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



Rustem Ospanov

Roma Tre University & NFN Sezione di Roma Tre

November 23rd, 2021

Motivation

- o Muons are important signature for the physics programme of the ATLAS experiment at the LHC
 - Electroweak studies with W & Z bosons, Higgs boson measurements, searches for new phenomena...
 - Muon trigger signatures contributed \sim 10% of the total 100 kHz bandwidth of the Level-1 hardware trigger (L1)



Outline

- 1. ATLAS muon spectrometer (MS) and RPC detector
- 2. ATLAS L1 muon barrel trigger
- 3. Muon spectrometer upgrades for the High Luminosity LHC (HL-LHC)
- 4. Neural network regression model for RPC muon trigger
- 5. FPGA implementation and simulation

ATLAS muon spectrometer (MS)

- o 2 fast detectors for L1 trigger with position resolution of \sim 1 cm:
 - Resistive plate chambers (RPCs) in the barrel region ($|\eta| < 1.05$) subject of this talk
 - Thin gap chambers (TGC) in the endcap region (1.05 $<|\eta|<$ 2.4)
 - Fast measurements of muon transverse momentum (p_T) within the 2.1 μs latency of the L1 trigger
- o 2 precision detectors for high-level trigger (HLT) and offline muon reconstruction:
 - Muon Drift Tubes (MDT) for $|\eta| <$ 2.7 with position resolution of \sim 80 μm
 - Cathode Strip Chambers (CSC) ightarrow replaced with New Small Wheel detectors



Resistive Plate Counters

o RPCs were developed by Santonico and Cardarelli in early 80s

Careful study of different designs and many materials to arrive at a working prototype

o Two parallel electrodes producing high uniform electric field

Free electron \rightarrow avalanche \rightarrow streamer

- o High bulk resistivity reduces surface area for ionisation discharge \rightarrow suppresses streamers
 - ightarrow RPCs use phenolic resin known as bakelite first synthetic plastic invented in 1907
 - $ightarrow \, {\cal O}(100 \ {
 m Hz/cm^2})$ counting rates and ${\cal O}(1 \ {
 m ns})$ time resolution
- o RPCs are low-cost detectors covering large surface areas and using gas at room pressure
 - $\rightarrow\,$ RPCs are used at the LHC by the ATLAS and CMS muon trigger systems
 - ightarrow Multi-gap RPCs are used as time-of-flight detectors, e.g. reaching \sim 40 ps resolution with 10 gaps for ALICE

ATLAS Resistive Plate Chambers

- o Parallel resistive plates (bakelite with 2 imes 10¹⁰ $\Omega \cdot cm$) are separated by 2 mm with insulating spacers
- o Induced signal is read out using orthogonal η and ϕ copper strips with 23-35 mm pitch
- o $\,\sim\,$ 1 ns total time resolution ightarrow excellent separation of proton bunches that are 25 ns apart
- o 320 MHz clock for detecting raising edge of the amplified avalanche signal ightarrow 3.125 ns wide time bins
- $\,\circ\,$ RPC operate in avalanche mode with average applied voltage of 9.6 kV \rightarrow working at the efficiency plateau



- \circ Non-flammable low-cost gas: tetrafluorethane $C_2H_2F_4$ (94.7%), iso-butane C_4H_{10} (5%), sulphur hexafluoride SF_6 (0.3%)
- o This mixture is a potent greenhouse gas ightarrow currently being phased out in EU ightarrow raising costs and environmental impact

ATLAS RPC detector

- o 3 concentric cylindrical shells of double-layer (doublet) chambers located at radii of 7, 8 and 10 meters
- o $\,\sim$ 3700 gas volumes with the surface area of \sim 4000 m^2 with \sim 360k readout strips
- o Provide 6 measurements in bending (r, z) plane and 6 measurements in non-bending (x, y) plane





