EPJ Web of Conferences **295**, 02005 (2024) https://doi.org/10.1051/epjconf/202429502005
CHEP 2023

Benchmarking Data Acquisition event building network performance for the ATLAS HL-LHC upgrade

Mikel Eukeni Pozo Astigarraga[1,*], *Matias* Bonaventura[1,**], *James* Maple[1,***], *Ezequiel* Pecker Marcosig[2,****], *Giacomo* Levrini[3,†], and *Rodrigo* Castro[1,2,‡]

[1]CERN
[2]University of Buenos Aires
[3]University of Bologna

Abstract. The ATLAS experiment's data acquisition (DAQ) system will be extensively updated to take full advantage of the High-Luminosity LHC (HL-LHC) upgrade, allowing it to record data at unprecedented rates. The detector will be read out at 1 MHz, generating over 5 TB/s of data. This design poses significant challenges for the Ethernet-based network, which will have to transport 20 times more data than during Run 3. The increased data rate, data sizes and number of servers will exacerbate the TCP Incast effect observed in the past, making it impossible to fully exploit the capabilities of the network and limiting the performance of the processing farm. We present exhaustive and systematic experiments to define buffering requirements in network equipment to minimise the effects of TCP Incast and reduce the impact on processing applications. Both deep and shallow buffer switches were stress-tested using DAQ traffic patterns in a test environment at approximately 10% of the expected HL-LHC DAQ system size. As the desired HL-LHC system hardware is not currently available and the laboratory size is significantly smaller, the tests aim to extrapolate buffer requirements to the expected operating point. A novel analytical formula and new simulation models have been developed to cross-validate the results. The results of these evaluations will contribute to the decision-making process for the acquisitions of network hardware for the HL-LHC DAQ.

1 Introduction

In the coming years (expected 2026 to 2028), the ATLAS experiment [1] at CERN will undergo a significant upgrade to adapt to the new conditions provided by the High-Luminosity Large Hadron Collider (HL-LHC) [2]. The new HL-LHC will generate as many as 200 proton-proton collisions per bunch-crossing and, to fully exploit its physics potential, trigger rates must be increased by a factor of 10 compared to previous runs. To accommodate

*e-mail: eukeni.pozo@cern.ch
**e-mail: matias.alejandro.bonaventura@cern.ch
***e-mail: jamesm@jamesmaple.co.uk
****e-mail: emarcosig@dc.uba.ar
†e-mail: giacomo.levrini@bo.infn.it
‡e-mail: rcastro@dc.uba.ar

these higher rates, upgrades to both the detectors and the Trigger and Data Acquisition system (TDAQ) will be required [3]. The TDAQ system is tasked with collecting data from the detectors, selecting the interesting fraction and recording the selected data. The HL-LHC TDAQ system is expected to read out detectors at 1 MHz, resulting in a total throughput of approximately 40 Tbps out of which only about 1% will be selected for recording.

The Data Acquisition (DAQ) network has to support 20 times more data compared to previous runs which becomes additionally challenging given the latest trends in the network hardware industry. In the last few generations of network devices, memory scaling has not kept pace with the scaling of the link speeds, making it challenging to increase packet buffers linearly with the port density and throughput of the routers [4, 5]. Due to the inherently bursty and synchronous nature of DAQ traffic, network buffers play a critical role.

Figure 1 shows the HL-LHC TDAQ architecture. Compared to previous systems, all data will be transported from the Readout (RO) to the Event Filter (EF) posing additional challenges for the DAQ network connecting them. Moreover, all data related to a single bunch-crossing is processed by a single EF server, generating an instantaneous burst of data from the multiple RO sources down to a single EF server. These conditions generate increased usage of intermediate network buffers which, if saturated, worsen the TCP Incast effect already observed in the past [6, 7] making it impossible to fully exploit the capabilities of the network, limiting the performance of the EF processing farm.

Ensuring seamless transfer of detector data to the EF requires a carefully designed network with enough buffer space to accommodate the future TDAQ data traffic, as packet drops can significantly impact the overall system performance. Over-provisioning network buffers can result in unnecessarily increased costs. However, estimating buffering requirements has always posed challenges, prompting extensive research to characterise various workflows and technologies [8]. The final system will not be available until a later stage, while the network needs to be set up in advance. Additionally, final hardware may not be available for testing as we are years ahead of the system installation. While an early-stage small-scale lab prototype can be evaluated, the challenge lies in conducting measurements at a reduced scale and using different hardware, which can hinder the accuracy of evaluations.

This paper proposes a threefold strategy to estimate network buffer requirements and the associated data acquisition event-building performance:

- **Systematic experiments:** a series of exhaustive systematic experiments are conducted to measure buffer utilisation in a controlled test environment, representing approximately 10% of the expected size of the HL-LHC DAQ system.
- **Analytical model:** a simple analytical model is developed specifically for the DAQ traffic patterns. This analytical model captures the overall system behaviour, providing the grounds for fitting and numerical extrapolation of lab measurements.
- **Simulation model:** a discrete-event model, which was widely used during previous ATLAS runs [9], is extended and validated against lab measurements. This simulation model is exercised to asses the performance and buffering requirements for the number of Readout nodes expected for the final HL-LHC system.

