Software based readout driver evolution towards 1 MHz readout as part of the ATLAS HL-LHC upgrade

Serguei Kolos, on behalf of the ATLAS TDAQ Collaboration

***Abstract*—The High-Luminosity Large Hadron Collider (HL-LHC), which should begin operation in 2029, aims to increase LHC luminosity by a factor of 10 beyond its original design. In preparation for this increase, the ATLAS experiment has planned a major upgrade program that is split into two steps. During the "Phase-I" step, that will be completed in 2022, several new trigger and detector systems have been introduced and a new software application, called Software Readout Driver (SW ROD) developed to read data from these new systems. The SW ROD receives and aggregates data from the detector front-end electronics via the Front-End Link eXchange (FELIX) system. For LHC Run 3, which begins in 2022, the SW ROD will be used in parallel with the legacy Readout System (ROS), both operating at an input rate of 100 kHz. For the 'Phase-II' upgrade step, to be completed after Run 3 in time for Run 4 in 2029, the legacy ROS will be completely replaced with a new system based on the next generation of FELIX and a new software application called the Data Handler. The Data Handler is an evolution of the SW ROD that has the same functional requirements but must be able to operate at an input rate of 1 MHz. This contribution describes the design and implementation of the SW ROD application for Run 3. It also presents and discusses the results of performance measurements which demonstrate that the SW ROD application already fulfils the Phase-II performance requirements by being able to process data at a 1 MHz rate for realistic Phase-II input configurations.**

***Index Terms*—Data acquisition, Data collection, Data transfer, Object oriented programming**

I. Introduction

For LHC Run 3, which started in July 2022, the ATLAS experiment [1] has several new detector components that use modern Front-End (FE) electronics and an updated readout system. This system is based on a new facility called the Software Read Out Driver (SW ROD) [2], which receives FE data via the FE Link eXchange (FELIX) system [3]. The SW ROD performs event fragment building and buffering as well as serving data on request to the High Level Trigger (HLT). For Run 3 the new readout system is used in parallel with the legacy readout, but for Run 4, which is planned to start in 2029, all detector components will be upgraded to the new readout, which will have to operate at much higher input rate. To facilitate this the SW ROD component will evolve to a new one called the Data Handler, that has the same functional requirements but must be able to operate at 1 MHz input rate.

II. FELIX System Evolution to Run 4

FELIX is a new generic detector readout system, that is based on a custom PCIe card, that can receive data from detector FE electronics via several types of optical links and copy them to the memory of a commodity computer via the PCIe bus. FELIX is being used during Run 3 to receive data either via GigaBit Transceiver (GBT) [4] or the in-house designed FULL mode protocol. Run 3 FELIX cards, that are used for reading data from detectors front-end electronics have 24 input optical links, which in GBT mode can be split into up to 192 virtual links called E-Links. The optical links can operate at speeds up to 9.6 Gb/s and 12.8 Gb/s in GBT and FULL modes respectively, but the overall bandwidth of a FELIX card is limited by the 128 Gb/s bandwidth of the 16 lane PCI Express 3.0 interface.

The new FELIX card for Run 4 is expected to have up to 48 optical input links with data rates up to 25 Gb/s per optical link. It will support PCI Express Gen 4 interface, which would rise the total per card bandwidth limit to 256 Gb/s. These are preliminary parameters, which have been used for the ongoing Run 4 studies to track technology evolution. A final decision on the new FELIX card hardware platform will depend on the results of these studies.

For distributing data that is received via the FELIX card a custom in-house protocol called NetIO [5] has been developed. The current NetIO implementation is done on top of the RDMA over Converged Ethernet (RoCE) protocol that is supported by many modern network cards. For Run 4 NetIO will be optimized to reduce the number of non-application bytes per data packet.

III. SW ROD for Run 3

The SW ROD facility is implemented by a number of homogeneous software processes running on a set of commodity computers as shown in Fig. 1.

Each process has a specific configuration, which defines a set of FELIX links to be subscribed for receiving data as well as the data processing parameters. Each process also uses a custom data aggregation procedure that is specific for a given detector. This is achieved by implementing such a procedure as a plugin that implements a well defined ROBFragmentBuilder interface specified by the SW ROD.

