Integration and Commissioning of the Software-based Readout System for ATLAS Level-1 Endcap Muon Trigger in Run 3

Kaito Sugizaki[1,*], on behalf of the ATLAS Collaboration

[1]The University of Tokyo, Department of Physics and International Center for Elementary Particle Physics

Abstract. The Large Hadron Collider and the ATLAS experiment at CERN will explore new frontiers in physics in Run 3 starting in 2022. In the Run 3 ATLAS Level-1 endcap muon trigger, new detectors called New Small Wheel and additional Resistive Plate Chambers will be installed to improve momentum resolution and to enhance the rejection of fake muons. The Level-1 endcap muon trigger algorithm will be processed by new trigger processor boards with modern FPGAs and high-speed optical serial links. For validation and performance evaluation, the inputs and outputs of their trigger logic will be read out using a newly developed software-based readout system. We have successfully integrated this readout system in the ATLAS online software framework, enabling commissioning in the actual Run 3 environment. Stable trigger readout has been realized for input rates up to 100 kHz with a developed event-building application. We have verified that its performance is sufficient for Run 3 operation in terms of event data size and trigger rate. The paper will present the details of the integration and commissioning of the software-based readout system for ATLAS Level-1 endcap muon trigger in Run 3.

1 Introduction

The Large Hadron Collider (LHC) [1] at CERN is the world's largest and highest-energy particle accelerator, which crosses proton bunches at the frequency of 40 MHz. During Run 2, it operated with center-of-mass energy of $\sqrt{s} = 13$ TeV, and its peak instantaneous luminosity reached 2.1×10^{34} cm^{-2}s^{-1}, delivering integrated luminosity of 156 fb^{-1} to the ATLAS experiment [2]. For Run 3 scheduled from 2022 to 2024, it is expected to operate with center-of-mass energy of $\sqrt{s} = 13$–14 TeV and produce more collision data by upgrading the injector chain and applying techniques of luminosity leveling.

The ATLAS [3], a multi-purpose detector located at one of the interaction points of the LHC, detects particles produced in proton-proton collisions and searches for signals of physics of interest. Due to limitations of readout data bandwidth and offline computing resources, it is necessary to select "interesting" collision events online. As shown in Figure 1, the ATLAS realizes this online selection and readout with the ATLAS Trigger and Data Acquisition (TDAQ) system [4], involving two levels of trigger systems. Level-1 (L1) Trigger is the first-stage hardware-based trigger, which reduces the event rate to 100 kHz with a fixed

*e-mail: kaito.sugizaki@cern.ch

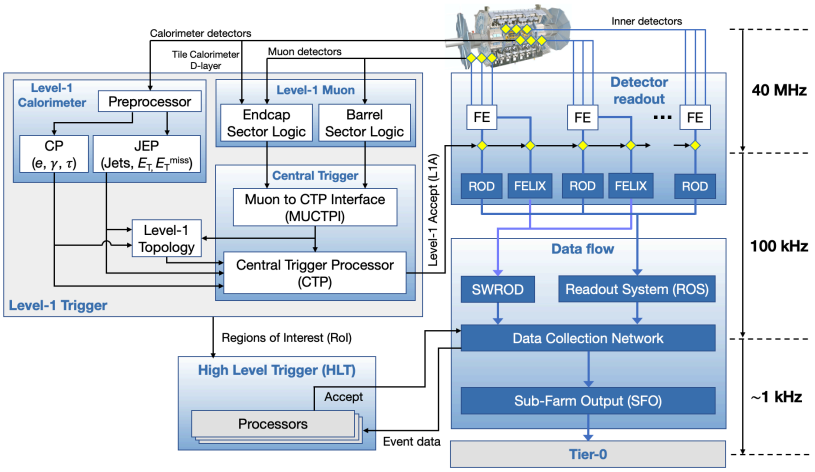

Figure 1: Schematic overview of the ATLAS TDAQ system in Run 3.

latency of ~ 2.5 μs, based on coarse information from muon and calorimeter detectors. Its final decision is made by the Central Trigger Processor (CTP), which outputs Level-1 Accept (L1A) signals to the detector front-end. The accepted event data are read out to the Readout System (ROS) or SWROD [5], new software based readout drivers replacing ROD and ROS. These data are then processed by the second-stage software-based trigger called the High Level Trigger (HLT), which selects events out of Level-1 accepted events at about 1.5 kHz. Finally, the selected event data are transferred to the Sub-Farm Output (SFO) and then to Tier-0 for temporary and permanent storage respectively for offline data analysis.

