipped with an Intel Xeon E5-1660v4 CPU (8 cores), 32 GB of memory, a Mellanox Connect-X5 network interface (25 or 100 Gb/s) with one or two FLX-712 cards.

2.2 Firmware

The FELIX firmware communicates with the front-end electronics (FE) using the GBT (GigaBit Transceiver) [12] serial data transmission protocol, with a link speed of 4.8 Gb/s, ensuring synchronization with the 40 MHz clock of the LHC, a crucial aspect for reliable data processing. A single GBT link carries multiple data streams, called *E-links*, with configurable bandwidth (80/160/320 Mb/s). On the detector front-end, the GBT protocol is used by the radiation-hard GBTx ASIC[13].

The FELIX firmware has two main flavours or modes, known as GBT and FULL. Both modes use the GBT protocol for the links towards the FE and support up to 24 links. The GBT mode uses the GBT protocol to receive data from the FE, while the FULL mode implements a light-weight protocol for this path, with a bandwidth of 9.6 Gb/s per optical link. FULL mode is intended for communication with other FPGA-based systems over the custom 8b/10b encoded protocol implemented as a single wide data stream with no handshaking or logical substructure (i.e. no *E-links*). LAr Digital Processing Blade and L1Calo trigger systems use FELIX in FULL mode, while NSW and BIS 7/8 subdetectors use GBT mode.

The main components of the firmware [14] are the GBT encoders/decoders, the data management firmware, known as the Central Router, the TTC decoder and Wupper [15], which is the PCIe engine with DMA interface. The Central Router passes data between *E-links* of the GBT links and the PCIe interface in both directions. It encodes and decodes *E-links* data with protocols such as 8b/10b, HDLC and various customs serials protocols.

2.3 FELIX Software framework

The FELIX software suite[1] includes drivers, low-level tools, routing software and interface libraries. The low-level tools are used to control and configure the FELIX I/O card. The *cmem_rcc* driver allows applications to allocate large contiguous buffers in the server memory, facilitating high-bandwidth access via DMA.

To efficiently distribute data between the buffers and subscribed applications through the network, the routing tool *felix-star* is used, as shown in Figure 3. *felix-star* runs as daemon on FELIX servers, is designed to be event driven, with asynchronous non-blocking operation and single-threaded with one processes per PCIe endpoint.

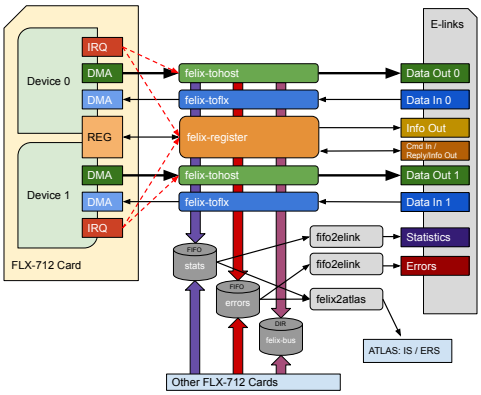


Figure 3: Architecture of felix-star data routing tool.

The subscribed applications such as a SWROD server, communicate with *felix-star* using the custom NetIO interface library based on libfabric [16]. It uses Remote Direct Memory Access (RDMA) to significantly reduce the number of data copies needed for data transfer, bypasses kernel context-switching, and minimizes CPU utilization, enhancing overall performance. NetIO also supports data coalescence, especially convenient when the data chunks to be transferred are small as in GBT mode, or a true zero-copy approach (no data coalescence) when the data chunks are large.

The performance of *felix-star* has been thoroughly tested and verified, with trigger rates of up to 200 kHz, surpassing the Run-3 requirements by a factor of two.

