PROCEEDINGS OF SCIENCE

# FELIX and the SW ROD: commissioning the new detector interface for the ATLAS trigger and readout system

**J. G. Panduro Vazquez, on behalf of the ATLAS TDAQ Collaboration**[a,*]

[a] *Royal Holloway, University of London,*
*Egham Hill, Egham, TW20 0EX, United Kingdom*

*E-mail:* j.panduro.vazquez@cern.ch

After the current LHC shutdown (2019-2022), the ATLAS experiment will be required to operate in an increasingly harsh collision environment. To maintain physics performance, the ATLAS experiment is undergoing a series of upgrades. A key goal of this upgrade is to improve the capacity and flexibility of the detector readout system. To this end, the Front-End Link eXchange (FELIX) system has been developed. FELIX acts as the interface between the data acquisition; detector control and TTC (Timing, Trigger and Control) systems; and new or updated trigger and detector front-end electronics. The system 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. The serial links may aggregate many slower links or be a single high bandwidth link. FELIX also forwards the LHC bunch-crossing clock, fixed latency trigger accepts and resets received from the TTC system to front-end electronics. 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. Commodity servers connected to FELIX systems via the same network run the new multi-threaded Software Readout Driver (SW ROD) infrastructure for event fragment building, buffering and detector-specific processing to facilitate online selection. This presentation will cover the design of FELIX and the SW ROD, as well as the results of the installation and commissioning activities for the full system in spring 2021.

**** The European Physical Society Conference on High Energy Physics (EPS-HEP2021), ****
**** 26-30 July 2021 ****
**** Online conference, jointly organized by Universität Hamburg and the research center DESY ****

*Speaker





PoS(EPS-HEP2021)825

## 1. Introduction

Over the next decade, the Large Hadron Collider (LHC) will undergo a series of upgrades to maximise its discovery potential for new physics processes, towards a final stage known as High Luminosity LHC or HL-LHC. The resulting increase in average collision luminosity of up to 7 times the original design value poses a significant challenge to the experiments serviced by the collider in terms not only of data volume and rate, but also event processing complexity. In order to prepare for this new operational environment, the ATLAS [1] experiment is undertaking a series of upgrades to the detector, trigger and data acquisition (DAQ) systems.

As part of this process, the Front-End LInk eXchange (FELIX) system has been developed as the primary detector interface between the front-end electronics and the DAQ system. By taking advantage of recent advances in server performance, FPGA capabilities and network throughput, FELIX will replace existing legacy hardware with a flexible routing platform able to receive data directly from front-end electronics and serve them to peers on a commodity switched network. FELIX will also serve as a relay for trigger and clock information from the Timing, Trigger and Control (TTC) [1] system to front-end electronics. It will also be possible to use FELIX to send general purpose control data to front-end electronics to manage modules throughout data taking and calibration.

With FELIX in place, all data processing, formatting and monitoring tasks previously performed in custom hardware will now take place in software running in commodity server farms, a facility known as the Software Readout Driver (SW ROD). The first wave of systems to be upgraded to use FELIX and the SW ROD will be those undergoing significant detector or readout upgrades during the 2019-2022 experimental shutdown ahead of the third major LHC data taking period (Run 3). These are the New Muon Small Wheels (NSW) [3], Liquid Argon (LAr) digital readout [4] and the calorimeter hardware trigger electronics (L1Calo) [5]. A smaller scale demonstrator for upgraded Barrel RPCs (BIS 7/8) [6] will also be installed during this shutdown. The remaining ATLAS systems will then be migrated to FELIX en-masse during the next long shutdown from 2025-2027, ahead of what will then be Run 4.

In this paper we will describe the design of the FELIX and SW ROD platforms, along with the status of the ongoing commissioning effort taking place throughout 2021.

