

Development of FPGA-based neural network regression models for the ATLAS Phase-II barrel muon trigger upgrade

Rustem Ospanov^{1,*}, Changqing Feng¹, Wenhao Dong¹, Wenhao Feng¹, and Shining Yang¹

¹Department of Modern Physics, University of Science and Technology of China, Hefei, China

Abstract. Effective selection of muon candidates is the cornerstone of the LHC physics programme. The ATLAS experiment uses a two-level trigger system for real-time selection of interesting collision events. The first-level hardware trigger system uses the Resistive Plate Chamber detector (RPC) for selecting muon candidates in the central (barrel) region of the detector. With the planned upgrades, the entirely new FPGA-based muon trigger system will be installed in 2025-2026. In this paper, neural network regression models are studied for potential applications in the new RPC trigger system. A simple simulation model of the current detector is developed for training and testing neural network regression models. Effects from additional cluster hits and noise hits are evaluated. Efficiency of selecting muon candidates is estimated as a function of the transverse muon momentum. Several models are evaluated and their performance is compared to that of the current detector, showing promising potential to improve on current algorithms for the ATLAS Phase-II barrel muon trigger upgrade.

1 Introduction

The ATLAS experiment at the Large Hadron Collider (LHC) is a general purpose detector observing high energy collisions of protons and heavy ions. The detector was designed for efficient detection of leptons, hadronic jets and missing transverse energy. The ATLAS physics programme includes measurements of the Higgs boson properties, discovered simultaneously with CMS in 2012 [1, 2], measurements of the Standard Model properties, and many diverse searches for new physics phenomena. Many of these measurements require efficient detection of targeted physics processes with small cross-sections. In particular, efficient selection of muon candidates is the important requirement of the ATLAS physics programme. This paper investigates new ideas for hardware-based trigger algorithms for identification of muon candidates.

The ATLAS detector [3] consists of several sub-detectors with the cylindrical geometry with the LHC beam line serving as the z -axis. The detector consists of one central barrel section and two endcap sections. The inner tracking detectors are immersed in 2 T magnetic field allowing precise measurements of the charged particle momenta. The electromagnetic and hadronic calorimeters are located outside the tracking detectors. The muon spectrometer is located outside the calorimeters and immersed in approximately 0.5 T magnetic field generated by three air-core toroidal magnets.

*e-mail: rustem@cern.ch

Interesting collision events are selected in real-time by the two-level trigger system using sophisticated data filtering algorithms (triggers). The first level trigger (L1) uses dedicated hardware algorithms in order to reduce the LHC collision event rate of 40 MHz to about 100 kHz. The second high-level trigger system uses software algorithms to further reduce the rate of selected events to about 1 kHz. The L1 trigger system uses the Resistive Plate Chamber detector (RPC) [4–6] in the central (barrel) region of the ATLAS detector for selecting muon candidates. The RPCs are fast gaseous detectors with space and time resolution of about 2 cm and 1 ns, respectively. The ATLAS RPCs are constructed as three concentric doublet layers located at radii ¹ of approximately 6.8 m, 7.5 m and 9.8 m, referred to as RPC1, RPC2 and RPC3 layer, respectively. The RPC muon trigger algorithms are implemented using application-specific integrated circuits (ASICs) that were developed specifically for the RPC trigger system [7].

The L1 muon trigger identifies muon candidates and measures their momentum using six momentum thresholds [8]. The primary single muon trigger corresponds to the transverse momentum threshold (p_T) of 20 GeV. The primary L1 muon trigger records events about 15 kHz rate at the highest instantaneous LHC luminosity in 2018. The majority of muons at the LHC are produced in decays of heavy flavour hadrons and decays of electroweak W and Z bosons [9]. The majority of the selected L1 trigger muon candidates are low momentum muons with mismeasured p_T [8].