Together, these three complementary strategies allow for cross-validation and provide a reasonable level of confidence in the performance predictions. Moreover, these predictions will guide the definition of specific network devices that will be deployed in production for the future DAQ HL-LHC system.

The rest of the paper is organised as follows. In Section 2 other related research is discussed together with an introduction to the simulation framework used. In Section 3 the main tools and methodology used to gather results are described. In Section 4 measurements and results are discussed. Finally, Section 5 presents future works and concludes the paper.

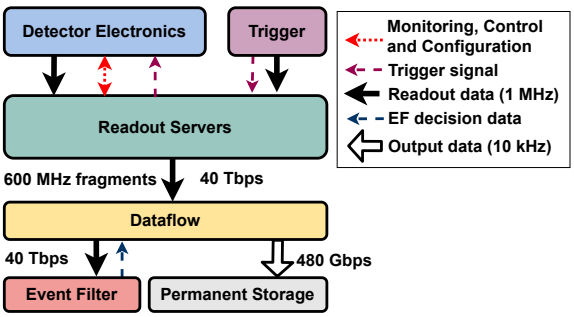

Figure 1. TDAQ components and architecture for the ATLAS HL-LHC. The Readout system (green) forwards trigger signals from the Trigger (purple) to detectors' electronics (blue) and forwards readout data to the Dataflow system (yellow). The data are provided to the Event Filter (red) upon request for analysis. Selected data are aggregated and eventually sent to permanent storage.

2 Related work

Packet buffers in network devices have been the subject of extensive research since the early days of the Internet. Despite this extensive body of work, which includes well-grounded theory [8], extensive simulations [10] and empirical experimentation, the optimal sizing of packet buffers remains a difficult discipline, where each real system deserves its own crafted sizing technique. Packet buffers significantly contribute to the uncertainty of network traffic: they cause queuing delays and jitter, packet losses (when overfilled), and they can degrade throughput. Operators typically follow the Bandwidth-Delay-Product rule-of-thumb [11] to provision buffers, an empirical rule conceived in the 90s for few long-lived TCP connections. This is the opposite scenario for the DAQ system, which supports thousands of micro-burst TCP flows over a high-speed network. Mathematical models span from dynamic differential equations to graph-theory, stochastic processes and probability theory [8]. Specific models have been derived for TCP incast problems in data centres [12], but are often tailored to specific workloads. They lack key parameters for DAQ-like systems, and estimating appropriate values for other parameters can be very challenging.

Previous data taking sessions in the ATLAS experiment have encountered similar challenges. Solutions to the TCP incast effect that are used in data centres were evaluated in the DAQ context [13] showing improved performance but requiring kernel recompilation or modifying network hardware. A server-based software switch with nearly limitless buffer to provide lossless operation showed good performance [7], but the port density and fault tolerance rendered this design inappropriate. A software-based traffic shaping solution showed to control buffer overflows when using small buffer switches but at the expense of increased latency and constrained maximum performance [6]. The traffic shaping algorithm implemented in the study focused on controlling dedicated per-port buffers. However, modern switches typically offer shared buffers across all ports, which diminishes the effectiveness of the algorithm. In a recent performance study conducted within the context of the LHCb experiment and utilising Ethernet RocEv2 (instead of TCP as in this study), a comparison between deep and shallow buffer switches was performed. The study revealed that the shallow buffer switch exhibited sub-optimal performance, while the deep-buffer switch demonstrated satisfactory performance but only when dealing with a limited number of data sources [14].

The simulation model for the TDAQ system relies on the Discrete-Event System Specification (DEVS) [15], the most general mathematical formalism for modelling discrete-event systems. DEVS is capable of representing discrete event, discrete time and continuous dynamics combined in a mathematically sound way. A model in DEVS is described as a hierarchical composition of behavioural (atomic) models and structural (coupled) models. This fact makes DEVS a natural choice to map the hierarchical system in Figure 1, where coupled and atomic models represent network components.

The PowerDEVS [16] simulation toolkit was chosen as it proved suitable to model data networks either from a microscopic (packet-to-packet) and macroscopic (fluid-flow) approaches [17, 18]. PowerDEVS allows defining model structure on a Graphical User Interface (GUI) suitable for non-experts, or using Python scripting to build large complex models [19]. Advanced users can build DEVS atomic behavioural models using C++.

The TDAQ model for the HL-LHC system is an extension to models extensively used during the previous ATLAS runs [9], which includes low-level model libraries (e.g. delays, priority queues, routing tables and egress ports), high-level model libraries (e.g. TCP/UDP senders/receivers, switches and routers implementing different congestion control mechanisms), as well as TDAQ application models. Most of the high-level components were validated against their real counterparts.