LHC delivered a half of the originally designed number of collisions

- $\rightarrow\,$ Study RPC detector performance to check for possible aging effects
- $\rightarrow\,$ RPC performance paper using 2018 data: JINST 16 (2021) P07029

RPC detector response

o Measure RPC detector response with offline probe muons produced in pp collisions

- Use Z boson decays to 2 muons one muon is tag and second is probe
- Propagate probe muons in magnetic field to predict an impact point on the RPC surface
- Offline probe muon candidates are reconstructed using primarily the MDT detector
- o Detect hits associated with muon induced avalanche \rightarrow hit time and multiplicity
 - Hit is a signal induced in one strip which is above a tunable threshold of the front-end electronics

Calibrated hit time for one RPC module Zero corresponds to time of pp collisions Hit multiplicity in response to muon passage for one RPC module Efficiency is a fraction of events with at least one detected hit







RPC detector efficiency

- o Muon detection efficiency = probability to detect avalanche with ≥ 1 hit
 - Measured using events containing a muon predicted to pass through a given chamber
 - Gas gap efficiency = probability to detect a muon induced avalanche using either η or ϕ strips
- o Average RPC detector efficiency to detect a muon is $\sim 94\%$
 - Excellent detector stability during data taking in 2018
 - About 10% of RPCs were off in 2018 due to gas leaks these chambers are not shown below





Study of FPGA-based neural network regression model

RPC counting rates and ionisation currents

o Measured RPC ionisation currents and counting rates as a function of instantaneous luminosity

- Scale linearly with instantaneous luminosity, as expected
- o Also measured the mean avalanche charge = $I/R_{counts} \approx 30 \text{ pC}$
 - Consistent with test beam results ightarrow confirmed with the full RPC detector



RPC integrated charge limits

o RPC detector was certified for up to 0.3 $\ensuremath{\text{C}/\text{cm}^2}$ integrated charge

- This corresponds to about 10 years of LHC operations at $\mathcal{O}(100~\text{Hz/cm}^2)$, equivalent to 30 $\mu\text{A/m}^2$
- Some chambers at high $|\eta|$ will exceed this limit for High Luminosity LHC ightarrow reduce HV and replace some RPCs



 \rightarrow RPC ionisation currents extrapolated to the HL-LHC instantaneous luminosity

RPC detector time resolution

- o Measure time resolution using time differences of muon signals recorded by two parallel RPC layers
 - Two layers are separated by \sim 20 mm \rightarrow negligible muon time-of-flight
 - Subtract time resolution component of the front-end electronics which is measured in-situ
- o Average measured RPC time resolution: $\sigma_{RPC}/\sqrt{2} \sim 1$ ns
 - Small differences between η and ϕ time resolution is due to differences in construction



 $t_{\text{layer 0}} - t_{\text{layer 1}}$

 σ_{RPC}

Search for slow-moving stable charged particles

- σ Time-of-flight and dE/dx energy loss are used to search for heavy stable charged particles
 - RPC is the most sensitive detector for measuring muon time-of-flight
- o Search for production of supersymmetric particles (stau, chargino, gluino, R-hadron)
 - Sensitive to other models producing heavy stable charged particles





ATLAS L1 muon barrel trigger

Level 1 muon barrel trigger



- L1 muon barrel trigger uses RPCs to detect muon trigger candidates at 40 MHz rate
 - Custom-built on-detector electronics making decision within 2.1 μs
 - 3328 detector regions with $\Delta\eta imes\Delta\phipprox 0.1 imes 0.1$

o 3 low p_T thresholds:

 - 3/4 coincidence within trigger road in the two inner doublet layers (RPC1 and RPC2)

o 3 high p_T thresholds:

 Require highest low-p_T trigger plus 1-out-of-2 coincidence in the outer doublet layer (RPC3)

L1 muon barrel trigger: coincidence matrix

o Coincidence matrix ASIC (CMA)

