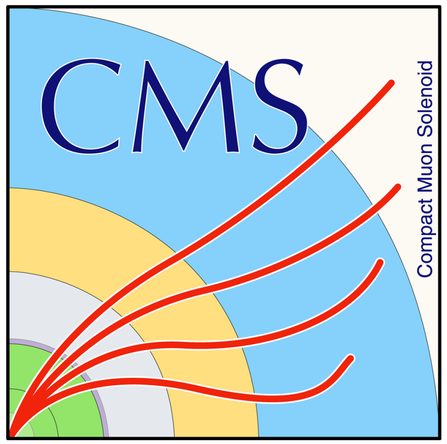
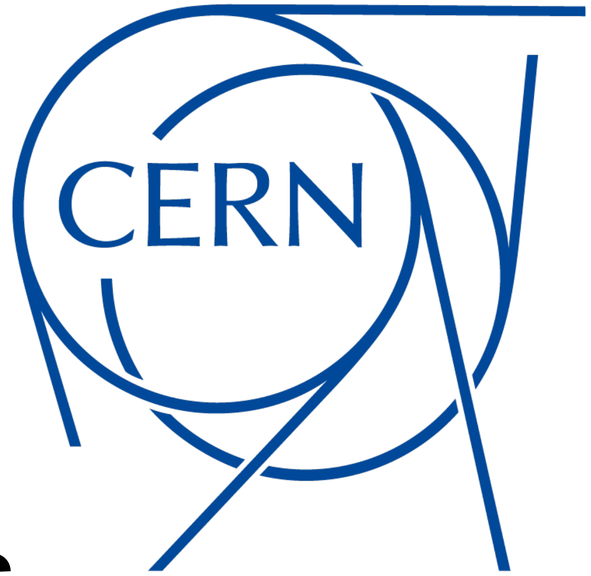


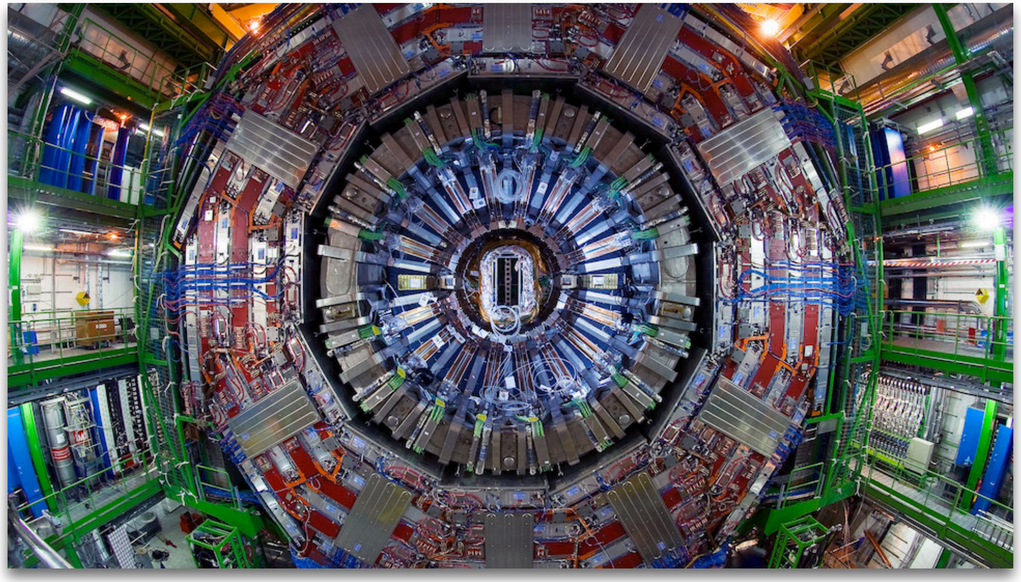

# 40 MHz Scouting with Deep Learning in CMS

Dejan Golubović on behalf of the CMS Collaboration

Connecting the Dots, 6th International Workshop, April 22-24, 2020

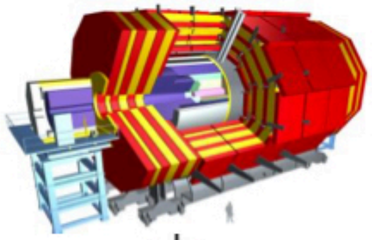
# Summary

- L1 Trigger scouting in CMS
  - Motivation, history, schedule
- Use of deep learning L1 scouting
  - Goals, datasets, methodology
- Results
  - Deep learning evaluation, hardware implementation

