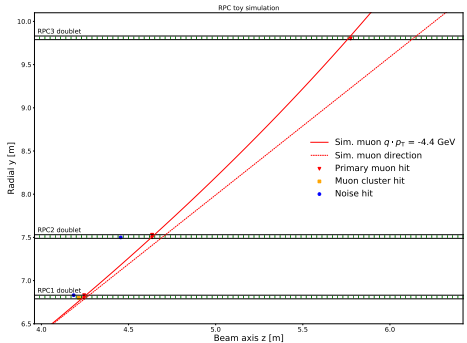
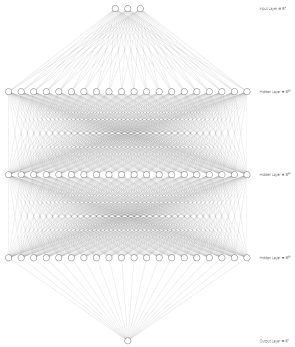
Development of resource-efficient FPGA-based neural network regression model for the ATLAS muon trigger upgrades





Rustem Ospanov, Changqing Feng, Wenhao Dong, Wenhao Feng, Kan Zhang, Shining Yang

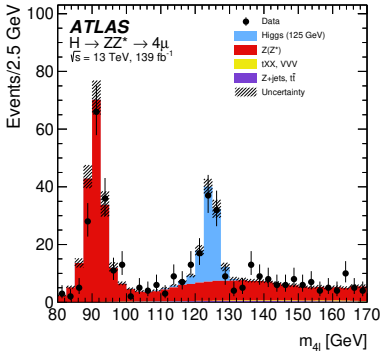
International Conference on High Energy Physics 2022 - Bologna, Italy

July 8th, 2022

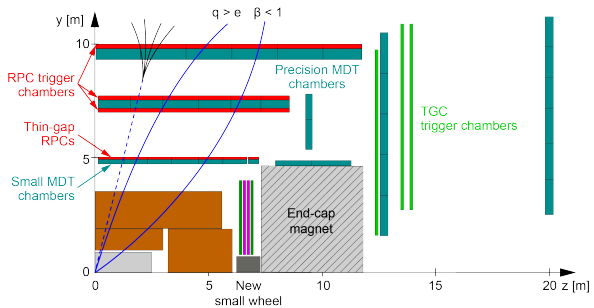
Motivation

- Muons are important signature for the physics programme at the Large Hadron Collider (LHC)
  - Electroweak studies with W & Z bosons, Higgs boson measurements, searches for new phenomena...
- Detector & trigger upgrades for High Luminosity LHC create new unique opportunities for improved trigger performance
  - *Use neural network regression to measure more precisely muon transverse momentum (p_T)*
  - *FPGA-based hardware machine learning algorithms can be used for new exotic trigger signatures*

$H \rightarrow 4\mu$

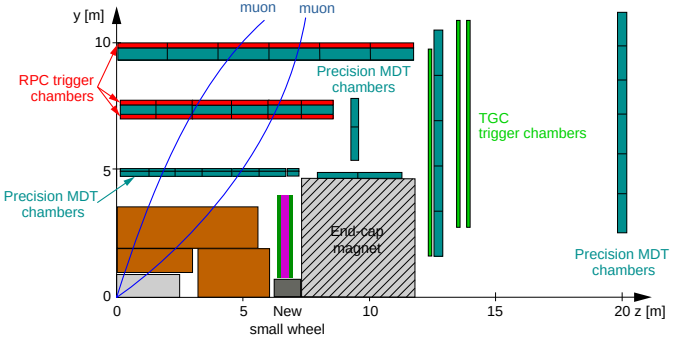


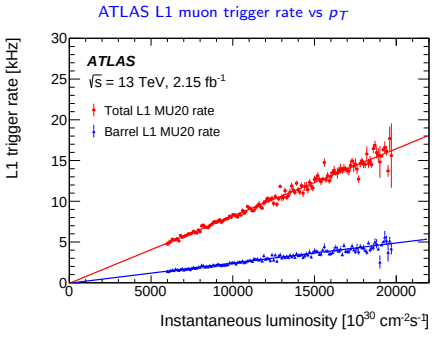
Potential new signatures for long-lived particle (LLP) searches:
LLP decays, highly ionising LLPs, slow-moving LLPs



ATLAS muon spectrometer (MS)

- 2 fast detectors for L1 trigger with muon position resolution of $\sim$ 1 cm:
  - Fast measurements of muon p_T within the 2.1 μs latency of the L1 trigger
  - Muons add $\sim$ 10% of 100 kHz of the Level-1 ATLAS trigger rate $\rightarrow$ dominated by muons with mismeasured p_T
  - Resistive plate chambers (RPCs) in the barrel region ($|\eta| < 1.05$) - *subject of this talk*
- 2 precision detectors for high-level trigger (HLT) and offline muon reconstruction:
  - 1 barrel and 2 endcap air-core toroid magnets with muon p_T resolution $\Delta p_T / p_T \approx 4\%$

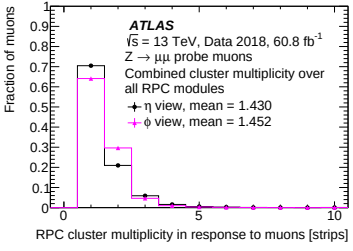




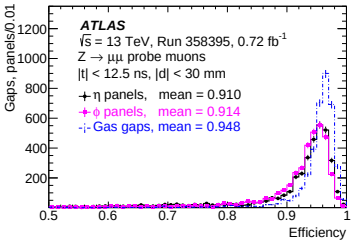
ATLAS RPC detector

- 3 concentric cylindrical shells of double-layer (doublet) chambers located at radii of 7, 8 and 10 meters
- $\sim$ 3700 gas volumes with the surface area of $\sim$ 4000 m^2 with $\sim$ 360k readout strips
- Provide 6 measurements in bending (r, z) plane and 6 measurements in non-bending (x, y) plane
- ATLAS RPC performance paper using 2018 data: JINST 16 (2021) P07029