- Application-specific integrated circuit (ASIC) to check coincidence of hits between two RPC layers within a cone
- 6 programmable roads (cone sizes) correspond to 6 trigger thresholds for muon p_T



Level 1 muon barrel trigger: efficiency

o MU20 is the primary L1 muon trigger threshold for selecting muons with $p_T > 20$ GeV for physics data taking

- Highly efficient for detecting muons produces in decays of W and Z bosons
- RPC acceptance holes and detector inefficiency lead to the efficiency plateau at 70% for MU20 trigger
- Steepness of the efficiency curve determines trigger rates \rightarrow dominated by muons with mismeasured p_T
- o Steepness of the efficiency curve determines trigger rates
 - Accepted MU20 events are dominated by low- p_T muons produced in $b\bar{b} + c\bar{c}$ events



L1 muon barrel trigger: (in)efficiency and rates

- o RPC acceptance holes and detector inefficiency lead to the efficiency plateau at 70% for MU20 trigger
 - Will install three new RPC layers in the inner barrel region for HL-LHC operations to increase acceptance
- o RPC muon trigger rates are dominated by low- p_T muons with mismeasured momentum
 - New Small Wheel detectors will reduce the endcap muon trigger rate by a factor of ~ 3
 - Barrel RPC muon trigger rates would then contribute a significant fraction of L1 events
 - Our study aims to improve p_T resolution of the future RPC trigger by using a neural network regression



Muon spectrometer upgrades for High Luminosity LHC

Muon spectrometer upgrades for High Luminosity LHC

o Current RPC:

- 6 layers with $\eta imes \phi$ grid of 3 cm wide strips
- Custom ASICs for muon trigger electronics
- Total L1 bandwidth is 100 kHz
- L1 latency to process an event: 2.1 μs

o After HL-LHC upgrades in 2025~2026:

- Higher background \rightarrow higher trigger rates
- 3 new inner RPC layers with better time resolution \rightarrow Thin-gap RPCs in inner barrel (BI)
- L1 \rightarrow L0: 1 MHz bandwidth & 10 μ s latency
- New FPGA-based electronics for L0 muon trigger
- MS Phase-2 Upgrade Technical Design Report
- TDAQ Phase-2 Upgrade Technical Design Report



FPGAs in future ATLAS trigger system

o Field-programmable gate array device (FPGA)

- Integrated circuit configurable after manufacturing
- Programmable logic blocks and interconnects
- Use software to programme computing hardware
- o L0 muon trigger:
 - Input: $\sim 0.1~\text{MB}$ at 40 MHz $\approx 4\text{TB/s}$
 - Fixed L0 muon latency \sim 4 $\mu \text{s} \rightarrow$ too fast for CPUs
 - Use FPGAs for hardware trigger algorithms
- High-level software-based trigger system (HLT) :
 - Input: \sim 2 MB at 1 MHz
 - Partial event reconstruction in regions of interest
 - R&D to use FPGAs to accelerate HLT algorithms



Neural network regression model for RPC muon trigger

Neural network regression model: goals

- 1. Our first goal is to measure muon q/p_T in order to improve $|p_T|$ resolution of the RPC trigger
 - Idea is to include muon charge $q \rightarrow$ narrower trigger road \rightarrow better p_T resolution and smaller background
 - Essentially, we use the neural network regression model to fit q/p_T

2. Design requirements

- Aim for fast enough network with small FPGA resource usage << resources of proposed XCVU13P FPGA
- Aim for neural network latency << 10 μ s latency of the future L0 trigger system
- If these goals can be achieved, neural networks can be also used for new exotic triggers long lived particles, etc

3. Advantages of using neural networks for hardware trigger

- Machine learning algorithms allow to reach higher signal efficiency and smaller background acceptance
- Same circuit can be used for different detector elements \rightarrow differences encoded via training weights
- Same circuit can be used for different triggers, for example to trigger on long lived particles
- o Collaboration with Prof. Changqing Feng, and Wenhao Dong, Wenhao Feng, Kai Zhang, Shining Yang
 - Preliminary results reported at CHEP 2021, today showing updates from our upcoming paper
 - Ours is different approach than Convolutional Neural Networks ightarrow presented at CHEP 2019 by Stefano Giagu