3 Methodology and Tools

The ATLAS TDAQ system exhibits a bursty and synchronised traffic pattern that places significant demands on network device buffers [6]. Data generated at the ATLAS detector undergo temporary buffering in the RO system. At the HL-LHC, data associated with each bunch-crossing (called **event** and estimated to be 5 MB) will be buffered in approximately 600 RO servers. Subsequently, a single EF node sends a request to all RO servers to collect the information associated with a specific event to process it and decide whether to discard or transfer it to permanent storage. Responses from all RO servers are tightly synchronised in time since they stem from the same request. Moreover, the RO and core networks exhibit significantly higher link speeds compared to the access links of the EF nodes (see Figure 2). Consequently, this leads to instantaneous data aggregation at network devices which need to buffer fast incoming data from multiple nodes into a single slower link.

As this effect multiplies with the rate at which Events are requested, increased processing rates require increased buffers in network devices. The detrimental effects of buffer overflow on performance are significant, as it leads to packet drops and subsequent TCP retransmissions. Although no data loss occurs, this process causes substantial performance degradation, preventing full utilization of the network link capacity. Consequently, this performance degradation has a direct impact on the event processing rate, ultimately diminishing the overall efficiency of the experiment. Therefore, accurate estimates of buffer utilisation are essential for making informed decisions regarding buffer provisioning in the network infrastructure and guiding future hardware acquisitions. This section describes three methods and tools employed to estimate buffer usage in network devices for the HL-LHC.

3.1 Lab test environment

Figure 2 shows the network and hardware configuration of the experiments' test environment. The lab comprises 48 RO nodes, and 2 racks with 32 EF servers. The devices under test (DUT), are the top-of-rack (ToR) switches that aggregate the EF nodes. Only one of the EF racks was used in each experiment to measure the performance of either the Juniper QFX5120-48Y ToR (QFX) [20] or Juniper ACX7100-48L ToR (ACX) [21]. Both switches have an aggregated uplink capacity of 800 Gbps to the core routers, and an access link capacity of 25 Gbps to each of the 32 EF nodes. In comparison, the network for the HL-LHC system is anticipated to have a similar topology, the same link capacities, central core routers will possess sufficient capacity to accommodate the expected traffic, approximately 600 RO nodes (lab represents 8%), and thousands of EF servers hosted in up to 100 racks. For these experiments the EF racks are considered to be largely independent, so conducting experiments with a single rack can provide reasonable estimates of the buffer usage.

The QFX switch utilises a shallow buffer architecture which includes 5 MB of dedicated packet buffer distributed across all ports, and additional 27 MB which can be configured as shared or dedicated buffer. Shared buffers are used by traffic traversing all ports and dynamically allocated based on the traffic demands across different ports. Dedicated buffers are assigned and owned by individual ports, providing isolation and ensuring that each port has its own buffer resources. The ACX deep buffer switch has 8 GB of High Bandwidth Memory packet buffer and uses a virtual output queue architecture which maintains separate buffers for each ingress/egress port combination. Since the ACX costs approximately 2.5 times as much as the QFX, there is a strong preference to deploy shallow buffer switches in the final system. For the purposes of this study, the ACX deep buffer switch serves as a reference device, representing the desired behaviour of the system under the best-case scenario.

Regarding the software, existing DAQ applications and emulators were adapted to match the expected HL-LHC system data-flow patterns. For the EF, applications distribute the load amongst all EF nodes and generate the requests to the RO servers. These applications were configured to perform full-event building: upon receiving an event identifier, EF applications initiate requests to all RO servers to gather all the associated data of the Event. It is the responses to these requests that generate micro bursts and lead to buffer saturation of the ToR switches, which are the specific phenomena we are interested in studying. Emulators were used in place of the EF applications responsible for analysing, processing, and making decisions regarding the permanent storage of detector's data. To stress the network, these emulators were configured to process up to 32 events concurrently per server and with processing times that guaranteed processing power not to be a bottleneck in any test. All events are immediately discarded after processing to asses only event collection rates. Event identifiers are also emulated with a configurable rate set differently in each test. For the RO, emulators were used to generate data upon receiving requests from the EF nodes. The data aggregated from all RO servers determines the event size, which follow a stochastic distribution with a configurable mean set differently on each test. This configuration produces synthetic event sizes, which may not perfectly mirror those of actual data-taking sessions. Instead, the focus is on generating consistent data patterns with predictable and easily configurable event sizes. Communication between applications use the TCP/IP protocol and an asynchronous custom library based on Boost Asio.

