EPJ Web of Conferences **245**, 01037 (2020) https://doi.org/10.1051/epjconf/202024501037
CHEP 2019

FELIX: the new detector interface for ATLAS

William Panduro Vazquez[1,*] on behalf of the ATLAS TDAQ Collaboration

[1]Royal Holloway, University of London Egham Hill, Egham TW20 0EX

Abstract. After the current LHC shutdown (2019-2021), the ATLAS experiment will be required to operate in an increasingly harsh collision environment. To maintain physics performance, the ATLAS experiment will undergo a series of upgrades during the shutdown. 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 results of the installation and commissioning activities for the full system.

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 seven 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 [2]. As part of the re-evaluation of the architecture of the DAQ system, it became clear that the evolution of technology (e.g. larger and faster FPGAs, the advent of multi-core CPUs and high performance networking) made it possible to move tasks which were previously performed in customised hardware into the more flexible firmware and software domains.

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

As a culmination of this new approach, the Front-End LInk eXchange (FELIX) system has been developed as the primary detector interface between the front-end electronics and the DAQ system. FELIX will replace existing legacy hardware with a flexible routing platform able to receive data directly from front-end electronics and serve it to peers on a commodity switched network. FELIX will also serve as a relay for trigger accept 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 be able to take place in software running in commodity server farms, a facility known as the Software Readout Driver (SW ROD). The first set of systems to be upgraded to use FELIX and the SW ROD will be those undergoing significant detector or readout upgrades during the 2019-2021 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) [2]. Smaller scale demonstrators for upgraded Barrel RPCs (BIS 7/8) [5] and the Tile Calorimeter [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 be the first HL-LHC run.

In this paper we will describe the design of the FELIX platform and its relationship to the rest of the DAQ system. Results of the most recent performance testing will be presented along with the prospects for installation and commissioning in early 2020.

2 ATLAS DAQ System Overview and Run 3 Upgrade

The legacy ATLAS DAQ System [7] (used for Runs 1 and 2) consisted of detector specific front-end electronics read out via point-to-point links to custom components known as Readout Drivers (RODs). Each detector system implemented its own ROD design, typically using VMEbus technology. RODs performed numerous tasks, from data formatting to error checking and correction, reporting and monitoring. Once processing was completed, data were then passed to and buffered in the first common element in the DAQ chain, known as the Readout System (ROS). Nodes in the High Level Trigger (HLT) farm were then able to sample data from these ROS buffers in order to perform a final event selection step before saving accepted events to permanent storage for further analysis. Data arrived in the DAQ system at the Level-1 hardware trigger accept rate of 100 kHz, and were recorded to permanent storage at an average rate of 1.5 kHz.

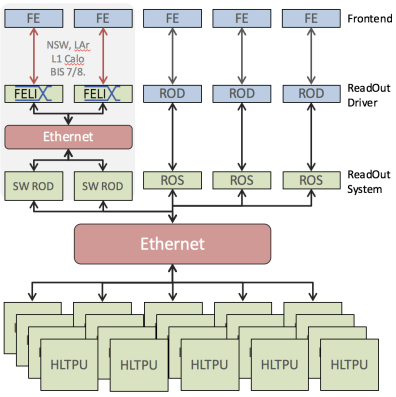

Figure 1. Diagram of the ATLAS DAQ System in Run 3. The new FELIX and SW ROD Components (left) operate alongside the legacy (ROD and ROS) system on the right. HLT processing units (HLTPUs) are able to sample event data from both readout paths via an identical interface.

In the Run 3 system, all new readout paths will use the combined FELIX and SW ROD system. 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 hardware. 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 the one originally taking place in the ROS. The software interface between the HLT and the SW ROD for event selection purposes is required to be identical to the one 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 remain at 100 kHz and 1.5 kHz respectively.

3 High-Level Architecture

FELIX systems are able to interface with front-end electronics over one of two optical link protocols: GigaBit Transceiver (GBT [8]), 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; 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 @ 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 (of order six times more) will be deployed in the 2025-2027 shutdown to service all remaining ATLAS systems.