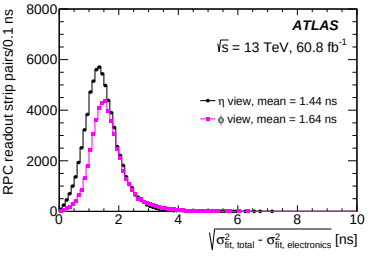
RPC cluster hit multiplicity in response to muons



Efficiency to detect muon passage



RPC time resolution $\sigma_{RPC}^{\text{gauss.}}$ [ns]



Neural network regression model for RPC muon trigger

Neural network regression model

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

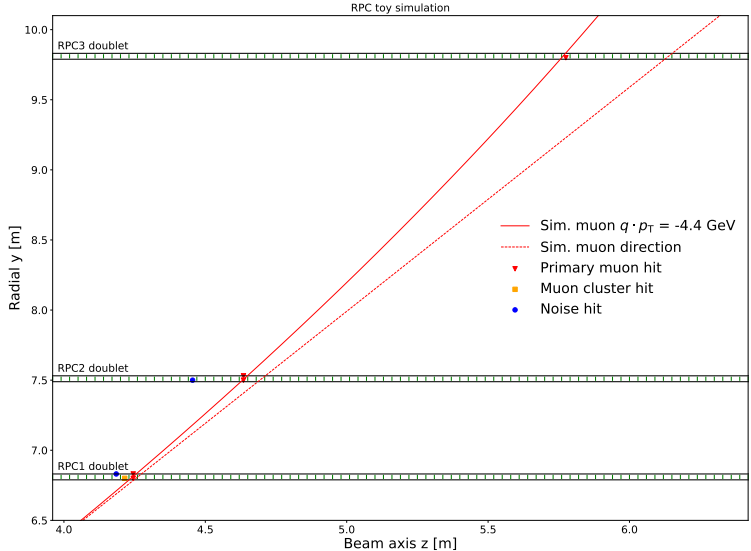
2. Design requirements
   - Aim for fast enough network with small FPGA resource usage $<<$ resources of proposed XCVU13P FPGA
   - Aim for neural network latency $<<$ 10 μs latency of the future L0 trigger system
   - *If these goals can be achieved, neural networks can be also used for new exotic triggers - 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

- Collaboration with Prof. Changqing Feng, and Wenhao Dong, Wenhao Feng, Kai Zhang, Shining Yang
  - *Presenting today our just published results:* EPJC 82, 576 (2022)

RPC toy simulation model

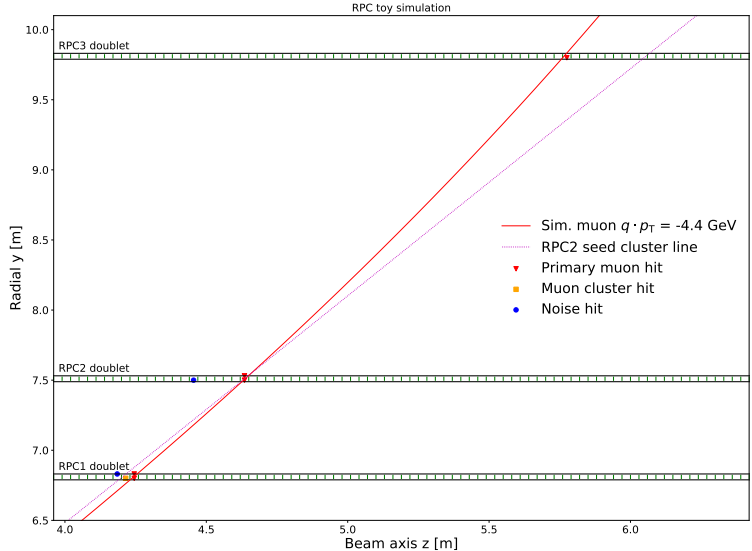
- RPC detector toy model for first studies:
  - Model existing RPC to allow comparisons with current ATLAS RPC performance
  - Uniform 0.5 T toroidal B-field
  - 6 RPC layers with perfect acceptance
  - Parallel 3 cm wide strips
  - Only bending (η) detector view
  - 95% efficiency to produce muon hit
  - 25% prob. to produce 2-strip muon cluster
  - 0.1% prob. per strip to make noise hit
  - No detector material (multiple scattering is small because of the air-core toroids)
- Muon simulation parameters:
  - Flat muon p_T: 3 to 30 GeV
  - Flat muon angle: 50 to 85 degrees to z-axis
  - Python code in GitHub: https://github.com/rustemos/MuonTriggerPhase2RPC



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*

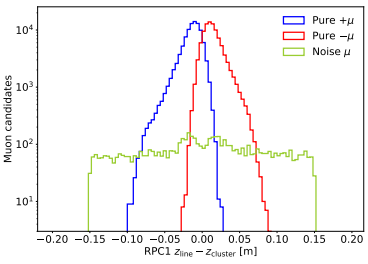
- With a window of ± 20 strips to make candidates, muons with $p_T < 3$ GeV bend outside this window
- 2 candidates when a noise hit is reconstructed as a muon cluster