RPC toy simulation model



Study of FPGA-based neural network regression model

Rustem Ospanov

Candidate muon reconstruction

- 1. In each single layer, reconstruct nearby contiguous hits as one cluster
- 2. In each doublet layer, merge overlapping single-layer clusters into one super-cluster
- 3. In RPC2 doublet layer, draw a straight line through each RPC2 super-cluster (seed line)
 - 3.1 In RPC1 and RPC3 doublet layers, select super-cluster closest to this line
 - 3.2 If the selected super-clusters are within ± 20 strips to seed line, make a muon candidate
- o With a window of ± 20 strips to make candidates, muons with $p_T < 3$ GeV bend outside this window
- 2 candidates when a noise hit is reconstructed as a muon cluster



Study of FPGA-based neural network regression model

Candidate muon reconstruction: muon deflections

- o Deflections from the straight line are due to muon curvature in the magnetic field
 - Computed with respect to the straight line from the collision point (origin) to the RPC2 seed cluster
- o RPC3 deflections from the seed line are plotted below as a function of muon qp_T



 \rightarrow RPC3 muon deflections start to be comparable to strip width of 3 cm for $p_{T}\gtrsim 10~{\rm GeV}$

Neural network inputs

o 3 inputs for the neural network training:

- 1. RPC2 seed cluster z position (gives muon angular direction to NN)
- 2. RPC1 cluster Δz to seed line for $|\Delta z| < 0.15$ m
- 3. RPC3 cluster Δz to seed line for $|\Delta z| < 0.6$ m
- Using differences improved NN convergence and performance



Study of FPGA-based neural network regression model

Neural network regression model: design

o Scan several network architectures

- ightarrow Select 3 hidden layers with 20 nodes each & ReLU activation
- o Network size is driven by RPC resolution with 3 cm wide strips
 - \rightarrow Little benefit from a larger network size
- o Linear loss function to improve training convergence
 - $\rightarrow~$ Mean of |differences| between simulated and predicted q/p_T
- o Network training with PyTorch:
 - 100k events without noise to improve convergence & performance





Rustem Ospanov

Neural network performance



o Excellent performance for predicting q/p_T for pure muons

- Noise μ shown in orange
- Evaluated with statistically independent events
- Contributions from noise muons are small
- Also developed quality criteria to suppress noise muons

Neural network performance: trigger efficiency

- $_{\odot}$ Compute efficiency for selecting muon candidates with $p_{T}>20$ GeV:
 - Compare to MU20 trigger efficiency in data as shown earlier
 - Toy simulation has perfect acceptance ightarrow scale efficiency curve to match the data plateau
- o Obtained much steeper efficiency curve than data potentially leading to lower muon trigger rates
 - Missing many effects present in the real RPC detector \rightarrow still looks interesting enough to study further...





Rustem Ospanov

FPGA implementation and simulation

FPGA implementation

- o Implemented the full neural network regression model in Vivado HDL
 - Data processing logic not yet implemented important for final prototype
- o Serial data pipeline between layers
 - Reduce a number of connections between layers using distributor ightarrow smaller latency
 - 5 clock cycles for signal handshake, final adder & ReLU operations and transmission
- o Process in parallel 20 neurons of each layer
 - Processing element (PE) implements logic for one neuron node \rightarrow next page





Study of FPGA-based neural network regression model

Rustem Ospanov

FPGA implementation: neuron processing element

- o Neuron node is implemented in processing element (PE):
 - Output = ReLU($\sum_{i=1}^{20} x_i \cdot \text{weight}_i + \text{bias}$)
 - Process serially 20 data inputs from the previous layer
 - Latency = (N_{input} + 4) $imes \Delta t_{\mathrm{clock}}$ = 24 $imes \Delta t_{\mathrm{clock}}$
- Multiply-add-accumulate (MAC) unit:
 - Implemented using one digital signal processor (DSP)
 - 3 clock cycles for multiplication and 2 clock cycles for addition
 - Odd/even inputs are processed independently: $IA_i * W_i + AC_{i-2}$