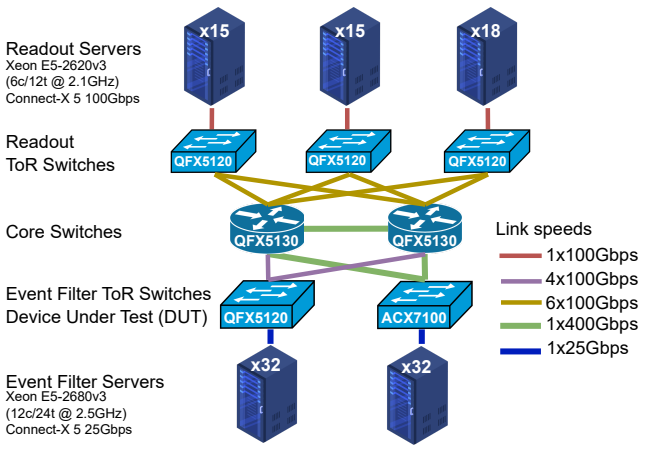

Figure 2. Topology of the Data Acquisition system prototype built in the lab. Provides access rates as expected for the HL-LHC system, except for the Readout servers that will be connected with 2x 100GbE. Only one of the devices under test (QFX and ACX) is experimented at a time.

3.2 Experiment methodology and automated tools

The objective is to determine the minimum shared buffer size required on the ToR switches for the system to operate with minimal packet drops. However, the internal metrics provided by the network devices themselves are not sufficiently accurate or fine-grained to capture the occurrence of micro bursts and short spikes, which are prevalent in this particular system. Hence, the experiments aim to estimate the minimum required buffer by setting the system conditions, sensing packet drops, and iteratively adjusting the available buffer in the ToR switch. The goal is to identify the buffer capacity at which no packet drops are observed.

Moreover, the buffers in the QFX switch are insufficient to operate at the target conditions so it is impossible to obtain the required buffer directly from the measurements. Consequently, experiments systematically vary the system conditions to observe the trend in which the minimum required buffer size at the ToR switches increases under different conditions. By sweeping the system and analysing the resulting data, the relationship between system conditions and the corresponding required buffer size can be extrapolated.

New automated tools were developed to perform long-running experiments, sweeping a big range of system conditions. These tools aided in carefully setting conditions on hardware and distributed applications, obtaining measurements from multiple sources, and running the system for sufficient time to reach equilibrium before taking measurements. Moreover, automated tools allowed for reproducible results, unattended execution to maximise the utilisation of the lab resources, flexible configuration of network devices and DAQ application parameters. They also allow different types of sweeps combining multiple parameters. Relevant to these experiments are binary searches, which enable efficient search of the minimum required buffer size with minimal iterations. Multiple executions of the same condition are performed to assess measurement errors. Automated tools are written in Python and were designed to be used beyond the scope of the experiments on this paper. The tools interact with ATLAS online systems, monitoring infrastructure and hardware devices. ATLAS control systems can set and access the initial system configuration, dynamically set parameters across multiple distributed applications, and provide metrics on running applications. Network monitoring time-series databases provide live metrics such as packet drops.

3.3 Analytical model

An analytical model of the EF ToR switch buffer occupancy profile was developed. This turned out to be very useful to exclude upfront potential candidate platforms due to insufficient buffer space, and to gain confidence about the suitability of the mathematical surface used for data extrapolation (see Section 4).

The regularity of the system parameters at a macroscopic level (i.e. event rates and event sizes) allow some degree of predictability on how much data are injected into the system, at which speed the buffer is filled (i.e. input speed at maximum) and at which speed it is emptied (i.e. output speed for a given event). With this in mind, the simplified model was built under the strong assumption that for a given event all the RO servers would reply with the corresponding data fragments almost simultaneously, resulting in a micro-burst that loads the ToR switch buffer at the speed of the switch input link. This simplification of the system is acceptable as long as the time skews of the different processes involved are below the time it takes to load the buffer. The model also ignores the impact of the TCP congestion control window, the number of EF nodes in the rack, queuing delays, stochastic processes, and assumes infinite capacity on the upstream network. These are, however, realistic assumptions due to 1) the modest amount of data that each Readout server sends and 2) the significant amount of network bandwidth installed on the core network layer.

The buffer occupancy then is modelled for each event with a triangular function: buffer occupancy increases at the speed of the input link during each event arrival; and decreases at the speed of the output link. The event rate determines the start time of these functions and their sum yields the shared buffer occupancy. The maximum is given by:

$$B_N(t)\,|_{\max} = N\left[\frac{(I-O)}{I}S - \frac{N-1}{2}\times\frac{O}{f}\right] \qquad \text{with} \quad N = \left\lfloor\frac{S\times f}{O}\right\rfloor + 1$$

where: $S \equiv$ Event size plus protocol headers and network overheads; $f \equiv$ Event frequency that needs to be absorbed by a given Event Filter ToR switch; $O \equiv$ Output bandwidth (speed at which a given event is flushed from the buffer); $I \equiv$ Input bandwidth (speed at which a given event arrives to the buffer). The formula is valid in general if $I > (N-1)O$ (i.e. the system does not push more data than the input link bandwidth). For additional details on how the formula is derived and its mathematical constraints refer to [22].

3.4 Simulation model

The simulation model for the HL-LHC TDAQ system (available at [23]) is an extension of the model in [9] to include new applications, network devices and the expected topology.