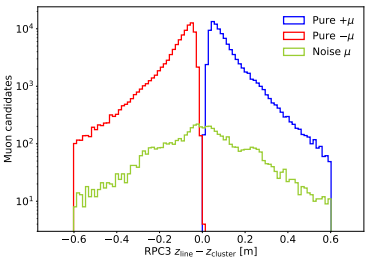
Neural network inputs

- 3 inputs for the neural network (NN) training:
  1. RPC2 seed cluster z position (provides muon angular direction information to NN)
  2. RPC1 cluster Δz with respect to seed line - require $|\Delta z| < 0.15$ m
  3. RPC3 cluster Δz with respect to seed line - require $|\Delta z| < 0.6$ m
  - *Using differences improved NN training convergence and performance*

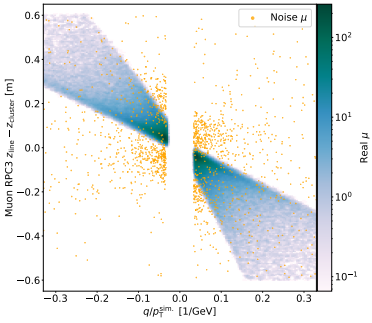
RPC1 differences between $z_{cluster}$ and seed line



RPC3 differences between $z_{cluster}$ and seed line

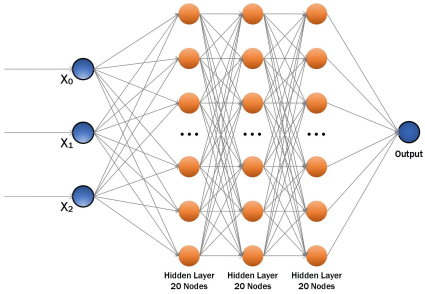


RPC3 deflections from the seed line vs. muon qp_T

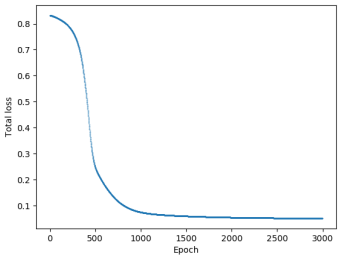


Neural network regression model: design

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

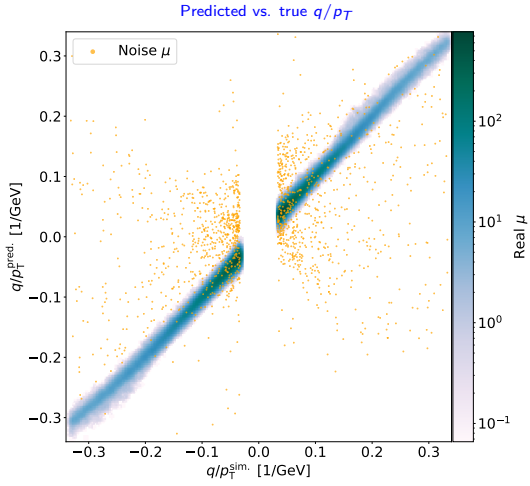


Loss function vs. training epoch



Neural network performance

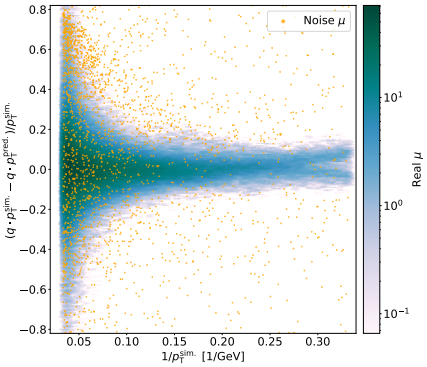
- Excellent performance for predicting q/p_T for pure muons
  - Noise μ shown in orange
  - Evaluated with statistically independent events
  - Contributions from noise muons are small
  - Also developed quality criteria to suppress noise muons



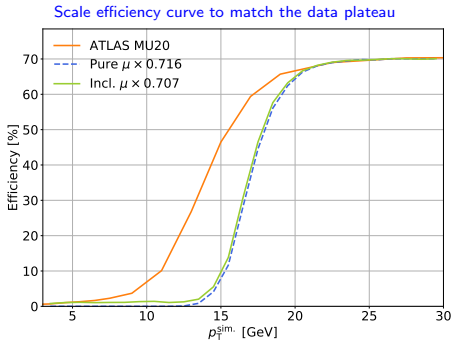
Neural network performance: trigger efficiency

- 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 → scale efficiency curve to match the data plateau
- Obtained much steeper efficiency curve than data - potentially leading to lower muon trigger rates
  - Missing many effects present in the real RPC detector → still looks interesting enough to study further...

Relative p_T error vs. true q/p_T
Achieved about 10% p_T resolution at 10 GeV



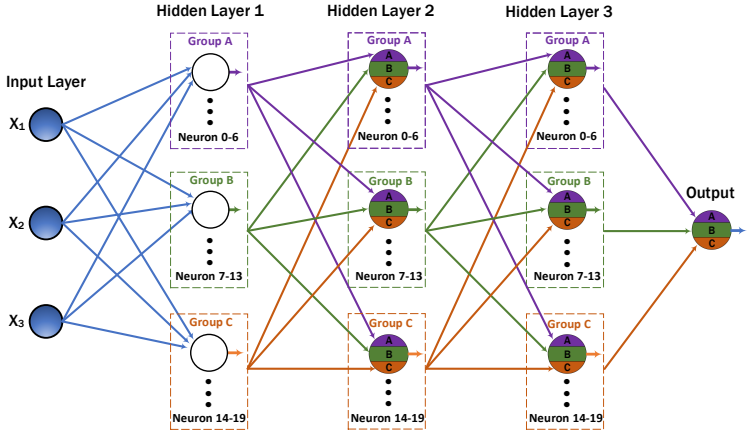
Efficiency to trigger muon with $p_T > 20$ GeV
Scale efficiency curve to match the data plateau