Figure 2 (left) shows the detector layout for ATLAS Level-1 endcap muon trigger, which identifies muon candidates with high p_T in the endcap regions ($1.05 < |\eta| < 2.4$) using a two-step coincidence logic: Big Wheel (BW) Coincidence and Inner Coincidence. In the former, the p_T of muon candidates are roughly calculated by taking coincidence within the three

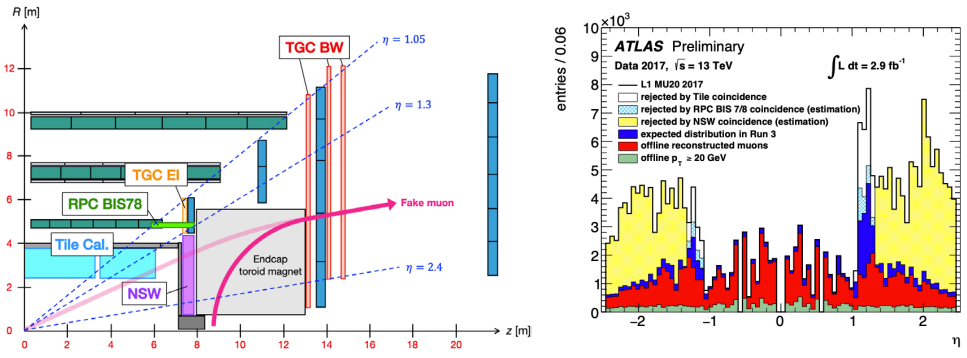

Figure 2: (Left) Detector layout of ATLAS Level-1 endcap muon trigger. RPC BIS78 and New Small Wheel (NSW) are newly introduced from Run 3 to further reduce fake triggers. (Right) Pseudorapidity distribution of muon candidates collected by single muon trigger with a threshold of $p_T > 20$ GeV at Level-1 (L1_MU20) in Run 2 and expected in Run 3 [6]. In $|\eta| > 1.05$, more than half of the muon candidates selected online were not reconstructed offline, suggesting fake dominance in the endcap regions.

outer stations of Thin Gap Chamber (TGC), called Big Wheels. In the latter, coincidence between TGC BW and inner muon detectors are taken to reduce contaminations of muons below the p_{T} threshold and fake backgrounds due to protons emerging from the beam pipe and endcap toroid magnets. A good performance of the reduction is mandatory to maintain the p_{T} threshold low to keep the physics acceptance high. For Inner Coincidence during Run 2, TGC Endcap Inner (EI) and the outermost layer of Tile Calorimeter were used in $1.05 < |\eta| < 1.3$ and TGC Forward Inner (FI) [1] was used in $1.3 < |\eta| < 1.92$. Nevertheless, muon triggers from the endcap regions were still fake dominant as shown in Figure 2 (right). Therefore, from Run 3, new detectors called RPC BIS78 and New Small Wheel (NSW) will be introduced in the Level-1 endcap muon trigger system to reinforce the Inner Coincidence logic. RPC BIS78 will cover $1.05 < |\eta| < 1.3$ in the "small sectors", which do not have TGC EI due to the layout of the barrel toroid magnets. NSW will replace the current Small Wheel including TGC FI and provide high-resolution track measurements with new angle information in $1.3 < |\eta| < 2.7$. It is expected that 90% of fake muon triggers will be rejected in Run 3, resulting in approximately 45% reduction of Level-1 muon trigger rate [4, 7].