## 2. ATLAS DAQ System Overview and Run 3 Upgrade

In the Run 3 system, all new readout paths will use the combined FELIX and SW ROD system, with the rest of the detector read out via the legacy (Run 1 and 2) path, as shown in Figure 1. FELIX receives Trigger signals from the ATLAS TTC system over a dedicated optical link and relays them to front-end electronics, causing event data to be read out. Data are then received over point-to-point optical links and routed to peers (typically commodity servers) connected via an Ethernet network. The primary peer on this network will be the SW ROD, which will perform the majority of the data processing functions previously performed in custom hardware modules themselves called Readout Drivers (RODs), which FELIX and SW ROD replace in the data taking path. Other systems connected to the network will also be able to subscribe to FELIX and SW ROD data streams for the purposes of monitoring and calibration control. The SW ROD will also implement

local buffering functionality analogous to that which takes place in the equivalent component in the legacy system, the Readout System (ROS) [5], which received data directly from the legacy ROD modules and is the second component replaced in the new architecture. The software interface between the High-Level Trigger (HLT) and the SW ROD for event selection purposes is required to be identical to that which exists between the HLT and the ROS. Thus, both legacy and upgraded readout paths will be able to operate side-by-side, as shown in Figure 1. The overall input trigger and output recording rates for the DAQ system will not change from Run 2, operating at 100 kHz and 1.5 kHz respectively.

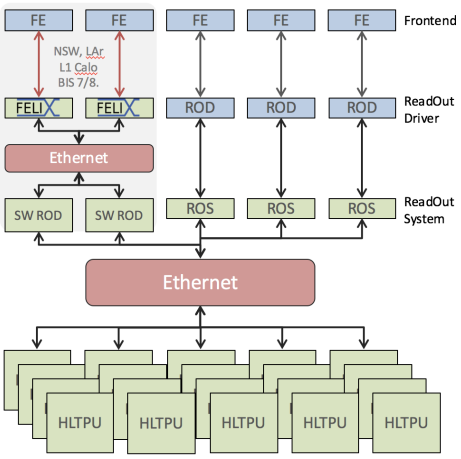


**Figure 1:** Diagram of the ATLAS DAQ System in Run 3 [7]. The new FELIX and SW ROD Components (left) operate alongside the legacy (ROD and ROS) system on the right. HLT processing units (HLTPUs) are able to sample event data from both readout paths via an identical interface.

## 3. High-Level Architecture

FELIX systems are able to interface with front-end electronics over one of two optical link protocols: GigaBit Transceiver (GBT [2]), a radiation-hard standard developed at CERN, where multiple lower speed links (known as E-links) from separate pieces of electronics can be aggregated into single 4.8 Gb/s link; and FULL mode, an in-house design with no link substructure for higher bandwidth (9.6 Gb/s) communication between FPGAs. Data streams for either protocol can be configured to use different encoding, although 8b10b is typically used for normal dataflow. Each FELIX server hosts custom I/O cards, known as FLX-712, with firmware to interface with either of the two link protocols. For the GBT case each server hosts two cards, whereas for the higher bandwidth FULL mode case each server hosts one card (driven primarily by the number of available PCIe lanes). Each server also hosts high bandwidth network interface cards (25 GbE for GBT, 100 GbE for FULL mode). Each FELIX server has an Intel® Xeon® E5-1660 V4 CPU (8 cores running at 3.2 GHz) and 32 GB of ECC RAM.

In the first upgrade phase towards HL-LHC approximately 60 FELIX servers hosting a total of 100 FLX-712 cards will be deployed, routing data to 30 SW ROD systems. A significantly larger number (6 times more) will be deployed in the 2025-2027 shutdown to service all remaining ATLAS systems.

## 4. FELIX Design

The FLX-712, shown in Figure 2, supports a 16-lane PCIe Gen 3 interface, able to reach a throughput of up to 100 Gb/s. The interface to the front-end is via either MTP 24 or 48 couplers, after which the light is internally routed to one of eight MiniPOD transceivers. A maximum of 48 bi-directional optical links can be connected to each board. A Xilinx® Kintex Ultrascale (XCKU115) FPGA provides the platform for all on-board firmware features. An interchangeable mezzanine card provides an interface for a number of timing and control systems. Each board also hosts dedicated electronics for configuration and diagnostics.