The model includes representations for the EF and RO applications which buffer, transport and process data (as described before). The EF's data analysis processes are modelled as a sequence of stochastic delays. RO response times and event rate are modelled with stochastic distributions whose parameters are determined experimentally. The simulation network topology mimics that of Figure 2. The simulation model represents network traffic at the packet level, including protocols, network transfer delays, and OSI layers 3 and above.

New simulation models for the network devices were added to incorporate the shared-buffer scheme. Additional details specific to the QFX device were added based on insights gained through experimentation. For example, the buffer in the QFX device is divided into cells of fixed size, which reduces the effective usable space, and there are limitations on the amount of buffer that a single port can utilise. These are crucial details that were taken into account to increase the simulation model accuracy in representing the operational characteristics of the QFX device. Not included in the simulation model are explicit representations of real data (which is reflected only by packet sizes), EF algorithms that analyse data (represented by their stochastic processing times), CPU and memory utilisation in the servers, and physical network links characteristics (represented only by their delays).

One key strength of the simulation model is its flexibility to represent multiple scenarios and capture fine-grain details across network and application metrics (for example, in the transition from the lab configuration to the final HL-LHC system). In contrast to the analytical model, the simulation model can capture internal and emergent behaviours for all connected elements in the system. For instance, the model has been used to study load balancing mechanisms [17] and traffic shaping algorithms [9].

To determine the minimum buffer to prevent packet drops, the lab tests are replicated in the simulation. Simulated ToR switches are configured with an infinite buffer capacity and their maximum utilisation is measured.

4 Results

This section shows the results for the lab measurements, the analytical formula and different simulation runs. We adopt a target operational requirement of a **5 MB event size** and 1 MHZ event collection rate for the HL-LHC system. Considering an estimate of 83 EF racks in the

final system, we assume a **12 kHz event collection rate** is required for each EF ToR switch, as these switches operate independently.

Figure 3 shows the maximum achievable performance, without packet drops, as a function of the event size for the two different DUT switches. In the case of the QFX switch, the automated tools employ a binary search on the event generation rate to determine the maximum rate at which no significant packet drops occur. Conversely, for the ACX switch, no rate limitation is necessary as the deep buffer capacity ensures that no overflow occurs. The performance bottleneck for the ACX switch is primarily associated with the application and server performance rather than the network itself. The ACX switch is utilised as a baseline for comparison and to demonstrate the feasibility of the system.

The plot shows the extent of the performance degradation due to the reduced buffer capacity of the QFX switch. At the target event size of 5 MB, the system is able to collect events at 14.2 kHz with the deep buffer, surpassing the required 12 kHz target rate. In contrast, with the QFX shallow buffer switch the event collection rate drops to a mere 1.8 kHz. Such a decreased rate renders this configuration unfeasible for meeting the system's requirements. These results highlight the critical importance of an adequate buffer capacity, as the choice of switches significantly impacts the system's ability to achieve the desired data collection rates.

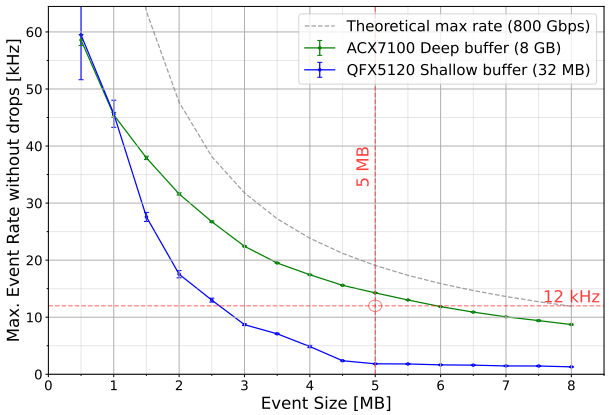

Figure 3. Comparison of ACX and QFX switches showing maximum event rate as a function of the event size without packet drops. The intersection of dashed lines indicate the expected operating point. Error bars show the standard deviation of multiple measurements. The lower performance of the QFX device is due to buffer overflows and the resulting TCP Incast effect.

The estimated **minimum required buffer sizes** to avoid packet drops on the QFX switch are shown in Figures 4 and 5 as a function of the event size and event rate respectively. Experiments cover event rates in the range [1-20] kHz and event sizes in the range [1-8] MB, but fewer cases are included in the plot for simplicity. Regarding lab measurements, the plots show that the system can not cope with high event rates or event sizes due to lack of buffer. The target rate of 12 kHz can only be achieved with events smaller than 2.5 MB. Also, the required buffer as a function of the event size fits well to a second order polynomial, and a linear function for the case of the event rate.