Study of FPGA-based neural network regression model

FPGA implementation: latency and resource usage

o Latency for the full network: 98 clock cycles

245 ns @400 MHz << 10 $\mu \rm s$ latency of L0 trigger system

o Deadtime for the full network: 24 clock cycles

60 ns @400 MHz < 3 LHC bunches = 75 ns

o Resource usage for implementation on Xilinx FPGA XCKU060:

LUTs	Registers	DSPs
9949 (3.15%)	10257 (1.55%)	68 (2.36%)

- o This corresponds to $\sim 0.5\%$ of resources of XCVU13P FPGA
 - 32 such devices will be used for muon barrel trigger upgrade

o Much lower resource usage than <u>hls4ml</u> with \times 3 latency

- Latency can be further reduced by using more DSPs



FPGA implementation: fixed point arithmetic

- o Our FPGA implementation uses 16-bit binary fixed-point numbers
 - Scan several options for fractional part precision
 - Compute relative p_T error between full precision and fixed-point precision plotted below
 - Chosen 10 bits for the fractional part and 6 for the signed integer part



FPGA simulation

- o Full neural network circuit has been tested using simulation:
 - Simulation test project was developed using Questa Advanced Simulator and SystemVerilog
- o Compare results from PyTorch and FPGA simulation for the same events:
 - Percent level errors from using fixed point 16-bit arithmetic
 - Efficiency curve for the FPGA implementation is nearly identical that obtained with PyTorch



Potential applications for FPGA-based neural networks

o HL-LHC searches for long lived particles (LLPs)

- L1 trigger was optimised for detecting SM particles
- FPGAs allow development of dedicated exotic triggers
- Can neural networks be used to trigger on exotic signatures? LLP decays, slow-moving LLPs, highly ionising LLPs

o Hardware accelerators for ATLAS High Level Trigger (HLT)

- HLT runs on a large CPU farm that will process events accepted by L0 at 1 MHz rate
- Is it possible to use FPGAs or GPUs to accelerate CPU-intensive (track) reconstruction steps?
- Ongoing R&D to answer this question by 2025, plan to use commercial FPGA or GPU cards plugged in PCIe slots
- Main points: cost, power, cooling, flexibility, usability



Summary and outlook

- o Effective trigger selection of muon candidates is crucial for the ATLAS physics programme
- Excellent performance of the ATLAS RPC detector and L1 muon barrel trigger with 2018 data
- o Extensive muon spectrometer & trigger upgrades are planned for the HL-LHC
- o All new FPGA-based L0 muon trigger electronics will allow more sophisticated trigger algorithms
- We developed resource efficient FPGA-based neural network regression model
 - Regression model is trained with toy RPC simulation to measure muon q/p_T
 - Promises better performance than the current L1 system \rightarrow steeper muon efficiency curve
 - Implemented this neural network in FPGA code: 245 ns latency and very low resource usage
- o Results look promising and warrant further studies using more accurate simulation
 - Plan to develop dedicated triggers to search for new long-lived particles using the muon spectrometer

Thank you for your attention!