3 FELIX performance in Run-3

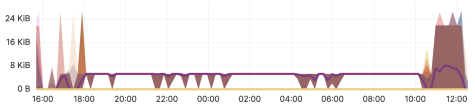
For Run-3, 64 FELIX PCs with 105 FLX-712 cards were installed into the ATLAS DAQ system. The configuration and readout of FELIX cards is orchestrated by an application based on Supervisord [17]. It automatically starts and restarts *felix-star* applications and the different configurations can be seen via a web interface. The monitoring is integrated in the ATLAS infrastructure, publishing to the ATLAS Error Reporting System (ERS) [18] and using Grafana dashboards [19] for operation monitoring [20].

For the operational performance of FELIX, we present the three largest applications of FELIX participating in the data-taking sessions during Run-3, specifically LAr Digital Processing Blade, Level-1 Calorimeter trigger and New Small Wheel. Operation of a large system is always a challenge, especially when the DAQ needs to interface a variety of FEs. Overall, the data losses because of FELIX have been negligible and, during commissioning and early data-taking, FELIX software was improved to prevent it. Also, FELIX has been able to provide firmware and software tools to diagnose issues in the FEs.

[1] https://gitlab.cern.ch/atlas-tdaq-felix

(a) Max. Throughput ~ 40 Gb/s

(b) Rate ~ 100 kHz



(c) Max. Message size 22 kB

Figure 4: Examples of FELIX readout of LDPB: Throughput, Rate and Message Size.

(a) Throughput ~ 8 Gb/s

(b) Rate ~ 96 kHz

Figure 5: FELIX performance of L1Calo readout: Throughput and Rate examples for gFEX.

3.1 LAr Digital Processing Blade

LDPB uses FELIX in FULL mode and is the ATLAS system with the largest data throughput, although it is far from network maximum bandwidth of 100 Gb/s. Figure 4a shows, in normal data-taking conditions, the received data throughput from each PCIe endpoint as a function of time, reaching a total of ~ 40 Gb/s considering both devices. In turn, Figure 4b shows, for the same period of time, the rate close to 100 kHz, with stable performance. Finally, Figure 4c shows the maximum size of the messages, that can be up to 22kB. LAr is the only system that uses a true zero-copy approach, where the software does not accumulate messages in network buffers: the network device sends messages composed of fragments residing in the DMA buffer.

3.2 Level-1 Calorimeter trigger

The Level-1 calorimeter trigger also uses FELIX in FULL mode. This system is actually composed of gFEX (Global Feature Extractor), eFEX (Electron Feature Extractor), jFEX (Jet Feature Extractor) and TREX (Tile Rear Extension), the new additions to the ATLAS hardware trigger.

Figure 5a shows, for a high-rate 2023 run, the received data throughput for the gFEX board as a function of time, reaching ~ 8 Gb/s, when the trigger rate is close to 100 kHz, as shown in Figure 5b. The average message size is 3kB. L1Calo uses a feature called *streams*: each of the 16 FULL mode links in use carries up to 9 streams, each transmitting messages at a rate of 100 kHz.

3.3 New Small Wheel

The New Small Wheel muon detector uses FELIX in GBT mode. Each FELIX card serves up to 200 *E-links*, each transmitting messages at the L1 trigger rate. The average size of a NSW message is 40B.

The large number of *E-links* requires more computing resources to process NSW data. One of the major challenges on the software side, faced during the first year of Run-3, was to prevent late packet arrival from FELIX to SWROD. All messages were delivered, but with a latency occasionally exceeding the SWROD time window of O(10) ms. It was found that the leading cause was CPU saturation on the FELIX host, reaching 100% utilization for trigger rates above 80 kHz. Figure 6 shows the packet arrival efficiency for a 2022 run on the left, and for a 2023 run on the right after the SWROD was optimized to reduce CPU utilization.