72 New Sector Logic (SL) boards have been newly developed to handle data from five types of detectors (i.e. TGC BW, TGC EI, Tile Calorimeter, RPC BIS78, and NSW) and to process final trigger decisions for the Level-1 endcap muon trigger. The New SL is equipped with a XILINX Kintex-7 FPGA (XC7K410T) [8], which has about 20 times larger memory resource than its predecessor in Run 2, to process triggers using interface for high-speed optical serial links and the aforementioned two-step coincidence logic. The final trigger output, including muon p_{T} and Region-of-Interest (RoI) information, is sent to the Muon-to-Central Trigger Processor Interface (MUCTPI), which combines outputs from the Level-1 endcap and barrel muon triggers. In addition, six TTC Fanout Boards (TTCFOB) receive and distribute Timing, Trigger and Control (TTC) [9] signals to New SLs and also handle busy signals of the Level-1 endcap muon trigger system.

2 The software-based readout system for ATLAS Level-1 endcap muon trigger in Run 3

For offline validation and performance evaluation of the new Level-1 endcap muon trigger algorithm, it is necessary to read out the trigger logic input and output data of New SLs. The trigger readout system, formerly based on hardware, will be upgraded for Run 3 to improve flexibility and to cope with the large increase in trigger input data due to the introduction of RPC BIS78 and NSW. For the Level-1 endcap muon trigger in Run 3, we have introduced a new software-based readout system to realize trigger readout at 100 kHz.

Figure 3 shows the overview of the newly designed software-based readout system for the Level-1 endcap muon trigger in Run 3. One unit of the Level-1 endcap muon trigger readout system consists of 12 New SLs, one TTCFOB, an Ethernet switch, an event-building application called Software-based Readout Driver (SROD) [2] running on a commercially available high-performance server (SROD server), and PCIe S-LINK [10] interface card (PCIe S-LINK Card) on the SROD server. The entire readout system exploits six units, and one ROS is responsible for the readout from all of them.

When New SL and TTCFOB receive L1A signals, data for the corresponding events are sent via TCP/IP protocol to SROD server, using a hardware-based TCP processor fully

[1] Note that TGC FI will be replaced by New Small Wheel in Run 3 and therefore is not shown in Figure 2 (left).

[2] SROD is custom developed software for the Level-1 endcap muon trigger system and should not be confused with SWROD mentioned in Section 1 and Figure 1. SROD only replaces the legacy hardware-based ROD whereas SWROD replaces both ROD and ROS.

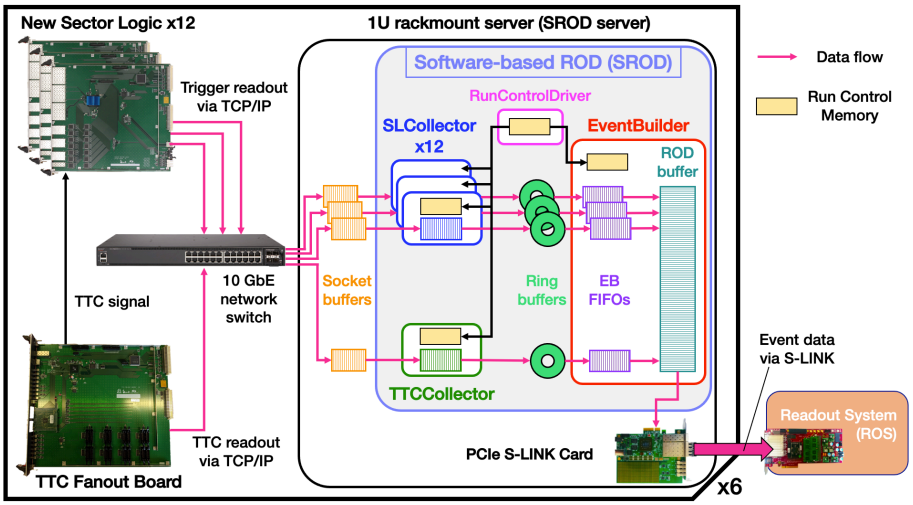

Figure 3: Overview of the software-based readout system for ATLAS Level-1 endcap muon trigger in Run 3. The system consists of New SLs, TTCFOBs, network switches, SRODs, S-LINK Cards, and ROS. The software structure of the SROD has been shown in detail.

implemented in the FPGA (SiTCP [11]) to enable high-rate output. SROD is required to build event data in accordance with the L1A rate at 100 kHz. The PCIe S-LINK Card transfers processed event data from SROD to ROS using S-LINK protocol via optical fibers. In this new trigger readout system, SRODs running on six servers play an essential role, as they build event data for the whole Level-1 endcap muon trigger.