BACKUP

Trigger timing calibrations

- RPC hits (muon signals) are calibrated online with 3.125 ns step
 - More than sufficient to identify individual LHC bunch crossings with 25 ns spacing
- 99.7% of muon candidates arrive within expected 25 ns time window
- Excellent stability of timing calibrations during data taking period



RPC trigger efficiency is reduced by $\approx 20\%$ by detector support structures



Study of FPGA-based neural network regression model

RPC trigger efficiency is reduced by another pprox 10% by inefficient modules (left plots)





Offline muon

Study of FPGA-based neural network regression model

HL-LHC studies

- ▶ RPC upper limit on current density is 30μ A/m² for HL-LHC at $\mathcal{L} = 7.5 \times 10^{34}$ cm⁻²s⁻¹
- Extrapolate current LHC data to high luminosity to study expected performance
 - Chambers with smaller radius and at high $|\eta|$ will exceed these limits
 - Plan to reduce HV to 9.2 kV and decrease front end thresholds to regain \sim 10% efficiency
- Scan FE discriminator thresholds at 9.6 kV (nominal) and 9.2 kV (proposed for HL-LHC)



RPC detector currents at different |n| stations

RPC detector efficiency

versus discriminator V_{FF}

RPC toy simulation: hit multiplicity



- On average, about 9 total hits per event
- On average, about 2 noise hits and 1.8 cluster hits

Rustem Ospanov

Neural network simulation: cluster multiplicity



- Super-clusters are reconstructed from single-layer clusters in one doublet layer that are within $1.5 \times \text{strip}$ width
- On average, about 8 single-layer clusters
- On average, about 5 super (double-layer) clusters
- Check: same number of hits for both cluster types

Study of FPGA-based neural network regression model

Rustem Ospanov

Neural network simulation: super-cluster position differences

0.100 0.6 0.075 Z_{cluster} [m] Ξ 04 0.050 Zrluster 0.025 0.2 0.000 0.0 -0.025 Zline -0.2 ° < ∠_{sim. µ} < 70° --0.050 < ∠_{sim. µ} < 70° m **RPC** Ä $70^{\circ} < L_{sim}, \mu < 50^{\circ}$ -0.4 $70^{\circ} < \angle_{sim}$, $\mu < 50^{\circ}$ -0.075 $^{\circ} < L_{sim}$, $\mu < 40^{\circ}$ 50 $< L_{sim} \mu < 40$ -0.100-0.6-30 -20 -1030 -30 -20 20 30 Pure muon $a \times p_{T}$ Pure muon $a \times p_{T}$ 0.15 0.6 z_{cluster} [m] ε 0.10 0 ctor 0.05 0.7 0.00 0.0 Zlin Zli -0.05 -0.2 RCI RPC3 -0.10 -0.4 -0.15 -0.6 -30 30 -30 20 30 Muon with noise $q \times p_T$ Muon with noise $a \times p_{T}$ Rustem Ospanov 47

- o Muon candidates:
 - Require \geq 1 clusters per RPC1, RPC2 and RPC3
 - Make muon candidate for each RPC2 (seed) cluster
 - Draw line through each RPC2 seed cluster
- o Clear correlations between p_T and Δz for muons without noise hits
- As expected, random deviations for muons containing noise hits

Study of FPGA-based neural network regression model

Candidate muons

o On average, we reconstruct 1 muon candidate per simulated event

- 0 candidates when one doublet layer is inefficient
- 2 candidates when a noise hit is reconstructed as a cluster
- In later plots, noise muon candidate (noise μ) contains at least one noise cluster



Number of muon candidates per event

Training events with noise

Training events without noise





FPGA implementation: multiply-add-accumulate (MAC) unit

- o MAC is implemented using one DSP with 3 clock cycles for multiplication and 2 clock cycles for addition
- o Odd and even input data elements are processed in parallel and independently: $IA_i * W_i + AC_{i-2}$



Simulated dose and particle flux

Hadrons flux

Neutron equivalent flux



HL-LHC upgrades of RPC detector and trigger

- o For HL-LHC data-taking, RPC will provide up to 9 measurements of $\eta \times \phi$
- o Inner barrel RPCs will increase detector acceptance
- MDT will be included in hardware muon trigger \rightarrow refine p_T measurement for candidates accepted by RPC
- o Order of magnitude better time-of-flight resolution with new on-detector electronics and faster thin-gap RPCs