The results obtained with the detailed simulation model (based on microscopic first principles) align very well with the simplified analytical model (based on completely different macroscopic assumptions). However, both results differ from the laboratory measurements by an empirically identified linear factor $k \approx 1.5$. Yet, said factor is remarkably consistent and robust across all studied setups. We suggest that this linear correction can be considered as a parameter consistently absorbing all unknown and unmodelled behaviours of the equipment (possibly also capturing sources of systematic measurement bias in the switches' reporting system). We consider that this finding does not undermine the strength of the conclusions, which remain valuable and consistent to inform equipment upgrade decisions (at least within the ranges of values considered in this study).

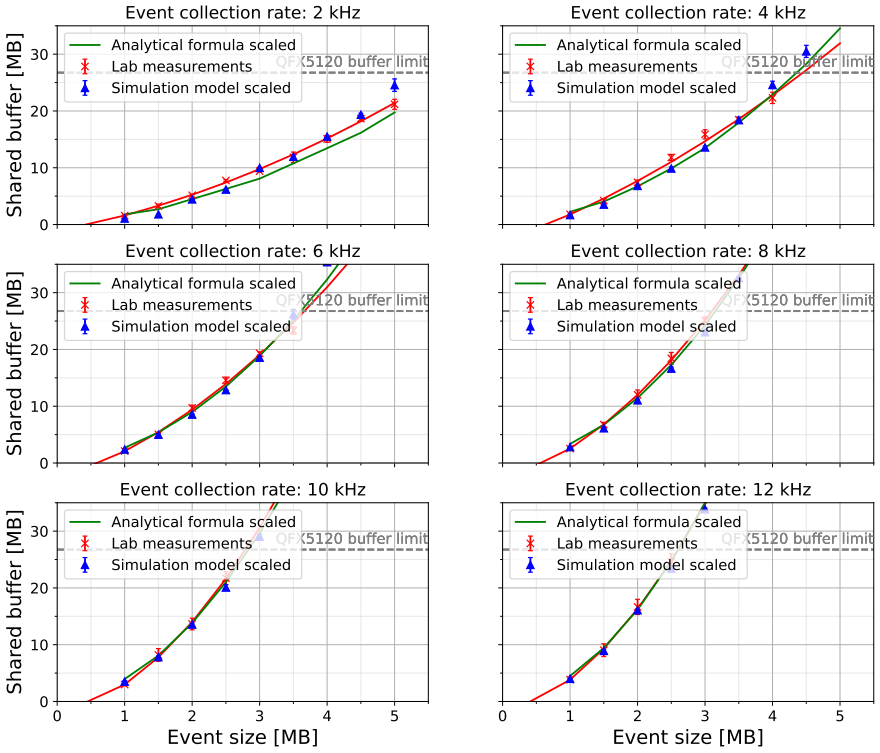

Figure 4. Minimum shared buffer size required to avoid packet drops on the QFX switch. Several options are presented for combinations of event size (x axis) and event collection rate (panels). Three different methods are applied (lab tests, simulation and analytical). Error bars depict standard deviations. Results of the simulation and analytical models are corrected by an empirically identified scaling factor $k = 1.5$ that proves consistent across all scenarios.

Finally, Figure 6 shows the estimated minimum required buffer to avoid packet drops on the QFX switch across the entire range of event sizes and event rates. Figures 5 and 4 are vertical slices of this plot. Regarding the lab measurements, the automated tools evaluated more than 300 data points, some not included in the plot as they always produced drops. With each data point requiring approximately 5 minutes and repeating 8 times, the full measurements took more than 8 days. Regarding the simulation, the plot includes 320 data points. The model was run for 5 simulation (virtual) seconds, as a stationary state is reached in approximately 1 second. The execution time of the simulation depends mainly on the traffic that traverses the network (a function of *Event Size x Event Rate*) and ranges from a couple of minutes up to 18 hours in the worst case. Single-threaded simulations were launched in batches and distributed across multiple nodes to execute concurrently.

The plot includes a statistical fit of the lab measurements, and the predictions of the analytical and simulation models, for system conditions that cannot be assessed in the lab. The fit is obtained minimising the *least squares error* of the measurements corresponding to the two dimensional input data grid (event sizes and event rates). The shape of the surface chosen for the fit is compatible with the analytical model, being a first order polynomial for a given event size and a second order one for a given event rate. The coefficient of determination obtained for the fit is significant ($R^2 = 0.99$) so the resulting surface could be