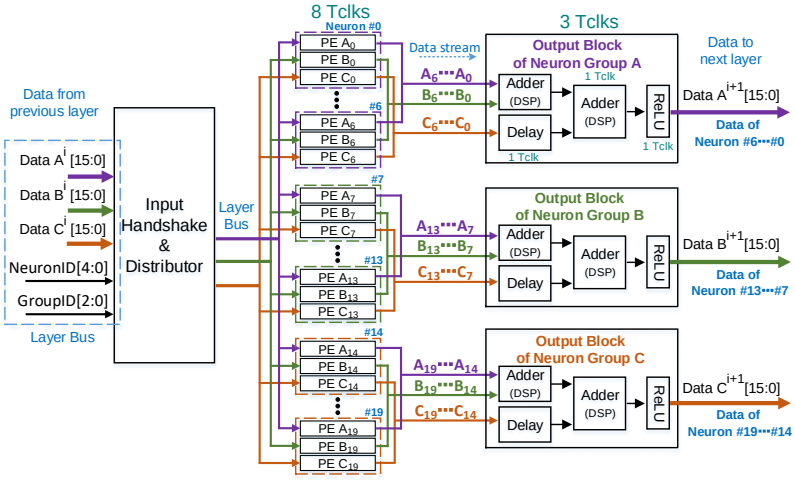
FPGA implementation and simulation

FPGA implementation

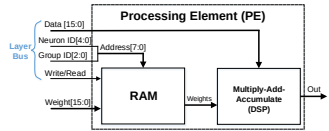
- Implemented full neural network regression model in Vivado hardware description language (HDL)
  - 3 serial data pipelines between layers sending data simultaneously to 20 neurons of next layer
  - Neuron node is implemented using 3 processing units (PEs), each receiving & processing data in parallel



FPGA implementation: hidden layer design



Neuron node processing element (PE)



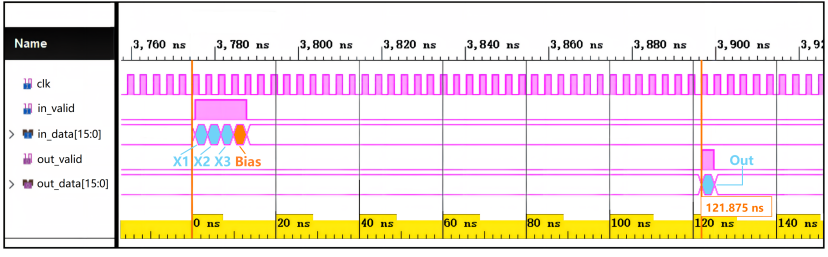
FPGA resource utilisation

	Deadtime (cycles)	Latency (cycles)	DSPs
Layer 1	5	7	20
Layer 2	8	11	66
Layer 3	8	11	66
Output layer	8	10	5

Neuron node output = $\mathrm{ReLU}(\sum_{i=1}^{20} x_i \cdot \mathrm{weight}_i + \mathrm{bias})$

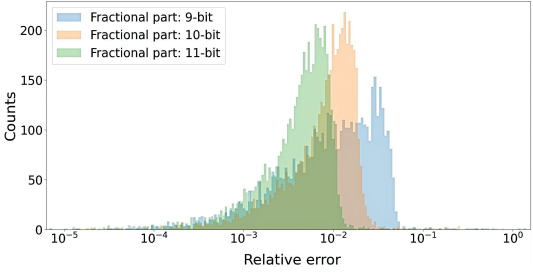
- Resource usage and simulated latency @320 MHz for Xilinx FPGA XCKU060 for 3 FPGA implementations
  - Optimised HDL design available at hdl4nn and simplified (baseline) HDL design using one PE per neuron
  - High Level Synthesis (HSL) design with hls4ml - "reuse" sets number of multiplications by each DSP
  - *A factor of 4 smaller resource usage for our HDL design with similar latency as HLS implementation*
  - *122 ns latency and 25 ns deadtime for our optimised HDL design using 157 DSPs and about 5,000 LUTs*

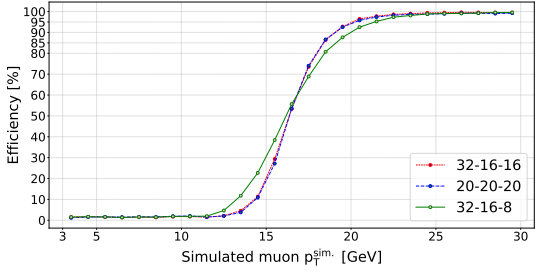
	Deadtime (cycles)	Latency (cycles)	DSPs	LUTs	Registers
Optimised HDL design	8	39	157 (5.7%)	4659 (1.4%)	7370 (1.1%)
HLS (reuse = 1)	2	23	677 (24.5%)	30952 (9.2%)	15507 (2.3%)
HLS (reuse = 2)	6	94	430 (15.6%)	38701 (11.5%)	8916 (1.4%)
HLS (reuse = 4)	8	100	225 (8.2%)	35972 (10.7%)	6301 (1.0%)
Simplified HDL design	24	98	65 (2.4%)	9949 (3.0%)	10257 (1.6%)