As shown in Figure 3, SROD, written in C++, adopts a multi-process architecture with TCP sockets and shared memories to realize required performance. Three types of applications run in parallel: Collectors, EventBuilder, and RunControlDriver. Collectors, 12 SLCollectors and one TTCCollector in particular, establish socket connections with 12 New SLs and TTCFOB for data collection. Collected data from socket buffers are then sent to ring buffers, which are prepared as shared memories to allow multi-process access. The ring buffers have been designed to absorb data arrival time differences between Collectors.

EventBuilder builds events using data fragments collected from ring buffers. Data collected from each ring buffer are first put into the corresponding EventBuilder FIFO (EB FIFO) to check event ID consistency between the fragments. Using the data fragments received in 13 EB FIFOs, event data are built in accordance with the ATLAS raw event format [12] and are put into ROD buffer. Finally, EventBuilder sends the event data in ROD buffer to ROS via S-LINK protocol by using the S-LINK Card. Since these steps are carried out sequentially, EventBuilder application has the heaviest load out of all SROD applications. Therefore, the performance of SROD and hence the Level-1 endcap muon trigger readout system are determined by the processing speed of EventBuilder. It is required that the sequential "build-send cycle" be processed within 10 μs on average to cope with the L1A rate at 100 kHz.

RunControlDriver enables synchronization of independent applications in SROD. Each application in SROD has a shared memory (Run Control Memory, RCMemory) to retain information of run control state and command. RunControlDriver receives run control information from the ATLAS central system [13] and propagates it to the RCMemories of all applications. Each application operates in obedience to the state and command stored in its RCMemory.

3 System integration, commissioning, development, and performance

3.1 System integration

The installation of six SROD servers in the ATLAS counting room (USA15) was completed in June 2020. We have adopted Supermicro SuperServer 5019P-WTR [14], 1U rackmount servers with 12-core CPU, for the SROD servers. The SROD applications have been successfully integrated into the ATLAS TDAQ system by an extension of the Object Kernel Support (OKS) [15], the ATLAS standard online software database, with the new SROD applications (i.e. Collectors, EventBuilder, and RunControlDriver). By defining the appropriate hierarchical structure of SROD application objects in the OKS, we have realized systematic control of the new readout system in terms of enabling and disabling, which has widely facilitated the readout system commissioning and development.

3.2 Commissioning and development

As mentioned in Section 2, EventBuilder is a single process and must complete its sequential build-send cycle within 10 μs on average. This has been achieved by scrupulous investigation of the EventBuilder processing time and behavior. For instance, code optimization to reduce the use of `std::stringstream` variable has saved 5 μs of processing time per event. In addition, some faults in New SL and TTCFOB FPGA firmware were tracked down by rigorous data checks implemented in EventBuilder. Frequent exception handling caused by these faults also slowed down the EventBuilder processing speed. Corrections of these problems and many other improvements in the readout system have been carried out to realize stable operation at 100 kHz. Detailed performance of EventBuilder will be explained in Section 3.3.

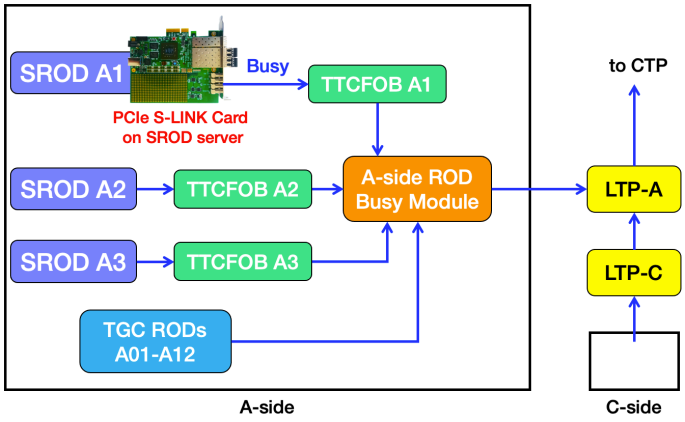

Figure 4: Schematic overview of the busy propagation system of the Level-1 endcap muon trigger. Only the A-side system has been shown, since the C-side system has a symmetrical structure.