The ATLAS detector will undergo extensive upgrades over the next several years in order to cope with higher collision rates and correspondingly higher backgrounds of the High Luminosity LHC environment. In order to maintain efficient trigger selections, the entirely new hardware trigger system will be installed that will reduce the LHC collision rate to about 1 MHz [10]. This new hardware system is referred to as the L0 trigger. As part of the trigger upgrades, three new layers of the RPC detector will be installed in the inner region of the muon spectrometer [11]. In addition, new muon trigger algorithms will be executed on field-programmable gate arrays devices (FPGAs) that will allow more flexible re-programmable trigger logic.

The present work is motivated by the availability of large FPGA devices that will be used for the L0 muon trigger system. These FPGA devices will allow implementation of more sophisticated trigger algorithms. The previous work has studied application of the convolutional neural networks for the task of identifying muon candidates [12]. In the present approach, different neural network model is applied with the goal to achieve smaller FPGA resource usage and therefore allowing implementation of several different models, including models searching in real-time for signs of new physics phenomena. The ultimate goal of our work is development of FPGA-based machine learning algorithms for identification of new physics signatures, such as decays of long lived particles in the muon spectrometer [13] or identification of slow-moving heavy charged particles [14].

This paper develops FPGA-based neural network regression models for the new L0 muon trigger system. As the first step, the aim is to improve on the current RPC system by measuring more precisely the candidate muon p_T . Achieving this result would demonstrate potential of the presented approach for reducing muon trigger rates and for freeing up bandwidth resources. The current work does not consider improvements due to the installation of three new RPC layers, aiming instead to benchmark the neural network regression model against the current ATLAS RPC detector.

¹Precise chamber position varies with ϕ -coordinate.

2 Methods

A simple simulation model of the current RPC detector has been developed [15]. The geometry of the currently existing detector is used for the present study in order to allow direct comparisons with its performance [6]. This model includes three cylindrical doublet RPC layers, with each doublet layer containing two parallel detector surfaces separated by 2 cm. Active detector surface is simulated by contiguous parallel strips with 3 cm width. Only one layer of strips is included that measures muon deflections due to the magnetic field in the (r, z) plane. Approximately 100,000 of positively and approximately 100,000 of negatively charged muons are produced at the detector origin at random angles. Muons are simulated with an uniform transverse muon momentum distribution in the range between 3 GeV and 30 GeV. Muons are then propagated through the uniform toroidal magnetic field of 0.5 T for the radius greater than 6 m. No material scattering effects are included in this simple model, although they are expected to be smaller than the RPC strip width.

A probability for a muon passing through a given strip to produce a hit is taken to be 95%. In addition, the muon has the average probability of 25% to produce one hit in a closest nearby strip and 5% probability to produce two hits in nearby strips, referred to as cluster hits. The cluster hit probability is 0% at the strip's centre and linearly increasing to 50% at the strip's edge. Finally, a probability to produce a noise (background) hit in each strip is 0.1%. These noise hits correspond to ionisation events in the real detector due to background particles, such as low momentum photons and neutrons. The above values for cluster and noise hit probability were chosen to approximate actual detector effects, based on experience of the authors with the ATLAS RPC detector [6]. More precise evaluation of these effects will be subject of the followup work. Figure 1 shows an event display of one negative muon traversing the simulated RPC detector model.

Each simulated event is processed to reconstruct clusters from adjacent hits in each doublet layer. A candidate muon corresponds to one cluster in the RPC2 layer and two matching clusters in the RPC1 and RPC3 layers that are closest to the straight line starting at the origin and passing through the centre of the RPC2 cluster (referred to as the seed cluster line). The cluster centre is the mean position of the adjacent hits within the cluster. The centres of the selected RPC1 and RPC3 clusters are required to be within a window of 0.15 m and 0.6 m for the RPC1 and RPC3 layers, respectively. The sizes of these windows were chosen to collect the majority of muons that curve in the magnetic field but still remain within the window. In events with noise hits on the RPC2 layer, two or more muon candidates are typically reconstructed.