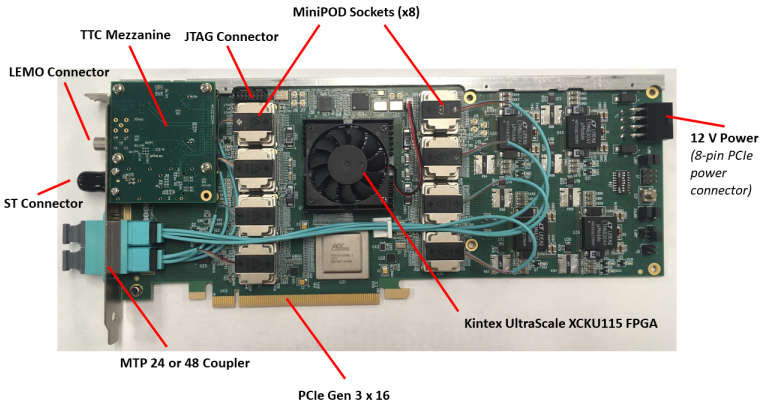


**Figure 2:** Photograph of a FLX-712 card with key components labelled [7].

The FELIX firmware is designed to be modular and flexible. Separate components manage different key functions, such as the link wrapper (GBT or FULL mode) and the PCIe and DMA engines. Between these lies a routing module, where data arriving over different links are prepared for DMA to the host server's memory. A separate module interacts with the TTC system, routing trigger and clock signals to front-end electronics as well as inserting an information packet into the data stream to the SW ROD for each L1 trigger accept to synchronise its operation. Should operating conditions require a pause in dataflow, the module can also relay a BUSY signal back to the central trigger on demand.

For typical dataflow use cases the chosen FPGA permits 24 GBT or FULL mode channels to be serviced by each card, but dedicated builds exist which handle 48 links purely for the purposes of distributing TTC signals to the front-end, without any dataflow in the other direction.

The FELIX software suite comprises high- and low-level components. Alongside a dedicated device driver, low-level tools make it possible to use all firmware features in a laboratory setting and debug any issues which may arise. At a higher level, a high performance daemon operates in an 'always on' fashion in order to receive data from the FLX-712 over DMA and provide onward routing. The software daemon is designed to be event driven [8], able to react to hardware interrupts from the card indicating incoming data, or signals from the network interface or operating system. The design is such that copies of the data in memory are kept to a bare minimum to maximise throughput. High performance network interface software, called NetIO [8], is responsible for the final stage of routing to connected network peers.

## 5. FELIX Performance and Commissioning

FELIX systems have been in use across multiple subdetector testing and development sites for numerous years, including large scale surface integration with, for example, the ATLAS New Muon Small Wheel, where the entire detector was commissioned on the surface before installation in the ATLAS cavern. Along with dedicated performance testing by the FELIX developers, this experience has made it possible to robustly test all firmware and software features, as well as to stress test the system at high rate. In all cases FELIX has been demonstrated to exceed the Run 3 performance requirements with sufficient margin to absorb evolutions in collision environment and subdetector configuration.

FELIX is now being used successfully by all subdetectors taking part in the Run 3 upgrade, both on the surface and more recently the final installation for ATLAS data taking itself. The LAr calorimeter has demonstrated stable control and operation of their entire new timing system (LTDB), with data from the new digital processors (LDPB) integrated into the main ATLAS data stream via FELIX. The New Small Wheels are now in the process of being installed in the ATLAS cavern, with the A side installed earlier this year and the C side scheduled to be completed before the end of 2021. In both cases FELIX has been used successfully in both control and data taking modes. The other Run 3 systems, namely the L1 Calorimeter trigger and BIS 7/8, are also in the process of final hardware installation in ATLAS, with successful integration with FELIX both on the surface and in the cavern.