A. SW ROD Performance Requirements

In Run 3, the SW ROD has to operate at an input rate of 100 kHz, matching the ATLAS Level 1 Trigger accept

S. Kolos, is with University of California, Irvine, CA 92697-4575, USA (e-mail: serguei.kolos@uci.edu)

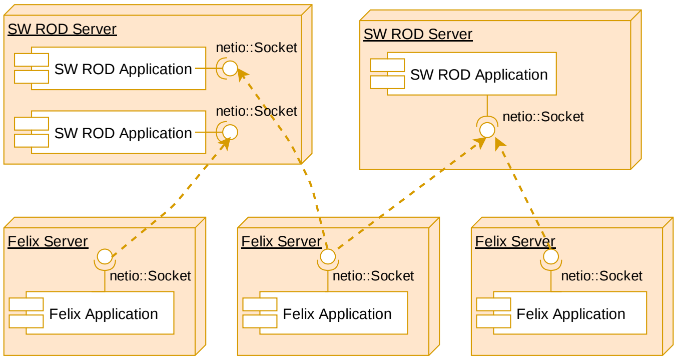

Fig. 1. SW ROD deployment.

rate. The number of input links and the overall data rates are defined by the detector front-end electronics requirements. Table I summarizes these numbers for a single FELIX card with 24 input links used in GBT mode.

TABLE I
FELIX CARD OUTPUT RATES FOR GBT MODE IN RUN 3

E-Links per card	Packet Size (B)	Packet Rate per Link (kHz)	Packet Rate per card (MHz)	Data Rate per card (GB/s)
192	40	100	19.2	0.77

B. SW ROD GBT Fragment Builder Algorithm

In GBT mode the SW ROD receives data from multiple E-Links and is responsible for aggregating them into so called ROB fragments by aligning data packets by their identifiers which were assigned by the L1 Trigger system. For Run 3 it is required that the Fragment Builder algorithm must scale for an arbitrary number of E-Links, from which data are aggregated. To address this requirement the algorithm implementation has been split into two stages as shown in Fig. 2.

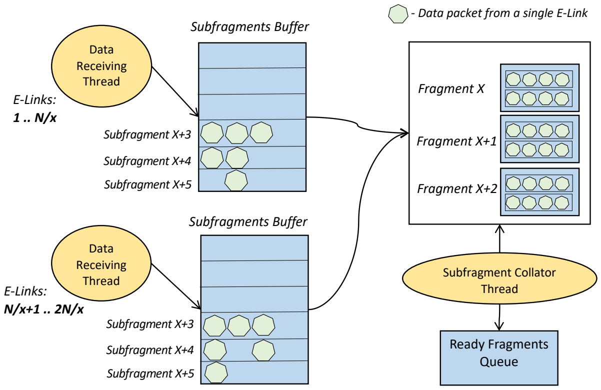

Fig. 2. SW ROD Fragment Builder algorithm.

The first stage of the algorithm is fully multi-threaded with all input E-Links being split between a configurable number of threads, such that each thread aggregates data from a unique subset of these E-Links into subfragments. In the stage two the subfragments produced by different threads are collated together into complete fragments. The first stage threads are completely independent and don't interact with one another until submitting completely built subfragments to the second stage. This approach scales extremely well as the bulk of the workload is effectively split among arbitrary number of isolated threads. This has been verified by performing a series of tests dedicated to studying the SW ROD scalability for the Run 3 requirements. Results of these tests are presented in the next section.

C. SW ROD Performance Tests

Scalability of the SW ROD fragment builder has been verified in a series of tests which have been performed on a test bed that replicates the same hardware configuration that will be used by the readout system during Run 3:

- SW ROD application running on a computer with a dual-socket Supermicro motherboard with 2 Intel(R) Xeon(R) Gold 5218 CPUs and 96 GB of DDR4-2667 RAM. Each CPU has 16 physical cores with a base frequency of 2.3 GHz.
- Input data for the tests generated by a FELIX card software emulation application running on another computer with an Intel Xeon E5-1660 v4 CPU with 3.2 GHz base frequency and equipped with 32 GB DDR4 2667 MHz memory.
- Both computers were equipped with Mellanox ConnectX-5 100 GbE network adapters, which were connected via a 100G network switch. Data were sent to the SW ROD application via the FELIX NetIO protocol.