Choosing fixed point arithmetic precision and neural network configuration

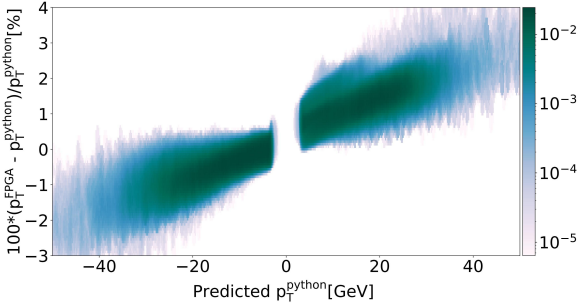
- HDL FPGA implementation uses 16-bit binary fixed-point numbers
  - Scan several options for fractional part precision
  - Compute relative p_T error between full precision and fixed-point precision - plotted below
  - *Chosen 10 bits for the fractional part and 6 for the signed integer part*
- Also scanned several neural network configurations
  - Selected 20-20-20 configuration as providing best muon momentum resolution

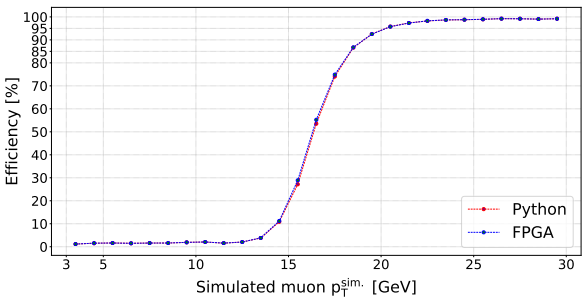




FPGA simulation

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





Summary and outlook

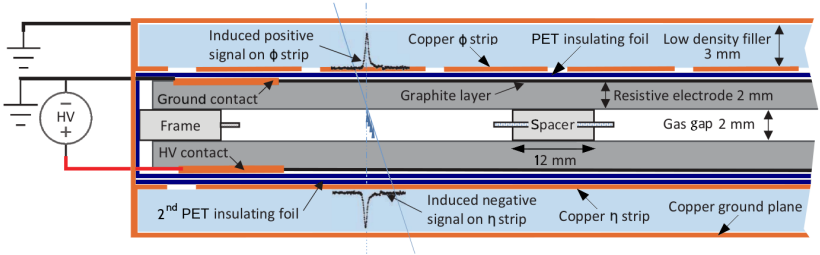
- ✓ Effective trigger selection of muon candidates is crucial for the LHC physics programme
- ✓ Extensive muon spectrometer & trigger upgrades are planned for the HL-LHC
  - New FPGA-based muon trigger electronics will allow more sophisticated ATLAS trigger algorithms
- ✓ *We developed resource-efficient FPGA-based neural network regression model - EPJC 82, 576 (2022)*
  - Neural network is trained to measure muon q/p_T using data from the toy RPC simulation
  - This model promises better performance than the current L1 system $\rightarrow$ steeper muon efficiency curve
  - Important caveat: toy simulation does not include many effects present in the real RPC detector
- ✓ This neural network was implemented in HDL code (hdl4nn) with 122 ns latency and 25 ns deadtime
  - *A factor of 4 smaller resource usage and similar latency compared to the implementation obtained with hls4ml*
- ✓ Results look promising and warrant further studies with ATLAS data and simulation
  - Also aim to develop dedicated L0 triggers to search for long-lived particles using the ATLAS muon spectrometer

Thank you for your attention!

BACKUP

ATLAS Resistive Plate Chambers

- Parallel resistive plates (bakelite with $2 \times 10^{10}\ \Omega \cdot cm$) are separated by 2 mm with insulating spacers
- Induced signal is read out using orthogonal η and ϕ copper strips with 23-35 mm pitch
- $\sim$ 1 ns total time resolution $\rightarrow$ excellent separation of proton bunches that are 25 ns apart
- 320 MHz clock for detecting raising edge of the amplified avalanche signal $\rightarrow$ 3.125 ns wide time bins
- RPC operate in avalanche mode with average applied voltage of 9.6 kV $\rightarrow$ working at the efficiency plateau

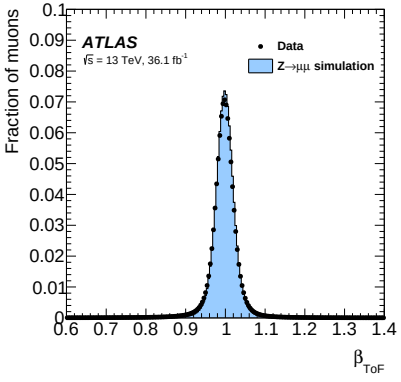


- Non-flammable low-cost gas: tetrafluorethane $C_2H_2F_4$(94.7%), iso-butane C_4H_{10}(5%), sulphur hexafluoride SF_6(0.3%)
- This mixture is a potent greenhouse gas $\rightarrow$ currently being phased out in EU due to environmental impact $\rightarrow$ raising costs

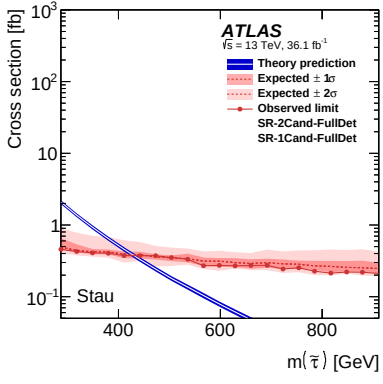
Search for slow-moving stable charged particles

- 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
- Search for production of supersymmetric particles (stau, chargino, gluino, R-hadron)
  - Sensitive to other models producing heavy stable charged particles

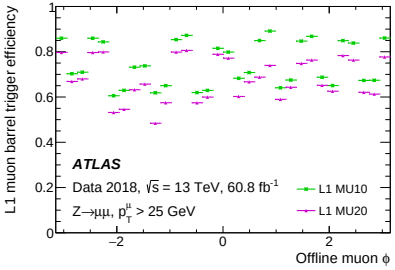
RPC has most precise β_μ resolution of $\sim 2\%$