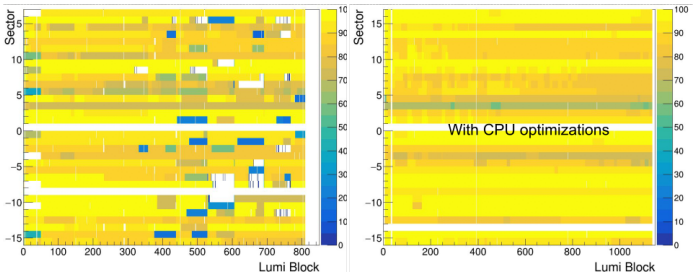


Figure 6: NSW readout efficiency before and after CPU optimizations to correct for late packet arrival.

4 ATLAS DAQ in Run-4

In Run-4, the Level-0 trigger rate will increase a factor 10 with respect to Run-3, reaching a nominal value of 1 MHz. The data volume per bunch crossing is also expected to increase due to the threefold increase in the proton-proton interactions per bunch crossing or pileup. The estimated data throughput flowing through the DAQ system is 4.6 TB/s, over 20 the typical Run-3 throughput. For these reasons, an upgrade of the TDAQ infrastructure is necessary. A diagram of the upgraded system is shown in Figure 7.

FELIX will readout all the subdetectors, meaning that legacy ROD/ROS systems will be replaced entirely with FELIX. There will be around 14000 optical links with bandwidth up to 25Gb/s per link. FELIX will support new detector-specific data transmission protocols. The SWROD will evolve into the Data Handler, that is currently under development.

4.1 Future FELIX: Prototypes, firmware and software upgrades

To take advantage of the technological evolution and to increase the maximum link speed up to 25 Gb/s, instead of the current 10 Gb/s, new FELIX I/O cards have been prototyped. Two prototypes have been produced so far: FLX-181 and FLX-182, shown in Figure 8. These cards are equipped with an AMD Versal Prime FPGA and 24 FireFly bidirectional optical data links. The boards incorporate a 16-lane PCIe Gen 4 interface.

As part of the development process for the Run-4 upgrade, the FELIX firmware is supporting a broader range of detector requirements, in addition to higher link and PCIe interface speeds, and multiple buffers per PCIe endpoint in computer memory. Aside from the GBT protocol, the firmware will support the lpGBT protocol (the evolution of GBT), encoders and decoders for the Inner Tracker Pixel and Strip [21, 22] and the 64b/67b-encoded Interlaken protocol.

With respect to the software upgrades, the architecture will remain similar as in Run-3, but with a different deployment scheme.

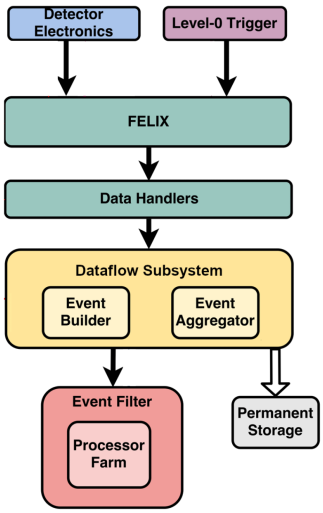


Figure 7: TDAQ architecture for Run-4.

(a) FELIX card prototype: FLX-181

(b) FELIX card prototype: FLX-182

Figure 8: Run-4 FELIX card prototypes.

4.2 Integration with new systems: Inner Tracker

The ITk is an all-silicon tracker detector, with custom ASICs covering detector acceptance up to $|\eta| < 4$. ITk is designed to surpass performance of the current tracker up to a pileup of 200.

Because its higher channel density, ITk is one of the major consumers of the DAQ bandwidth. Due to its complexity and intense development schedule its integration with FELIX began in 2019 and, by now, the FELIX readout of ITk is well advanced.

The FELIX firmware has been successfully tested for Pixel and Strip using Run-3 I/O card. FELIX is being used in ITk system tests, large scale detector prototypes [23].