4 Hardware

The FLX-712, shown in Figure 2, is a PCIe card supporting a 16-lane Gen 3 interface, able to reach a throughput of up to 100 Gb/s. An MTP 24 or 48 coupler provides the interface to external data fibres, after which the light is internally routed to one of eight MiniPOD transceivers handling 12 links each (four for receipt and four for transmission). A maximum of 48 bi-directional optical links can therefore be connected to each board. A Xilinx® Kintex Ultrascale (XCKU115) FPGA provides the platform for all on-board firmware features (see Section 5). An on-board PEX8732 PCIe switch [9] makes it possible to map two separate PCIe 8-lane endpoints into one 16-lane bus (as described in Section 5). A JTAG connector is provided to facilitate FPGA configuration, though this may also be stored in an on-board FLASH chip. FPGA programming and card health monitoring and control are also possible over PCIe.

Finally, an interchangeable mezzanine card provides an interface for a number of timing and control systems. The current ATLAS TTC system requires an ST interface for incoming

trigger and clock information and a LEMO connector for outgoing BUSY signals. An SI5345 jitter cleaner [10] on-board the FLX-712 itself ensures a sufficiently good quality clock for all FELIX use cases.

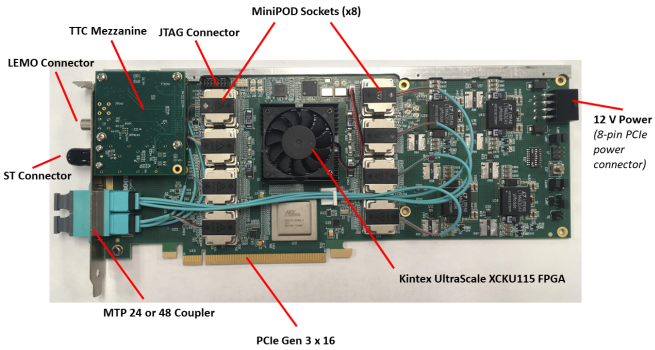

Figure 2. Photograph of a FLX-712 card with key components labelled.

5 Firmware

The FELIX firmware, a diagram of which can be found in Figure 3, 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 (a module known as Wupper). Between these lies the Central Router module, which performs the most data intensive workload. Here data arriving over different links are decomposed according to protocol and converted into regular 1 kB elements for optimal DMA transfer to the host server's memory. In order to optimise FPGA resource utilisation and timing, the FELIX firmware deployed in the FLX-712 consists of two duplicate paths with identical modules, each servicing half the input links and reading out to an 8-lane PCIe interface. As such, each FLX-712 card appears to the host server as two 8-lane devices.

Alongside the primary dataflow path, a separate common module interacts with the TTC system, injecting trigger and clock signals into the data path and relaying BUSY signals back to the central trigger should operating conditions require a pause in dataflow. The TTC module is also responsible for generating an information packet for each trigger accept received on a special stream for downstream subscribers to use to facilitate event fragment building and synchronisation. Other common modules also manage configuration registers, clock control/distribution and general housekeeping.

By re-using the basic blocks above it is possible to flexibly produce firmware designs for different use cases. For ATLAS, separate designs are maintained for both GBT and FULL mode, where the primary differences are the link wrapper and the complexity of the Central Router (which is significantly lower for FULL mode). Due to FPGA resource utilisation constraints the maximum number of GBT links which can be supported for primary dataflow is 24. For FULL mode the limitation comes from the PCIe bandwidth of the FLX-712, which implies a maximum of 12 links. However, 24 link variants are typically built to permit a larger number of links to be operated at lower average occupancy.