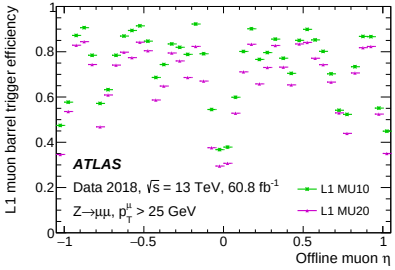


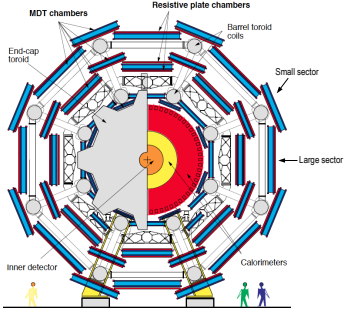
Exclusion limit on stau production

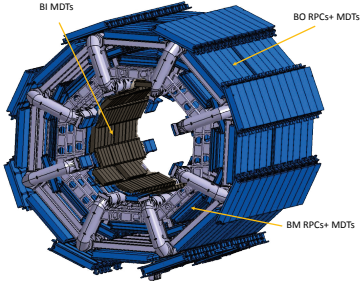


RPC trigger efficiency is reduced by ≈ 20% by detector support structures

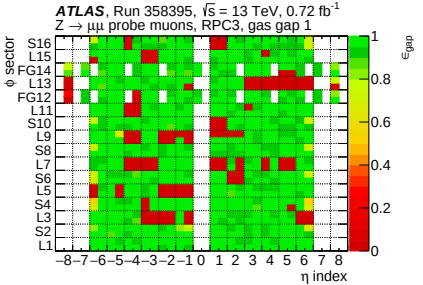


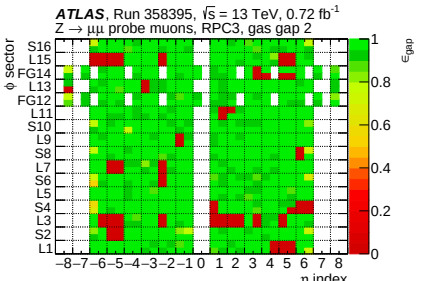


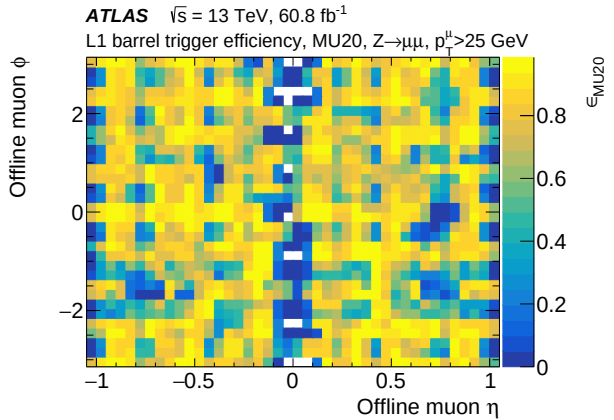




RPC trigger efficiency is reduced by another ≈ 10% by inefficient modules (left plots)



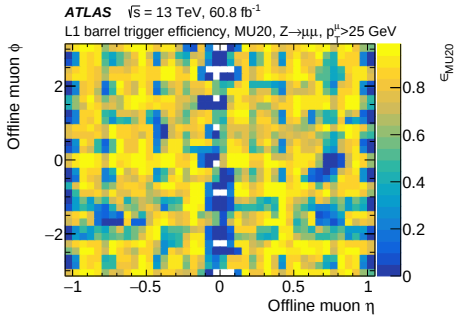




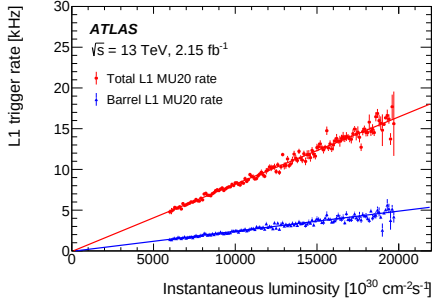
L1 muon barrel trigger: (in)efficiency and rates

- 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
- 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*

MU20 trigger efficiency in (ϕ, z) plane



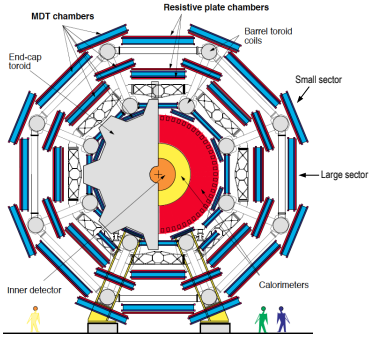
Muon trigger rate vs p_T



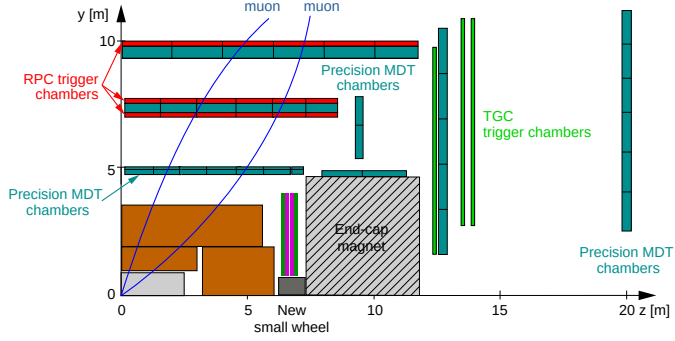
ATLAS RPC detector

- 3 concentric cylindrical shells of double-layer (doublet) chambers located at radii of 7, 8 and 10 meters
- $\sim$ 3700 gas volumes with the surface area of $\sim$ 4000 m^2 with $\sim$ 360k readout strips
- Provide 6 measurements in bending (r, z) plane and 6 measurements in non-bending (x, y) plane
- ATLAS RPC performance paper using 2018 data: JINST 16 (2021) P07029