As a handshaking mechanism with the central DAQ system, we have introduced proper busy handling procedures to realize stable operation. In the ATLAS DAQ system, the readout drivers are required to output busy signals to the CTP to avoid possible buffer overflow. The CTP stops L1A output in accordance with the busy signals, allowing time for the readout drivers to process accumulated event data. Figure 4 shows the busy propagation system

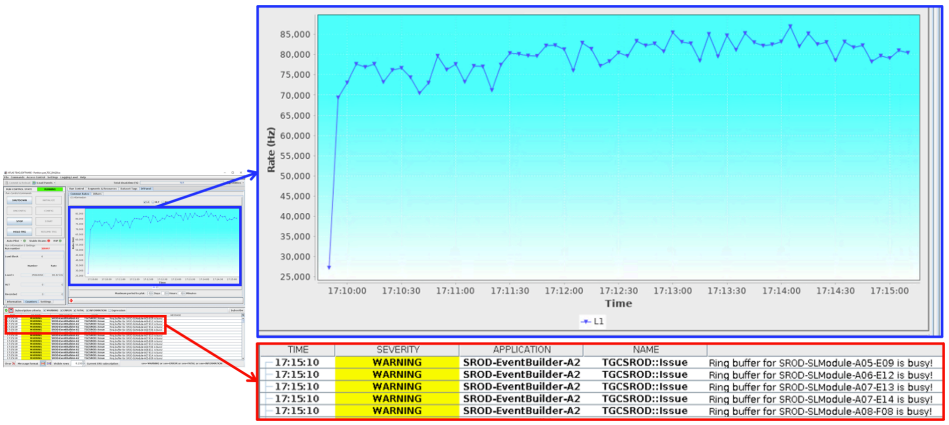

Figure 5: Run control panel of the ATLAS TDAQ software during the busy handling test. Dummy data from New SLs were adjusted to fix the SROD output data size to 11232 bits per event, which is too large for EventBuilder to handle at 100 kHz. Busy threshold of ring buffers were set to 50%. As expected, the EventBuilder application asserted busy signals, and the L1A rate was limited to about 80 kHz[3].

between SROD servers and the CTP. Busy signals asserted by SROD are first passed to the corresponding TTCFOB and then summed with busy signals from TGC RODs using ROD Busy Module. The combined busy of Level-1 endcap muon and TGC systems is propagated to the CTP via Local Trigger Processors (LTP). In SROD, EventBuilder application monitors the amount of unprocessed data in the ring buffers and asserts a busy signal when it exceeds a certain configurable threshold. The busy handling mechanism has been validated by standalone tests within the Level-1 endcap muon trigger system as shown in Figure 5.

In the course of system commissioning at various L1A rates, it was observed that readout errors occur at L1A rates lower than 10 kHz owing to Nagle buffering [16] algorithm in TCP/IP protocol, where buffers hold data until its size reaches the maximum segment size (MSS), or the data wait time reaches the Nagle timeout period to optimize overhead in packet transmission. For SiTCP, the MSS is set to 1460 bytes by default, and the Nagle timeout period is fixed to 4 ms. Considering that the average data size sent by each New SL and TTCFOB is $O(10$ bytes) per event, readout data for 50–100 events were temporarily stored in its TCP TX buffer at L1A rates higher than 250 Hz. An event request from the HLT to ROS is sent only 5–10 ms after the corresponding L1A is output by the CTP; events which are not sent to ROS in time are regarded as error events. Therefore, at L1A rates in the range of 250 Hz to about 10 kHz, the reduction of readout delay time was necessary to eliminate the error events. As possible solutions, disabling Nagle algorithm and the adjustment of MSS were both tested, however, these did not function as expected[4]. Therefore, we have introduced a "pad word" mechanism to eliminate delayed error events and to realize stable operation at all L1A rates up to 100 kHz. In this mechanism, New SLs and TTCFOBs monitor the amount of data in their TCP TX buffers. If the amount of data did not reach the MSS during a designated time period, New SLs and TTCFOBs add extra pad words after the actual readout data in their TCP TX buffers to send the data punctually in a dynamic manner. As for the

[3]The busy fraction observed in this test is lower than what is expected from the performance mentioned in Section 3.3. This is because XOFF [10] signal checks were omitted during data transfer to ROS in EventBuilder.