Coordinates of the reconstructed clusters are used to compute three input features for the neural network regression model. The absolute z -coordinate of the RPC2 seed cluster (in the middle doublet layer) is the first input feature. For RPC1 and RPC3 layers, the z coordinate differences between the impact point of the seed cluster line and the cluster position provide two more input features. These differences are shown in Figure 2 separately for positively and negatively charged muon candidates, and for muon candidates that include a reconstructed cluster due to the noise hit. Figure 3 shows these differences plotted as a function of the simulated muon p_T for the muon candidates without noise clusters (pure muons) and muon candidates with one noise cluster (noise muons). There are clear correlations between the z difference and muon p_T for the pure muons while for the noise muons these differences are more scattered due to the randomness of the noise hit position.

A neural network model is trained with the PyTorch library using the linear loss function. The selected network configuration includes three hidden fully connected layers with 20 nodes each and it uses the Rectified Linear Unit (ReLU) activation function. The three input features are transformed to produce similar standard deviation and mean values in order

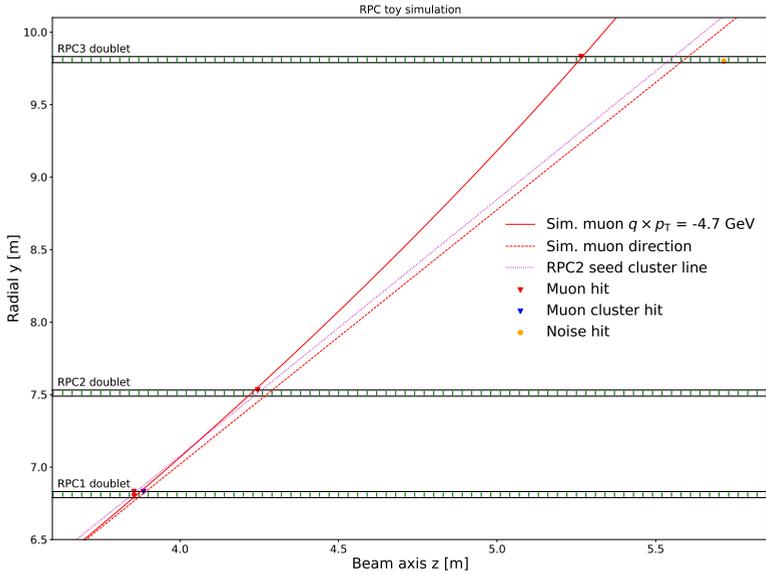


Figure 1. Illustration of one simulated muon candidate traversing the RPC detector. Six horizontal solid black lines represent six layers of the current RPC detector. Small vertical lines along these lines represent strip boundaries. Solid red line represents the muon track. Dashed red line represents the simulated muon direction at the origin. Dotted magenta line passing through the reconstructed cluster in the RPC2 layer is referred to as the seed line in the text. Coloured markers represent simulated RPC hits.

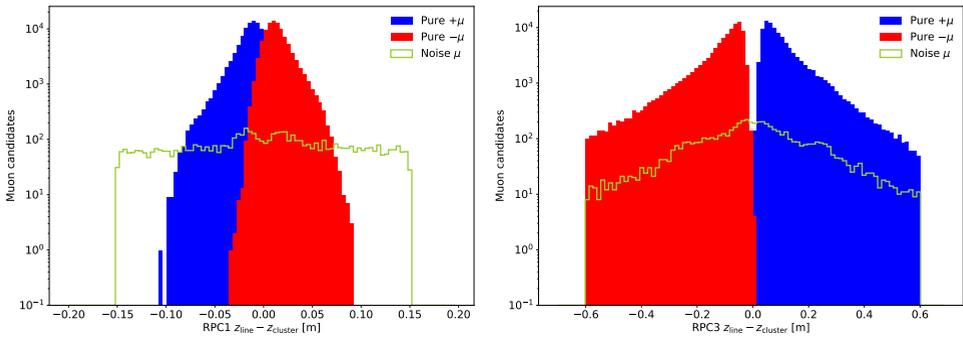


Figure 2. Differences of z positions between the impact point of the seed cluster line and the cluster positions in the RPC1 (left) and RPC3 (right) layers.