6 Software

The FELIX software suite comprises high- and low-level components. Alongside a dedicated device driver, low-level tools make it possible to test all firmware features in a laboratory

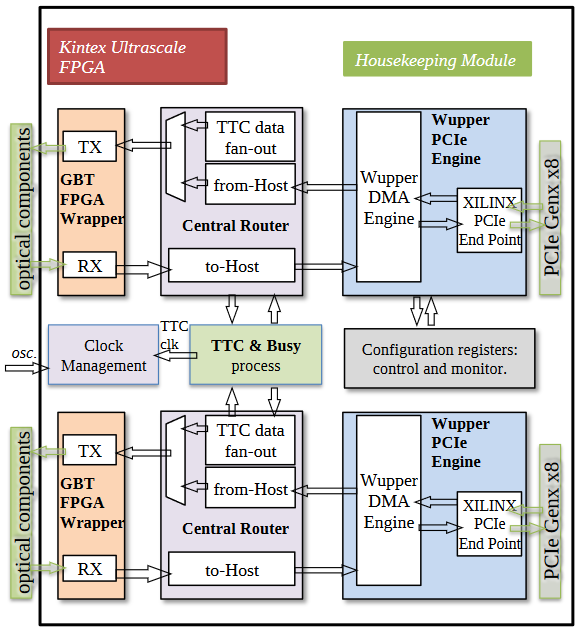

Figure 3. Diagram of the firmware deployed on the FPGA of the FLX-712 in GBT mode. In FULL mode the incoming (RX) GBT link wrapper is replaced with a dedicated FULL mode module. The Central Router is also much simpler in this case, as the data processing requirements are less severe.

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. DMA transfers are received via a separate ring buffer for each PCIe device visible to the host server (hence two per FLX-712). The software daemon is designed to be event driven [11], 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, known as NetIO [11], is responsible for the final stage of routing to connected network peers.

7 Performance Test Results

In this section, performance results will be presented for the most recent versions of the FELIX firmware and software, forming part of the approval process before mass production of hardware and installation in the ATLAS computing cavern. Tests were performed for each of the two use cases, namely GBT input links or FULL mode.

For GBT mode, a FELIX server hosting two FLX-712 cards was connected via a 25 GbE network to a data sink server. The input stage data to FELIX were provided via dedicated emulator hardware in a separate host server, providing a total of 48 GBT links to the FELIX server with configurable payloads. Trigger signals were provided using separate dedicated test hardware via the standard TTC interface.

As shown on the left in Figure 4, it was possible to operate all links with realistic payloads (a total of 384 individual E-links with packets of an average size of 40 Bytes) at the required 100 kHz rate for longer than 10 hours without any interruption. On the right of Figure 4, results are shown of a test activating the BUSY mechanism, showing that dataflow does indeed pause as expected. Finally, the trigger rate was increased to stress test the system, achieving an operating rate of 150 kHz and thus demonstrating the amount of design margin available (50%) with the current firmware and software.

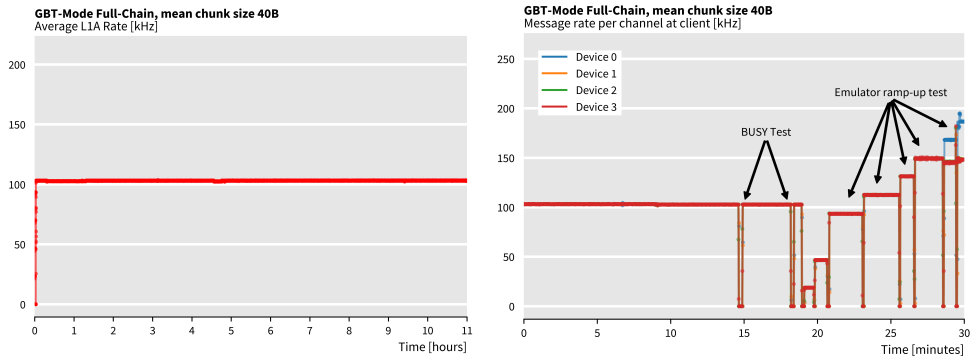

Figure 4. (left) Average L1 Accept rate in GBT mode performance test for all FELIX devices hosted by a single server. (right) BUSY and stress test results for GBT mode, split by FELIX device. Each device corresponds to 50 % of a FLX-712 card, as described in Section 5, hence four devices implies two FLX-712 cards hosted by the server. It was possible to run stably on all four devices up to a 150 kHz rate. The 'client' in the title refers to the application running on the data sink server.