[4]In cases when Nagle algorithm was disabled or MSS was set to $O(10$ bytes), data corruptions were observed at high L1A rates. The cause of this is under investigation.

SRODs, the pad words are ignored in EventBuilder, and the mechanism does not hinder the event-building process. The implementation of pad word mechanism has been completed in April 2021. It has been verified that delayed error events do not occur with the mechanism enabled, using the full readout system with six SRODs, six TTCFOBs, and 72 New SLs. Also, the performance of the pad word mechanism has been measured using one New SL and one SROD as shown in Figure 6. It has been observed that the mechanism works as expected.

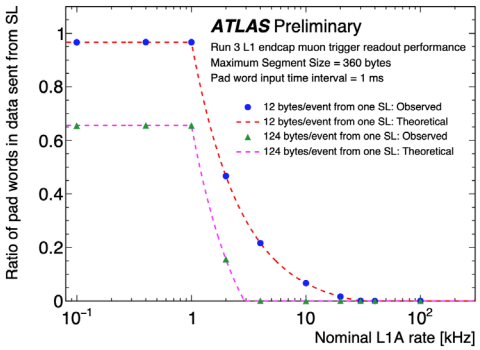

Figure 6: The ratio of pad words in data sent from one New SL to SROD. The MSS of New SL was set to 360 bytes, and fixed data of 12 bytes or 124 bytes per event were sent. Pad words were input to complement the event data in TCP TX buffer of New SL whenever further event data was not input for a time interval of 1 ms. Average of 200,000 events has been taken for each point. It has been shown that the pad word mechanism works as expected and does not hinder the event-building process, especially at high L1A rates.

3.3 Performance

As mentioned in Section 2, the performance of the trigger readout system is determined by the EventBuilder processing speed. Consequently, we have measured the EventBuilder processing speed by using the readout system which we have integrated in the ATLAS DAQ system. In particular, we conducted two tests to check

- the size of readout data which can be stably handled at the L1A rate of 100 kHz; and
- the L1A rate at which the system can stably read out data of the expected size in Run 3.

Since each of the six SRODs runs independently in the actual Run 3 setup, it is sufficient to only evaluate the performance of one of them. Therefore, we used a common setup for both tests where fixed dummy trigger data from 12 New SLs and TTC data from one TTCFOB were sent to one SROD when L1A signals were received by these boards. L1A signals were output randomly from LTP at a certain configurable average rate. Event data were built using the data fragments from the 13 boards and were sent to ROS by the SROD. 1% of the events processed by ROS were randomly selected and recorded in SFO. Consistency checks between the dummy input data sent from New SLs and the output data recorded in SFO were carried out to verify that the decoding procedures in SROD followed the correct data format. The setup of the SROD server and applications are summarized in Table 1.

Figure 7 (left) shows the EventBuilder "processing speed", which corresponds to the actual L1A rate considering the deadtime due to busy assertion by the SROD system, for various size of data sent from SROD to ROS with the nominal L1A rate (i.e. L1A rate that does not take the deadtime into account) fixed to 100 kHz. For each point, the EventBuilder

Table 1: SROD hardware and software setup used in the performance tests.

CPU	Intel Xeon Gold 6226 (12 cores, 2.70 GHz)
Memory	DDR4-2933 16 GB ×6
Network controller	Dual 10GBase-T LAN with Intel X722 + X557
Size of socket buffer	Max. 4 MB
Size of Collector buffers	12.8 kB
Size of ring buffers	1 MB
Size of EB FIFOs	1 MB
Size of ROD buffer	100 kB
Ring buffer busy threshold	50%

processing speed was calculated using 10 million events. EventBuilder was able to stably handle readout data of size up to about 4800 bits per event. For larger data, certain fractions of deadtime were observed. Since we have estimated the average data size in Run 3 to be approximately 1750 bits per event, this test has shown that the established system is able to read out data 2.5 times larger than its expected size without any deadtime due to the readout system.