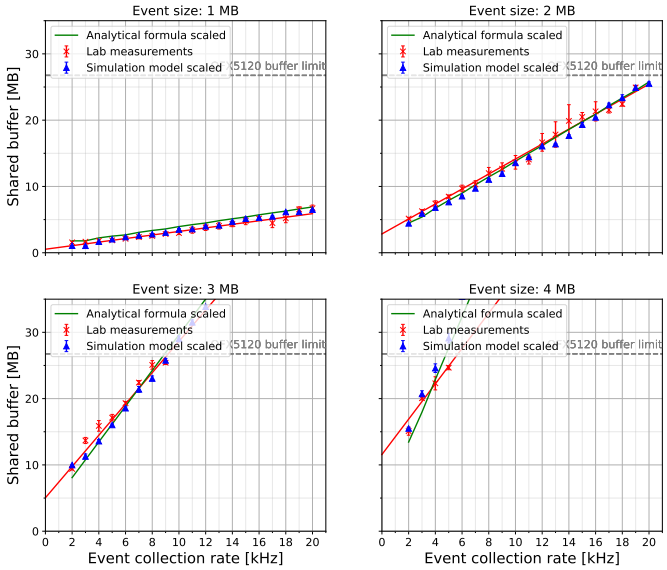

Figure 5. Minimum shared buffer size required to avoid packet drops on the QFX switch. Several options are presented for combinations of event collection rate (x axis) and event size (panels). Three different methods are applied (lab tests, simulation and analytical). Error bars depict standard deviations.

used to extrapolate to the target requirements and estimate the buffer needs to 71 ± 1 MB. The result is meaningful assuming the system scales without hitting other limitations first, in which case the buffer needs would be lower due to the traffic synchronicity would be lost. This value poses an optimistic lower bound for the predicted buffer size.

As described in Section 3.4, the simulation captures network and topology details, thus it is flexible in representing scenarios which can not be executed in the lab, nor captured by the analytical formula. Contrary to the previous figures, Figure 6 shows that for bigger event sizes and rates, the simulation model predicts a higher buffer utilisation than the analytical model. When the bandwidth utilisation (approximately $EventSize \times EventRate$) approaches its limit of 800 Gbps, the simulation model shows a rapid increase in buffer requirements due to network link saturation. This effect, which is expected in the real system, cannot be captured by the analytical formula or the statistical fit extrapolation. According to the model, the buffer size required for event rate and size of 12 kHz and 5 MB is 109 ± 4 MB This represents a conservative upper bound for the predicted buffer size. Ultimately, these bounds narrow down the design space to help engineers to take informed decisions.

Additionally, the simulation model was configured with the final number of 600 Readout nodes, providing a more realistic representation of the HL-LHC system compared to the 48 nodes available in the lab. Although the micro-burst sizes remain the same (as the event size is unchanged), the responses from the Readout nodes arrive more concurrently at the ToR, requiring additional buffering capacity (115 ± 3 MB).

5 Conclusions and Future Work

With the goal of characterising network device requirements for the HL-LHC system, this paper presents a comprehensive analysis of three distinct methodologies for benchmarking ToR switches in terms of their maximum achievable event building rate based on the buffer capacity offered. Through automated experimentation in a scaled-down laboratory it was found that a shallow buffer switch with a 32 MB buffer could not achieve the desired operating point. In contrast, a deep buffer switch with an 8 GB buffer achieved the desired performance, but at a higher cost, approximately 2.5 times that of the shallow buffer switch.

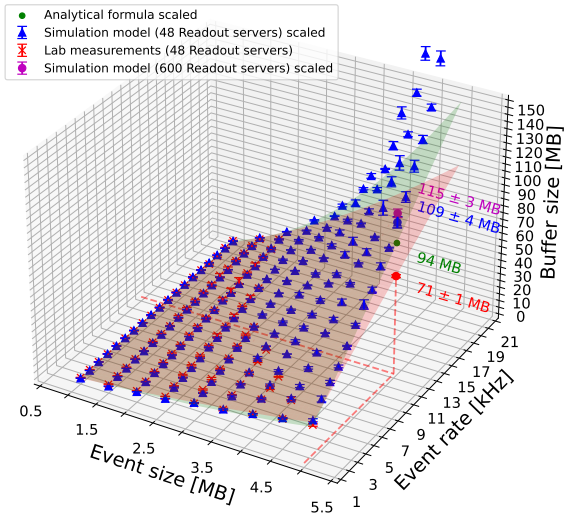

Figure 6. Minimum required buffer to avoid packet drops on the QFX switch when increasing the event rate and event sizes. Error bars depict the standard deviation of multiple measurements. Required buffer size predicted by all methods is highlighted for the desired operating point at 12 kHz and 5 MB.

An analytical formula and a discrete-event simulation model were developed to obtain design values at system scales not achievable with the laboratory hardware. The analytical formula was specifically developed for the ATLAS DAQ case. The simulation model extended a previously developed packet-level model for the ATLAS DAQ taking into consideration application algorithms, topology, network protocols, and queuing effects. New automated tools were used to measure the minimum buffer requirements under increasingly demanding conditions. Observing the trends in buffer needs confirmed the formula and simulation predictions, concluding that buffer requirements increase linearly with respect to the Event Rate and quadratically with respect to the Event Size.

Finally, to estimate buffer requirements for the expected working point of the HL-LHC, the formula and simulation were exercised on these conditions and measurements from the lab were extrapolated using a statistical fit. All three methods yielded similar results, indicating buffer requirements ranging from 71 MB to 109 MB. Furthermore, the simulation model was exercised using the final system setup of 600 RO servers, which predicted a slight increase in buffer requirement to 115 MB.