Non-bending (x, y) plane - RPC ϕ strips



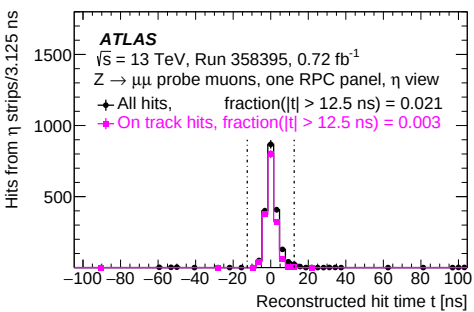
Bending (r, z) plane - RPC η strips



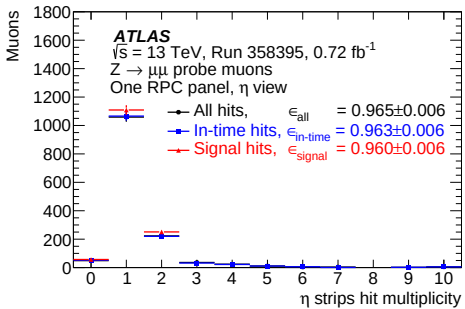
RPC detector response

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

Calibrated hit time for one RPC module
Zero corresponds to time of pp collisions

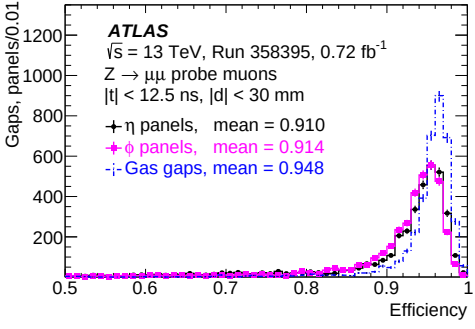


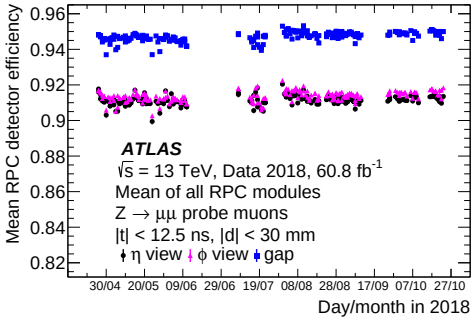
Hit multiplicity in response to muon passage for one RPC module
Efficiency is a fraction of events with at least one detected hit



RPC detector efficiency

- Muon detection efficiency = probability to detect muon induced avalanche producing ≥ 1 hit
  - Measured using events containing a muon predicted to pass through a given chamber
  - Gas gap efficiency = probability to detect avalanche using either η or ϕ strips
- 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

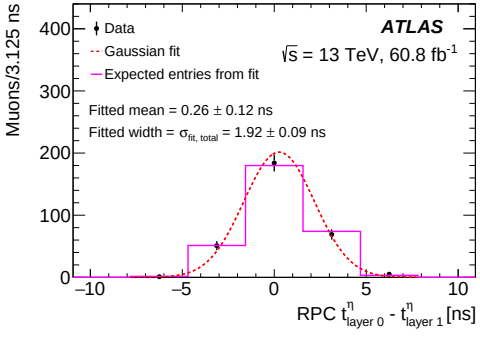




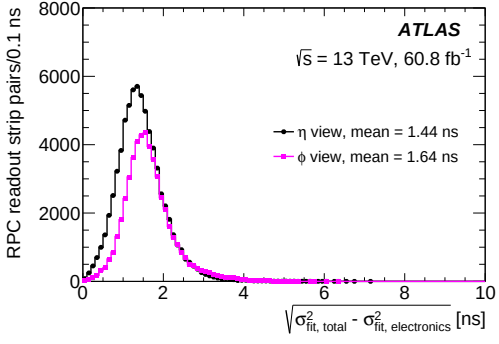
RPC detector time resolution

- 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
- 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

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

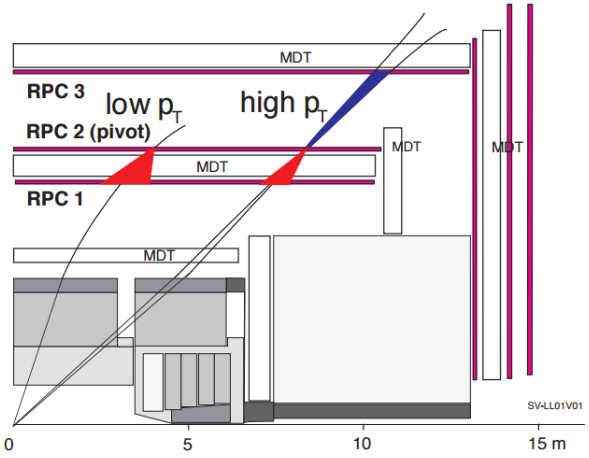


σ_{RPC}



ATLAS L1 muon barrel trigger

Level 1 muon barrel trigger

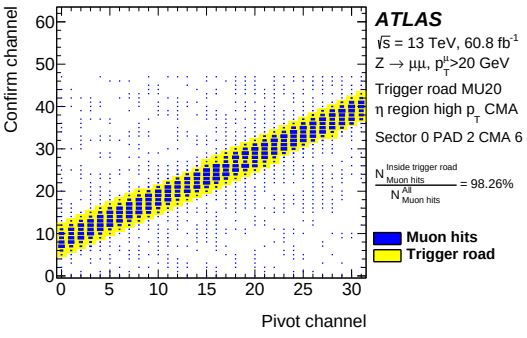


- L1 muon barrel trigger uses RPCs to detect muon trigger candidates at 40 MHz rate
  - Custom-built on-detector electronics making decision within 2.1 μs after each beam crossing
  - 3328 detector regions with $\Delta\eta \times \Delta\phi \approx 0.1 \times 0.1$
- 3 low p_T thresholds:
  - 3/4 coincidence within trigger road in the two inner doublet layers (RPC1 and RPC2)
- 3 high p_T thresholds:
  - Require highest low-p_T trigger plus 1-out-of-2 coincidence in the outer doublet layer (RPC3)