5 Conclusions and Outlook

FELIX is a novel readout system designed for the ATLAS experiment. In the ongoing LHC Run-3, FELIX is being used to readout the new sub-detector systems, reducing the amount of custom hardware, with respect to the previous architecture, in the data taking path. Both firmware and software are mature and are being used for data taking sessions, showing a good and stable performance for all the new systems (NSW, LAr and L1Calo). Looking forward, the HL-LHC will challenge the ATLAS DAQ system and, starting from Run-4, FELIX will readout the full detector. New FELIX cards are under development to replace FLX-712, according to the technical specifications. The firmware is also under development, and builds are produced for FLX-712. The current software architecture is being scaled for Run-4. FELIX is already part of the production and testing of the new Run-4 sub-detectors such as the Inner Tracker.

References

[1] ATLAS collaboration, *The ATLAS experiment at the CERN large hadron collider*, JINST **3** (2008)

[2] M. Abolins et al., *The ATLAS Data Acquisition and High Level Trigger system*, JINST **11**, P06008 (2016)

[3] A. Gabrielli, *Commissioning of ROD boards for the entire ATLAS Pixel Detector*, JINST **13**, T09009 (2018)

[4] R.-M. Coliban et al., *The Read Out Controller for the ATLAS New Small Wheel*, JINST **11**, C02069 (2016)

[5] L. Massa, *The BIS78 Resistive Plate Chambers upgrade of the ATLAS Muon Spectrometer for the LHC Run-3*, JINST **15**, C10026 (2020)

[6] D. Besin et al., *Design and Evaluation of the LAr Trigger Digitizer Board in the ATLAS Phase-I Upgrade*, IEEE Trans. Nucl. Sci. **66**, 2011 (2019)

[7] ATLAS collaboration, *Technical Design Report for the Phase-I Upgrade of the ATLAS TDAQ System*, **ATLAS-TDR-023** (2013)

[8] S. Ask et al., *The ATLAS central level-1 trigger logic and TTC system*, JINST **3**, P08002 (2008)

[9] S. Kolos et al., *New Software-Based Readout Driver for the ATLAS Experiment*, IEEE Trans. Nucl. Sci. **68**, 1811 (2021)

[10] G. Apollinari et al., *High-Luminosity Large Hadron Collider (HL-LHC): Preliminary Design Report*, **CERN-2015-005** (2015)

[11] N. Ilic et al., *FELIX: the new detector interface for the ATLAS experiment*, EPJ Web Conf. **214**, 01023 (2019)

[12] P. Moreira et al., *The GBT Project*, **CERN-2009-006**, 342 (2009)

[13] P. Leitao et al., *Test bench development for the radiation Hard GBTX ASIC*, JINST **10**, C01038 (2015)

[14] W. Wu, *FELIX: the New Detector Interface for the ATLAS Experiment*, IEEE Trans. Nucl. Sci. **66**, 986 (2019)

[15] A. Borga et al., *Wupper: PCIe DMA Engine for Xilinx FPGAs*, https://opencores.org/projects/virtex7_pcie_dma, accessed: 2023-08-10

[16] J. Schumacher, *Event-Driven RDMA Network Communication in the ATLAS DAQ System with NetIO*, **ATL-DAQ-SLIDE-2019-848** (2019)

[17] *Supervisor: A process control system*, http://supervisord.org, accessed: 2023-08-10

[18] S. Kolos et al., *The Error Reporting in the ATLAS TDAQ System*, J. Phys. Conf. Ser. **608**, 012004 (2015)

[19] *Grafana dashboards*, https://go2.grafana.com, accessed: 2023-08-10

[20] I. Soloviev et al., *ATLAS Operational Monitoring Data Archival and Visualization*, EPJ Web Conf. **245**, 01020 (2020)

[21] ATLAS collaboration, *Technical Design Report for the ATLAS Inner Tracker Pixel Detector*, **ATLAS-TDR-030** (2017)

[22] ATLAS collaboration, *Technical Design Report for the ATLAS Inner Tracker Strip Detector*, **ATLAS-TDR-025** (2017)

[23] B. Vormwald, *Performance of the ATLAS ITK Pixel detector prototype*, **ATL-ITK-PROC-2023-009** (2023)