to improve network training. The inverse of the simulated muon transverse momentum times its charge (q/p_T) is used as the regression parameter and the network output is used as the prediction. The q/p_T is used to improve numerical stability of the neural network parameter optimisation since deflections of muons with $p_T > 20$ GeV is of the order of the strip width and decrease with higher p_T values. Several network with different numbers of neurons per layer were tested and no strong dependence on the network size was observed for networks with 20 or more nodes per hidden layer. Statistically independent samples were used for training and testing different network configurations.

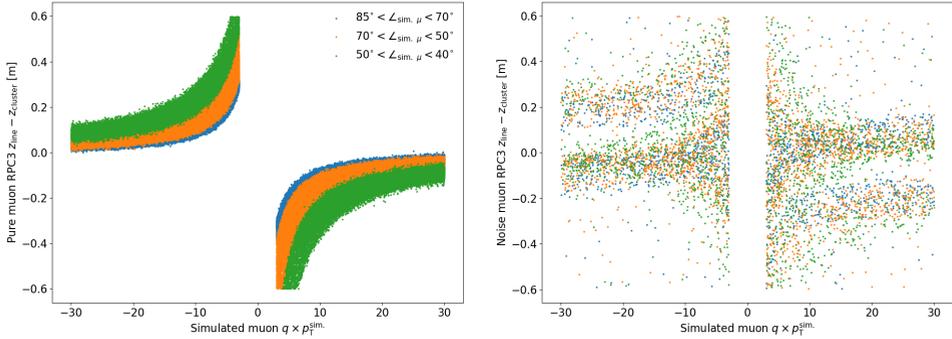


Figure 3. Differences of z positions between the impact point of the seed cluster line and the cluster positions in the RPC3 layer plotted as a function of simulated muon p_T for muon candidates without noise clusters (left) and muon candidates with one noise cluster (right).

3 Preliminary results

Two different network models were trained and tested. The first model uses training events with only muon hits and muon cluster hits but without noise hits. Figure 4 shows distribution of the predicted muon q/p_T plotted as a function of the true simulated q/p_T for this model, where test events include candidates with noise hits, shown separately as green dots. The second model uses training events with all three sources of hits. It was found that using the linear loss function allowed the training process for both models to converge to network configurations with similar performance. In initial tests, the mean squared error loss function was used leading to larger differences between two models, most likely because this loss function is more sensitive to outliers. The first model trained with the linear loss function is used for the present results.

The differences between true simulated muon $q \cdot p_T$ and $q \cdot p_T$ predicted by the model are plotted as a function of the simulated muon p_T in Figure 5. As expected, the p_T resolution is best at low p_T values and then degrades for $p_T > 20$ GeV because muon displacements due to the magnetic field became comparable to the strip width.

Performance of the neural network regression model was evaluated by computing the efficiency to select muon candidates with $p_T > 20$ GeV. Figure 6 shows the efficiency of selecting muons with $p_T > 20$ GeV for only pure muon candidates and for inclusive muon candidates that also contain candidates with noise hits. These efficiencies are compared to the reference efficiency for the so-called MU20 trigger which was measured in collision data with the current ATLAS RPC detector [6]. The efficiency of the MU20 trigger reaches the plateau at 70% due to the gaps in RPC detector coverage and inefficient RPC modules [6]. Since our simple simulation model does not include these effects, the efficiencies predicted by this model are scaled to the same plateau value as the MU20 trigger in order to allow clearer comparisons of shapes of the efficiency curves.