D. Test Results

Three series of tests were performed with the SW ROD application using one, two and three threads respectively to receive and aggregate data chunks from every group of 192 input links, which corresponds to input from a single FELIX card. The total number of emulated FELIX cards for different test series varied from 1 to 6, which made for a total number of input channels increasing gradually from 192 to 1152. The size of the generated data chunks was set to 40 bytes. The results of these tests are shown in Fig. 3.

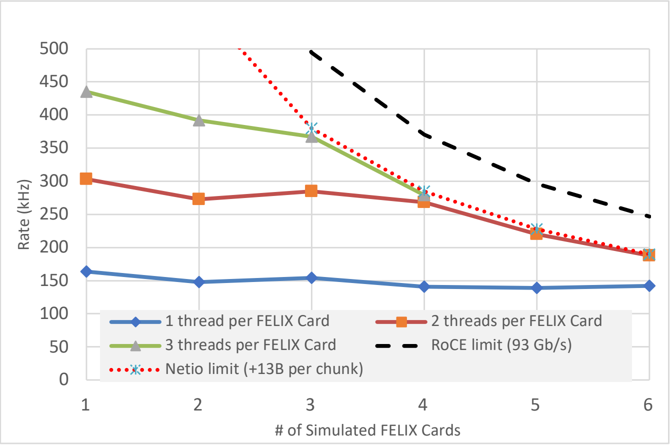

Fig. 3. GBT event fragment aggregation algorithm performance.

The dotted line shows the maximum theoretical input rate that can be obtained with the given hardware configuration, which is limited by the available network bandwidth. This line represents (1):

$$L = \frac{93 \times 10^9 Gb/s}{192 \times F \times (40 * 8 + 13 * 8)}, \qquad (1)$$

where L is the input rate, F is the number of simulated FELIX cards, 40×8 is the size of the data chunk in bits, 13×8 is the size of the NetIO protocol overhead per data packet in bits as well and $93 \cdot 10^9$ Gb/s is the maximum bandwidth that was achieved in the test setup using the RoCE protocol in Gb/s. This line shows that results of some tests with two and three reading threads were limited by the available network bandwidth. The test results demonstrate that the Fragment Builder algorithm scales very well with the number of worker threads.

E. Scalability Towards Run 4 Requirements

For Run 4 the ATLAS TDAQ system will have to receive data from trigger and detector electronics at an input rate of 1 MHz, which is defined by substantially increased instantaneous luminosity and the number of particle interactions per bunch crossing of the HL-LHC. As the readout system for Run 4 will be based on FELIX it is useful to study the limits of the current readout implementation, such that they can be addressed in the new TDAQ architecture. For this reason, a series of tests were performed with the SW ROD Fragment Builder algorithm to to find the number of worker threads that are required to perform fragment building at the rate of 1 MHz and how it scales with the number of input E-Links. For these tests the SW ROD application was used to aggregate data from all available input E-Links that a new FELIX card must be able to support. The configurations that were tested are summarized in Table II.

TABLE II
NEW FELIX CARD DATA OUTPUT PARAMETERS FOR RUN 4

Number of GBT links per card	24	48	48	48	48
Number of E-Links per GBT link	1	1	2	4	8
Number of E-Links per card	24	48	96	192	384
Average data packet size (B)	946	468	228	108	48

The average data packet sizes have been calculated via (2):

$$S_{\text{packet}} = \frac{1.86 \times 10^{11} Gb/s}{10^6 Hz} \div N_{\text{E-Links}} \div 8 - 13, \qquad (2)$$

where $1.86 \cdot 10^{11}$ Gb/s is the maximum bandwidth that was achieved in the test setup using 2×100 Gb/s network links, 10^6 Hz is the input rate, $N_{\text{E-Links}}$ is the number of E-Links for the given configuration and 13 is the overhead of the NetIO protocol per data packet in bytes.

Two rounds of tests were performed using different computers for the SW ROD application. In the first one SW ROD was running on the same computer that has been used for the tests described in the previous chapter. For the second round of tests an AMD-powered computer with AMD Epyc 7313P 3GHz (16 cores) CPU and 128 DDR4-2667 RAM has been used. Both computers used the same Mellanox ConnectX-5 100 GbE network adapters, which were connected to the same 100G network switch.