These findings help guiding future acquisitions of network devices, ruling out inadequate switch models. Although all three methods converged within approximately 15% of each other, there were discrepancies among them under low network load conditions. As a result, caution should be exercised when interpreting these findings and safety margins should be considered. The new simulation model and automated lab experimentation tools are valuable assets currently used to study varied aspects of the system, and will be important to assess future hardware versions. The simulation model allows cross-validating the system in terms of application design and scalability. The analytical and simulation models will be refined to better match lab measurements in low network load scenarios and to better explain the empirically identified scaling factor.

References

[1] ATLAS Collaboration, JINST **3**, S08003 (2008)
[2] ATLAS Collaboration, PoS **BEAUTY2018**, 055 (2018)

[3] ATLAS Collaboration, Tech. Rep. ATLAS-TDR-029, CERN, Geneva (2017)
[4] S. Yeluri, *Sizing router buffers - small is the new big* (2023), accessed on: 2023-05-08, https://community.juniper.net/blogs/sharada-yeluri/2023/02/22/sizing-router-buffers
[5] G. Appenzeller, I. Keslassy, N. McKeown, *Sizing Router Buffers*, in *Proc. of SIGCOMM'04* (2004), p. 281–292, ISBN 1581138628
[6] T. Colombo, on behalf of the ATLAS Collaboration, *Data-flow Performance Optimisation on Unreliable Networks: the ATLAS Data-Acquisition Case*, in *J. Phys.:Conf. Ser.* (2015), Vol. 608, p. 012005
[7] G. Jereczek, G. Lehmann Miotto, D. Malone, M. Walukiewicz, *A Lossless Network for Data Acquisition*, in *IEEE Trans. Nucl. Sci.* (2017), Vol. 64, pp. 1238–1247
[8] N. Bhatnagar, *Mathematical Principles of the Internet, Volume 1: Engineering* (CRC Press, 2018)
[9] M. Bonaventura, D. Foguelman, R. Castro, *Discrete event modeling and simulation-driven engineering for the ATLAS data acquisition network*, in *Comp. Sci. Eng.* (2016), Vol. 18, pp. 70–83, doi:10.1109/MCSE.2016.58
[10] K. Wehrle, M. Günes, J. Gross, *Modeling and tools for network simulation* (Springer Science & Business Media, 2010)
[11] C. Villamizar, C. Song, *High performance TCP in ANSNET*, in *SIGCOMM Compt. Commun. Rev.24* (1994), pp. 45–60
[12] M. Alipio, N.M. Tiglao, F. Bokhari, S. Khalid, *TCP incast solutions in data center networks: A classification and survey*, in *J. Net. Comp. App.* (2019), Vol. 146, p. 102421
[13] G. Jereczek, G.L. Miotto, D. Malone, *Analogues between tuning TCP for data acquisition and datacenter networks*, in *Int. Conf. Comm.* (IEEE, 2015), pp. 6062–6067
[14] R. Krawczyk, T. Colombo, N. Neufeld, F. Pisani, S. Valat, IEEE Transactions on Parallel and Distributed Systems **33**, 3640 (2022)
[15] B. Zeigler, A. Muzy, E.Kofman, *Theory of Modeling and Simulation*, 3rd edn. (Academic Press, San Diego, CA, USA, 2018), ISBN 0128133708
[16] F. Bergero, E. Kofman, SIMULATION **87**, 113 (2010)
[17] M. Bonaventura, R. Castro, *Fluid-Flow And Packet-Level Models Of Data Networks Unified Under A Modular/Hierarchical Framework: Speedups And Simplicity, Combined*, in *Winter Simulation Conf.* (2018), pp. 3825–3836
[18] R. Castro, E. Kofman, *An integrative approach for hybrid modeling, simulation and control of data networks based on the DEVS formalism*, in *M&S Comp. Net. Sys.* (2015), pp. 505–551, ISBN 978-0-12-800887-4
[19] E. Pecker–Marcosig, M. Bonaventura, E. Lanzarotti, L. Santi, R. Castro, *py2PowerDEVS: Construction and Manipulation of Large Complex Structures for PowerDevs Models via Python Scripting*, in *Winter Simulation Conf.* (2022), pp. 2594–2605
[20] Juniper, *Qfx5120 datasheet*, accessed on: 2023-05-15, https://www.juniper.net/content/dam/www/assets/datasheets/us/en/switches/qfx5120-ethernet-switch-datasheet.pdf
[21] Juniper, *Acx7100 datasheet*, accessed on: 2023-05-15, https://www.juniper.net/content/dam/www/assets/datasheets/us/en/routers/acx7100-cloud-metro-routers-datasheet.pdf
[22] M. Pozo Astigarraga, Tech. rep., CERN (2023), https://cds.cern.ch/record/2882692
[23] *Powerdevs atlas tdaq simulation model*, https://gitlab.cern.ch/tdaq-simulation/powerdevs/-/tree/chep2023 (2023)