For the FULL mode case a similar setup was used, but with a single FLX-712 card in the server, connected to a custom data source hosted separately. Furthermore, the FELIX server and data sink were this time connected with a 100 GbE network. In this case, as shown on the left in Figure 5, it was possible to reach a stable operating rate of 120 kHz with realistic link payloads (12 individual links with 4.8 kB packets). On the right of Figure 5, the results are shown of a stress test carried out without network, i.e. just transferring to the FELIX server across the bus. As can be seen, stable operation at an increasing trigger rate up to a final value of 300 kHz was achieved. The significant margin reflects the smaller processing overhead associated with the FULL mode protocol.

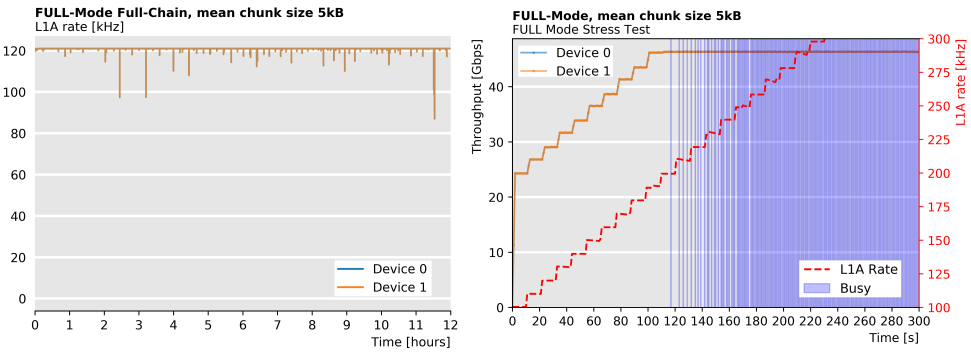

Figure 5. (left) Performance results for FULL mode, split by device (see Section 5) showing stable 120 kHz operation. (right) Stress test results for FULL mode, also split by device, where the L1 rate was increased until the FLX-712 asserted a signal known as XOFF, indicating saturation, indicated by the blue shaded region. In this case the combined throughout to the host across both devices was approximately 90 Gb/s.

8 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 component, 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 system is set to be deployed in the ATLAS computing cavern in early 2020 for commissioning in preparation for data-taking in 2021.

References

[1] The ATLAS Collaboration, JINST **3**, S08003 (2008)
[2] ATLAS Collaboration, *Technical Design Report for the Phase-I Upgrade of the ATLAS TDAQ System*, CERN-LHCC-2013-018 ; ATLAS-TDR-023 (2013)
[3] The ATLAS Collaboration, JINST **11**, C02069 (2016)
[4] ATLAS Collaboration, *ATLAS Liquid Argon Calorimeter Phase-I Upgrade Technical Design Report*, CERN-LHCC-2013-017 ; ATLAS-TDR-022 (2013)
[5] S. Biondi, PoS(EPS-HEP2015) p. 289 (2009)
[6] ATLAS Collaboration, *Technical Design Report for the Phase-II Upgrade of the ATLAS Tile Calorimeter*, CERN-LHCC-2017-019, ATLAS-TDR-028 (2017)
[7] ATLAS TDAQ Collaboration, Journal of Instrumentation **11**, P06008 (2016)
[8] P. Moreira et al., Topical Workshop on Electronics for Particle Physics, CERN-2009-006 pp. 342–346 (2009)
[9] *PEX 8732, PCI Express Gen 3 Switch, 32 Lanes, 8 Ports*, Broadcom Inc. (2010), `https://docs.broadcom.com/docs/12351851`
[10] *Si5345/44/42 Rev D Data Sheet*, Silicon Labs Inc (2016), `https://www.silabs.com/documents/public/data-sheets/Si5345-44-42-D-DataSheet.pdf`
[11] J. Schumacher et al., Proceedings of Computing in High Energy Physics 2019 (2020)