The main purpose of the present work is to improve resolution of muon p_T measurements by the RPC detector, which would lead to better rejection of low momentum muon candidates which dominate the acceptance rates by the RPC trigger [6]. Our regression model has a much sharper rise of the efficiency curve than the MU20 trigger. This sharper efficiency curve could lead to a lower fraction of the low momentum muons with incorrectly measured $p_T > 20$ GeV. This potentially better performance needs to be verified with data since our simulation model does not fully account for effects present in the real detector.

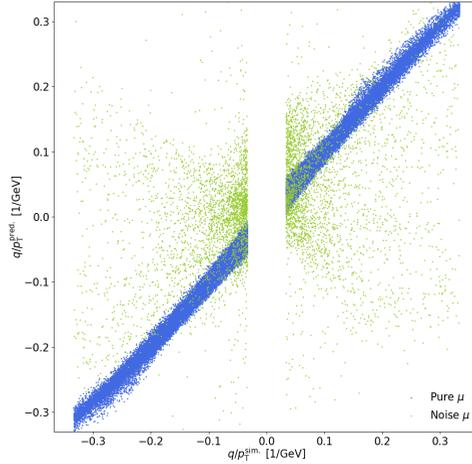


Figure 4. Output of the neural network plotted as a function of the inverse of the true simulated muon p_T using network trained with events including direct muon hits and muon cluster hits but no noise hits. Performance is evaluated using events with all three categories of hits. Blue dots show candidates with only muon and muon cluster hits. Green dots show candidates that include at least one cluster reconstructed from noise hits.

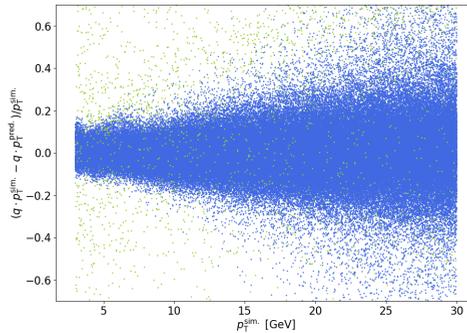


Figure 5. Resolution of predicting muon p_T plotted as a function of true simulated muon p_T using network trained with events including direct muon hits and muon cluster hits but no noise hits. Performance is evaluated using events with all three categories of hits. Blue dots show candidates with only muon and muon cluster hits. Green dots show candidates that include at least one cluster reconstructed from noise hits.

The efficiency was also evaluated for different numerical precision of the network parameters and input features. These tests were performed by limiting the number of bits for floating point precision in the range between 9 and 14 bits when evaluating network performance with test events. The performance was similar for all tests cases with the exception of the 9 bit case where the model performance starts to degrade noticeably. These tests indicate that 16 bit floating numbers can be used for the FPGA implementation of the neural networks models, therefore reducing FPGA resource requirements.

This neural network regression model is currently being implemented in FPGA code. One network layer has been already implemented with each neuron requiring one digital signal

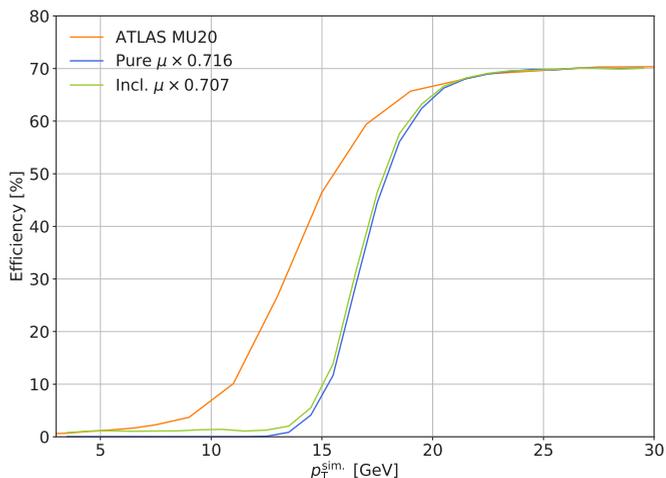


Figure 6. Efficiency of selecting simulated muon candidates plotted as a function of the true simulated muon p_T . Shown in orange is the ATLAS data efficiency for the MU20 trigger threshold of the present RPC detector [6]. Show in green (blue) is the simulation with (with) noise hits. Both curves include the effect of the cluster hits. The two curves are scaled to obtain the same efficiency at the plateau as the MU20 trigger.