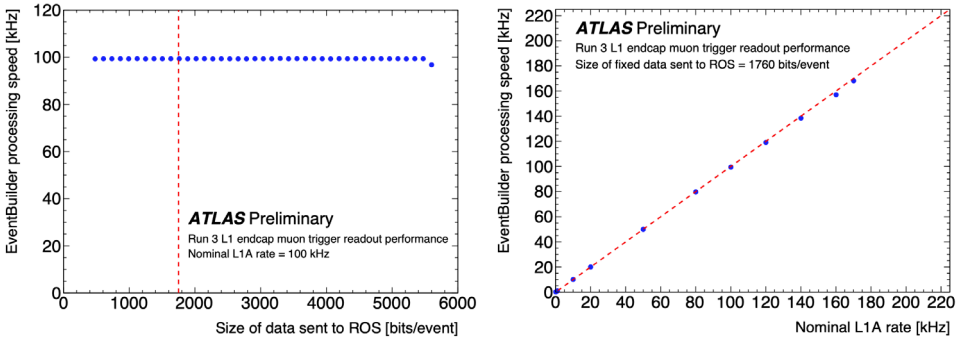

Figure 7: (Left) The relation between the EventBuilder processing speed and the size of data sent from SROD to ROS at the L1A rate of 100 kHz. The average data size in Run 3 is expected to be approximately 1750 bits per event, which is shown by the red dashed line. For events larger than 4800 bits, instantaneous busy signals were observed, despite the calculated average speed. Hence, it has been shown that the established readout system can stably handle up to about 4800 bits per event, more than 2.5 times larger than the expected size. (Right) The relation between the EventBuilder processing speed and the L1A rate for fixed readout data of 1760 bits per event. It has been proved that the Level-1 endcap muon trigger readout system can stably handle readout data of the expected size at L1A rates in the range of 100 Hz to 170 kHz without any deadtime.

Figure 7 (right) shows the relation between the EventBuilder processing speed and the L1A rate. In these tests, data sent from 12 New SLs were adjusted to fix the data size sent from SROD to ROS to 1760 bits per event, which is approximately the expected size in Run 3. The average L1A rate was varied in each run. The EventBuilder was able to stably process event data at L1A rates in the range of 100 Hz to 170 kHz without deadtime. At higher rates, deadtime due to busy signals from the SROD system was observed. This test has shown that the Level-1 endcap muon trigger readout system can withstand sudden increases in trigger rate during physics runs in Run 3.

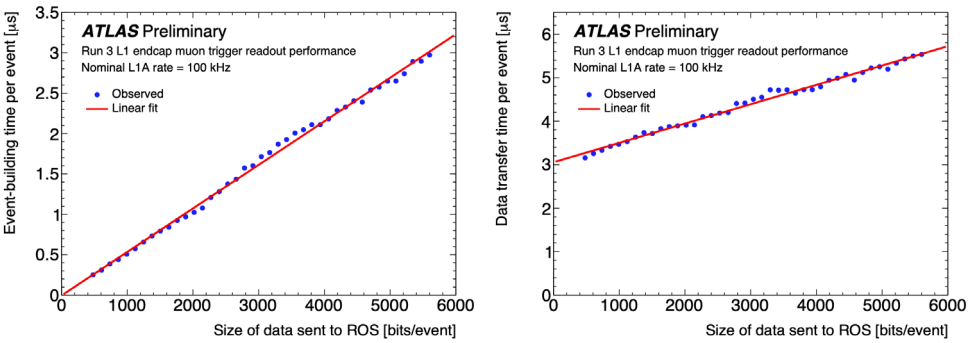

Figure 8: (Left) The relation between the event-building time per event in EventBuilder and the size of data sent from SROD to ROS at the L1A rate of 100 kHz. The event-building time is proportional to the size of readout data. (Right) The relation between the data transfer time per event in EventBuilder and the size of data sent from SROD to ROS at the L1A rate of 100 kHz. The duration of data transfer to ROS is in linear relation with the size of readout data with overhead of 3 μs.

Furthermore, the durations of processing steps in EventBuilder were studied for various data size, with the nominal L1A rate fixed to 100 kHz. Figures 8 show the event-building time per event and the data transfer time per event in EventBuilder respectively. For both steps, the durations per event increase linearly with the size of readout data. The EventBuilder performance for large-sized data is therefore bounded by these steps. Especially, the slope of Figure 8 (right) is approximately equal to 2 Gbps, which is determined by the bandwidth of S-LINK interface [17]. Overhead of 3 μs per event was observed in the data transfer step, which is suspected to be caused by some inefficiency in the S-LINK Card device driver. To achieve higher performance, system optimization to reduce the overhead is in progress as of February 2021.