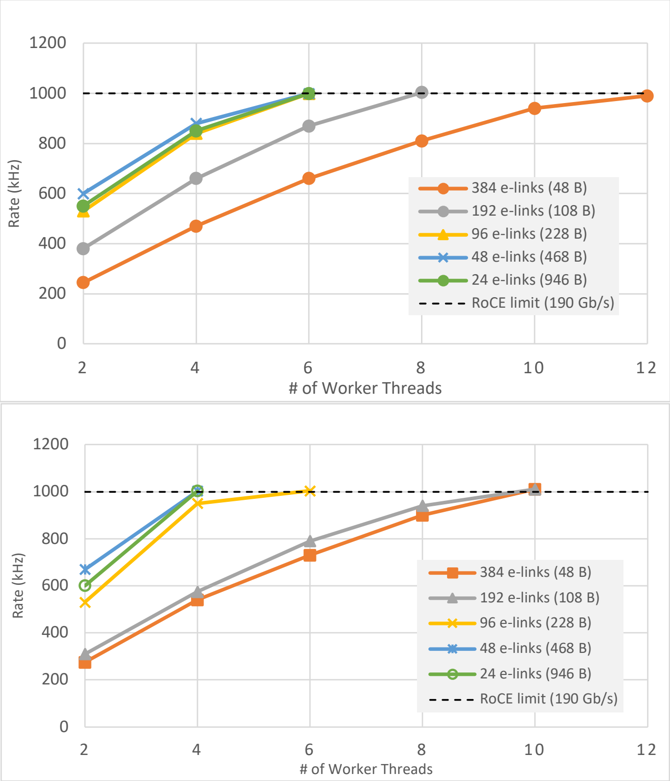

Fig. 4. SW ROD input rate on Run 3 SW ROD computer (a) and on AMD-powered computer (b).

Fig. 4a shows all results which were obtained using configurations with different numbers of input E-Links and data packet sizes. For each configuration the Fragment Builder algorithm was able to sustain 1 MHz input rate but required a different number of worker threads to be used. Results of the same tests performed on the AMD-powered computer are shown in Fig. 4b.

The test results demonstrate that the SW ROD Fragment Builder algorithm scales very well with the number of worker threads. Comparing results obtained for the two computers one can conclude that the AMD-powered configuration offers better performance due to a higher CPU frequency.

IV. CONCLUSION

The High-Luminosity Large Hadron Collider (HL-LHC), expected to enter in operation in 2029, aims to increase LHC luminosity by a factor of 10 beyond its original design. The new Readout system for the ATLAS experiment is based on the Front-End LInk eXchange (FELIX), introduced for some detectors in Run 3. A new component, called the SW ROD, has been developed to receive data from FELIX. The Data Handler component of the Run 4 DAQ system will be an evolution of the SW ROD, that will support the same functional requirements but must be able to operate at an input rate of 1 MHz to cope with the HL-LHC luminosity.

Performance testing to date demonstrates that the Run 3 SW ROD application is able to process data at 1 MHz rate for realistic Run 4 input configurations. It is expected that single CPU core performance should increase by at least 50% in the next 5 years, which will provide extra computing power and decrease overall cost of the new readout system for Run 4.

References

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

[2] S. Kolos, G. Crone, and W. P. Vazquez, "New software-based readout driver for the ATLAS experiment," *IEEE Transactions on Nuclear Science*, vol. 68, no. 8, pp. 1811–1817, aug 2021.

[3] S. Ryu, "FELIX: The new detector readout system for the ATLAS experiment," *Journal of Physics: Conference Series*, vol. 898, p. 032057, oct 2017.

[4] P. Moreira, R. Ballabriga, S. Baron, S. Bonacini, O. Cobanoglu, F. Faccio, T. Fedorov, R. Francisco, P. Gui, P. Hartin, K. Kloukinas, X. Llopart, A. Marchioro, C. Paillard, N. Pinilla, K. Wyllie, and B. Yu, "The GBT Project," 2009. [Online]. Available: https://cds.cern.ch/record/1235836

[5] J. Schumacher, C. Plessl, and W. Vandelli, "High-throughput and low-latency network communication with NetIO," *Journal of Physics: Conference Series*, vol. 898, p. 082003, oct 2017.