processor unit. It is estimated that the full network will use about 2% of logic resources of Xilinx XCKU060 FPGA, which is a smaller device than what is envisioned for the ATLAS muon trigger upgrade. Each neuron within the layer processes its input in parallel, therefore the layer latency is determined by the latency of one neuron. Simulation results show that latency of one neuron in our implementation is 23 clock cycles. Our goal is to achieve a latency of 100 cycles for the entire network, which corresponds to 250 ns latency for 400 MHz clock. Currently, we are implementing communication links between the neural network layers.

4 Discussion and outlook

The preliminary results presented in this contribution show promising potential for more precise measurements of muon candidate p_T using the neural network regression model with the ATLAS RPC detector data. The main difficulty of our current model is rejecting muon candidates that contain simulated noise (background) hits. These candidates with noise hits significantly degrade the performance of the regression model at low p_T values. To address this challenge, several algorithms are being developed in order to evaluate probability that a given muon candidate contains clusters originating from noise hits. For the upgraded detector, it is expected that effects of the noise (background) hits could be further suppressed by using three additional RPC layers and by using more precise timing information that will be available with the upgraded detector readout electronics [11]. We are currently working on the complete implementation of our regression neural network in FPGA code. We also plan to perform more realistic tests of the model performance using ATLAS data.

References

- [1] ATLAS Collaboration, Physics Letters B **716**, 1 (2012)

- [2] CMS Collaboration, *Physics Letters B* **716**, 30 (2012)
- [3] ATLAS Collaboration, *JINST* **3**, S08003 (2008)
- [4] R. Santonico, R. Cardarelli, *Nucl. Instrum. Meth.* **187**, 377 (1981)
- [5] R. Santonico, R. Cardarelli, A. Di Biagio, A. Lucci, *Nucl. Instrum. Meth. A* **263**, 20 (1988)
- [6] ATLAS Collaboration, *Performance of the ATLAS RPC detector and Level-1 muon barrel trigger at $\sqrt{s} = 13$ TeV*, arXiv:2103.01029 (2021), <http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/MDET-2019-01>
- [7] F. Anulli et al., *JINST* **4**, P04010 (2009)
- [8] ATLAS Collaboration, *JINST* **15**, P09015 (2020), 2004.13447
- [9] ATLAS Collaboration, *Phys. Lett. B* **707**, 438 (2012), 1109.0525
- [10] ATLAS Collaboration, *Technical Design Report for the Phase-II Upgrade of the ATLAS Trigger and Data Acquisition System*, CERN-LHCC-2017-020 (2018), <http://cdsweb.cern.ch/record/2285584>
- [11] ATLAS Collaboration, *Technical Design Report for the ATLAS Muon Spectrometer Phase-II Upgrade*, CERN-LHCC-2017-017 (2017), <http://cdsweb.cern.ch/record/2285580>
- [12] Stefano Giagu, *Fast and resource-efficient Deep Neural Network on FPGA for the Phase-II Level-0 muon barrel trigger of the ATLAS experiment*, ATL-DAQ-PROC-2020-008 (2021), <https://cds.cern.ch/record/2709652>
- [13] ATLAS Collaboration, *Phys. Rev. D* **99**, 052005 (2019), 1811.07370
- [14] ATLAS Collaboration, *Phys. Rev. D* **99**, 092007 (2019), 1902.01636
- [15] <https://github.com/rustemos/MuonTriggerPhase2RPC>