## 6. SW ROD Requirements, Design and Performance

The primary role of the SW ROD is to receive data from FELIX nodes and facilitate subdetector-specific processing (integrated via a plugin mechanism) before buffering the produced event fragments during the HLT decision, serving data to HLT nodes on demand. An individual SW ROD node can receive data streams from either individual FELIX nodes (systems using the higher bandwidth FULL mode links to FELIX from their front-end) or multiple nodes (for systems using GBT mode). In the latter case the SW ROD also makes it possible to aggregate data for different GBT E-links into combined event fragments according to a detector-specific algorithm. As well as the HLT, the SW ROD can also serve data to event sampling and monitoring systems to make it possible to track the performance of the system and overall data quality. Finally, to support testing and calibration, the SW ROD also makes it possible to write data to disc or serve them to dedicated calibration processing nodes.

Of the performance requirements driving the design of the SW ROD, the most significant is the large packet rate arriving at servers aggregating GBT links. Here, the large multiplicity of E-links means packet rates of up to 115 MHz may be experienced at the highest trigger rates. To handle this load, each SW ROD server hosts dual Intel Xeon Gold 5128 CPUs, each with 16 physical cores running at 2.3 GHz. These are complemented by 96 GB of DDR4 2667 MHz RAM as well as a 100 GbE network interface to FELIX and 40 GbE network interface to the HLT. The software infrastructure provided with each instance includes a highly optimised input stage, whereby the input packet rate is reduced as early as possible through multiple aggregation steps using temporary buffers. This means that the packet rate experienced by the subdetector processing code is nearer

the L1 trigger rate (i.e. up to 100 kHz in Run 3), thus maximising the available CPU performance for this step.

Testing in a laboratory setting has yielded promising results, well in excess of the ATLAS Run 3 requirements. Individual subdetector groups are also well advanced in their implementation of their processing plugins, demonstrating the robustness of the architecture and providing valuable input to drive additional optimisation ahead of integration for ATLAS data taking.

## 7. Conclusion

In this paper, the new readout interface for the ATLAS detector at the LHC at CERN has been presented. The new system, featuring the FELIX and SW ROD systems, is significantly more flexible than the legacy hardware it replaces. By making use of new technology it has been possible to move almost all data processing and monitoring operations previously performed in custom hardware into software running on commodity servers, with a common interface for all front-end electronics. The results of performance tests show that the new system is able to satisfy all ATLAS performance requirements for the upcoming run period with a healthy operational margin. The FELIX and SW ROD systems are in the process of being commissioned in the ATLAS computing cavern, on track for the start of Run 3 data taking in 2022.

## References

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

[2] P. Moreira et al., *The GBT Project*, Proceedings of the Topical Workshop on Electronics for Particle Physics 2009

[3] ATLAS Collaboration, *The Read Out Controller for the ATLAS New Small Wheel*, JINST **11**, C02069 (2016)

[4] ATLAS Collaboration, *ATLAS Liquid Argon Calorimeter Phase-I Upgrade Technical Design Report*, CERN-LHCC-2013-017, ATLAS-TDR-02

[5] ATLAS Collaboration, *Technical Design Report for the Phase-I Upgrade of the ATLAS TDAQ System*, CERN-LHCC-2013-018, ATLAS-TDR-023

[6] S. Biondi et al., *Upgrade of the ATLAS Muon Barrel Trigger for HL-LHC*, ATL-UPGRADE-PROC-2015-003.pdf

[7] ATLAS TDAQ Collaboration, *FELIX: the New Detector Interface for the ATLAS Experiment*, ATL-DAQ-PROC-2018-003

[8] J. Schumacher et al., *Event-driven RDMA network communication in the ATLAS DAQ system with NetIO*, Proceedings of Computing in High Energy Physics 2019