4 Conclusion

In Run 3 starting in 2022, new detectors will be introduced in ATLAS Level-1 endcap muon trigger system to enhance the rejection of fake muon triggers. For validation and performance evaluation of the new trigger algorithm, the inputs and outputs of the trigger logic will be read out using a new software-based readout system. We have integrated this readout system in the ATLAS DAQ system to facilitate commissioning and development in the actual Run 3 environment. Stable trigger readout has been realized for all input rates up to 100 kHz by the improvements of SROD EventBuilder application, implementation of proper busy handling, and developments of the pad word mechanism. Performance tests have shown that the new readout system is ready for stable operation in Run 3 in terms of event data size and trigger rate. Further developments and optimizations will be carried out to realize higher performance.

References

[1] L. Evans and P. Bryant, *LHC Machine*, JINST 3 S08001, 2008.

[2] ATLAS Experiment Public Results, *LuminosityPublicResultsRun2*. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2, accessed 2021-02-07.

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

[4] ATLAS Collaboration, *Technical Design Report for the Phase-I Upgrade of the ATLAS TDAQ System*, CERN-LHCC-2013-018, ATLAS-TDR-023, 2018.

[5] S. Kolos on behalf of the ATLAS TDAQ Collaboration, *New software based readout driver for the ATLAS experiment*, arXiv:2010.14884 [physics.ins-det], 2020.

[6] ATLAS Experiment Public Results, *Performance estimation of the Level-1 Endcap muon trigger at Run 2 and Run 3*. https://twiki.cern.ch/twiki/bin/view/AtlasPublic/L1MuonTriggerPublicResults, accessed 2021-02-21.

[7] H. Hibi on behalf of the ATLAS Collaboration, *ATLAS Level-1 Endcap Muon Trigger for Run 3*, PoS LeptonPhoton2019 146, ATL-DAQ-PROC-2019-015, 2019.

[8] XILINX, *7 Series FPGAs Data Sheet: Overview*, 2020. https://www.xilinx.com/support/documentation/data_sheets/ds180_7Series_Overview.pdf, accessed 2021-02-07.

[9] CERN, *Timing, Trigger and Control (TTC) Systems for the LHC*. https://ttc.web.cern.ch, accessed 2021-02-07.

[10] O. Boyle, R. McLaren, and E. v.d. Bij, *The S-LINK Interface Specification*, 1997. http://hsi.web.cern.ch/s-link/spec/spec/s-link.pdf, accessed 2021-02-07.

[11] T. Uchida, *Hardware-Based TCP Processor for Gigabit Ethernet*, IEEE Transactions on Nuclear Science, Vol. 55, No. 3, 2008.

[12] C. P. Bee, D. Francis, L. Mapelli, R. McLaren, G. Mornacchi, J. Petersen, and F. J. Wickens, *The raw event format in the ATLAS Trigger & DAQ*, ATL-D-ES-0019, 2016.

[13] The ATLAS TDAQ Collaboration, *The ATLAS Data Acquisition and High Level Trigger system*, JINST 11 P06008, 2016.

[14] Supermicro, *SuperServer 5019P-WTR*. https://www.supermicro.com/ja/products/system/1U/5019/SYS-5019P-WTR.cfm, accessed 2021-02-07.

[15] R. Jones, L. Mapelli, Y. Ryabov, and I. Soloviev, *The OKS Persistent In-memory Object Manager*, IEEE Transactions on Nuclear Science, Vol. 45, No. 4, 1998.

[16] J. Nagle, *Congestion Control in IP/TCP Internetworks*, Request for Comments 896, 1984. https://tools.ietf.org/html/rfc896, accessed 2021-02-07.

[17] A. Ruiz and E. v. d. Bij, *Implementation of S-LINK Using Physical Layer at 2.5 Gbps*, 2002. https://hsi.web.cern.ch/s-link/devices/hola/hw_spec.pdf, accessed 2021-02-07.