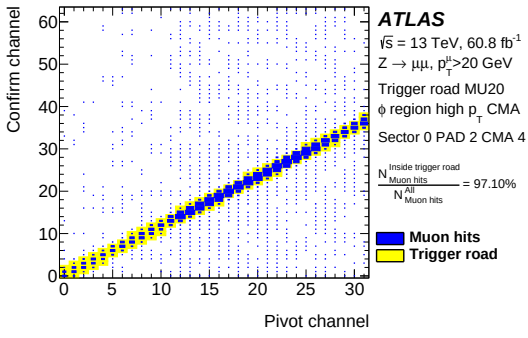
L1 muon barrel trigger: coincidence matrix

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

High-p_T MU20 road for **bending** η coordinate
Road width is determined by muon curvature in magnetic field



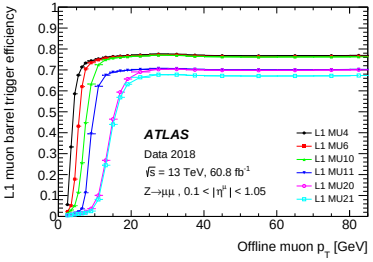
High-p_T MU20 road for **non-bending** ϕ coordinate
Road width is determined by RPC strip width



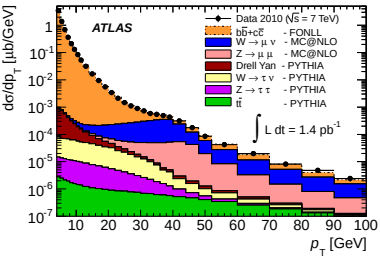
Level 1 muon barrel trigger: efficiency

- 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
- 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

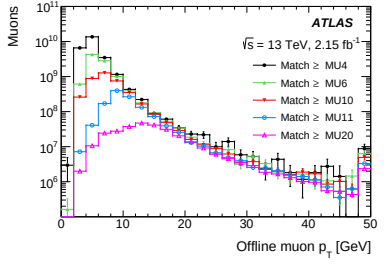
L1 barrel muon trigger efficiency $\epsilon_{\text{trigger}}$ vs. p_T



Inclusive muon p_T spectrum $P_\mu(p_T)$ at 7 TeV



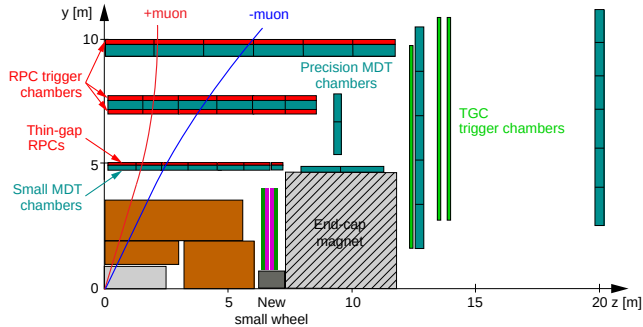
p_T of accepted MU20 events = $\int \epsilon_{\text{trigger}} \cdot P_\mu(p_T)$



Muon spectrometer upgrades for High Luminosity LHC

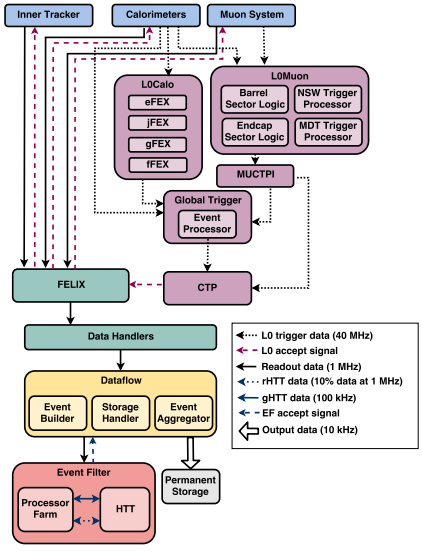
Muon spectrometer upgrades for High Luminosity LHC

- Current RPC:
  - 6 layers with $\eta \times \phi$ 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
- After HL-LHC upgrades in 2025∼2026:
  - Higher background → higher trigger rates
  - 3 new inner RPC layers with better time resolution → Thin-gap RPCs in inner barrel (BI)
  - L1→L0: 1 MHz bandwidth & 10 μs 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

- Field-programmable gate array device (FPGA)
  - Integrated circuit configurable after manufacturing
  - Programmable logic blocks and interconnects:
    lookup tables (LUTs), digital signal processors (DSPs), random-access memory (RAM), etc
  - *Use software to programme computing hardware*
- L0 muon trigger:
  - Input: $\sim$ 0.1 MB at 40 MHz $\approx$ 4 TB/s
  - Fixed L0 muon latency $\sim$ 4 μs $\rightarrow$ 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*



Potential applications for FPGA-based neural networks

- HL-LHC searches for long lived particles (LLPs)
  - L1 trigger was designed for detecting SM particles
  - FPGAs allow development of new dedicated exotic triggers
  - *Neural networks can be used to trigger on exotic signatures: LLP decays, highly ionising LLPs, slow-moving LLPs*