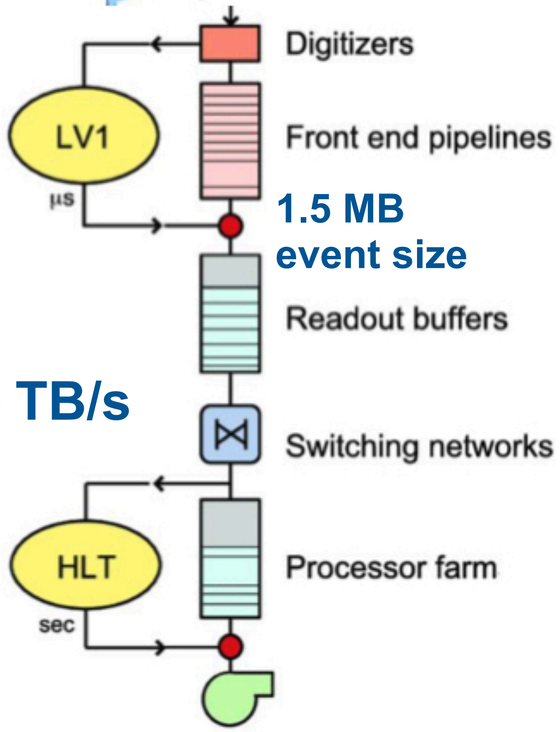


# CMS Trigger and Data Acquisition

## Phase 0&1: 2008 - 2024



Peak pile-up 60




40 MHz

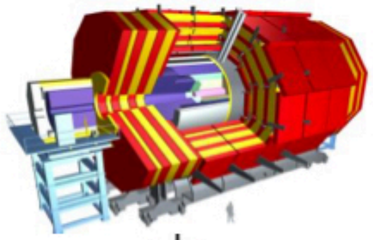
100 KHz

0.15 TB/s

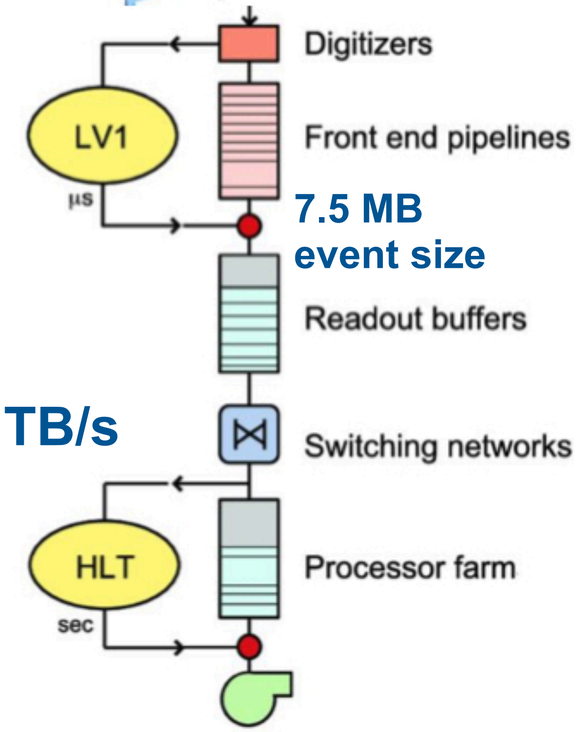
2 KHz

Phase-2
Upgrade
2027 - onwards

## Phase 2: 2027 - 2036



++ Peak pile-up 200




40 MHz

750 KHz

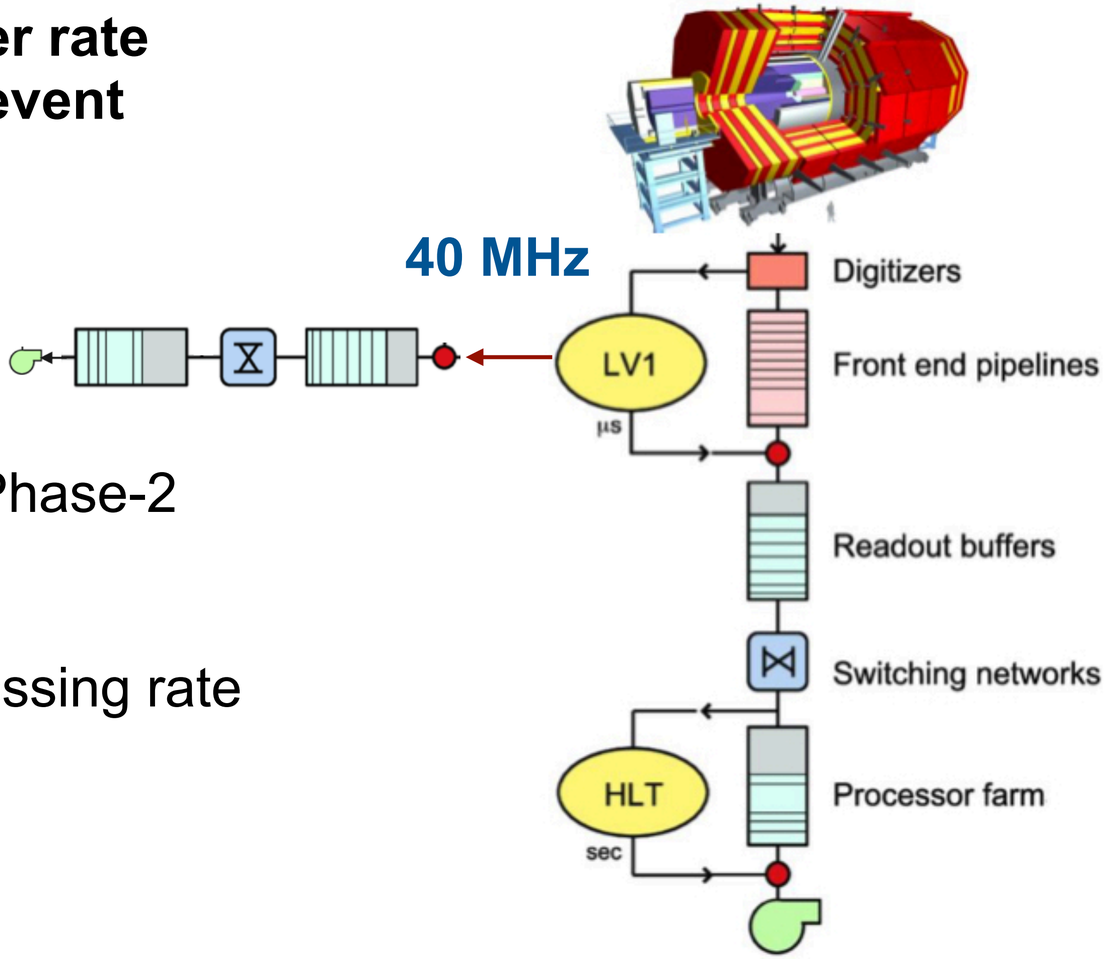
5.5 TB/s

7.5 KHz

# L1 Trigger Scouting Concept and Schedule

- **Analyse partial events at much higher rate than possible when reading out full event**

- Ongoing demonstrator development
- Further expanded system planned for Phase-2 (2027 onwards)
- Acquire L1 trigger data at full bunch crossing rate
- Analyse certain topologies at full rate
  - Semi real-time analysis
  - Storing of tiny event record



# L1 Trigger Scouting Motivation

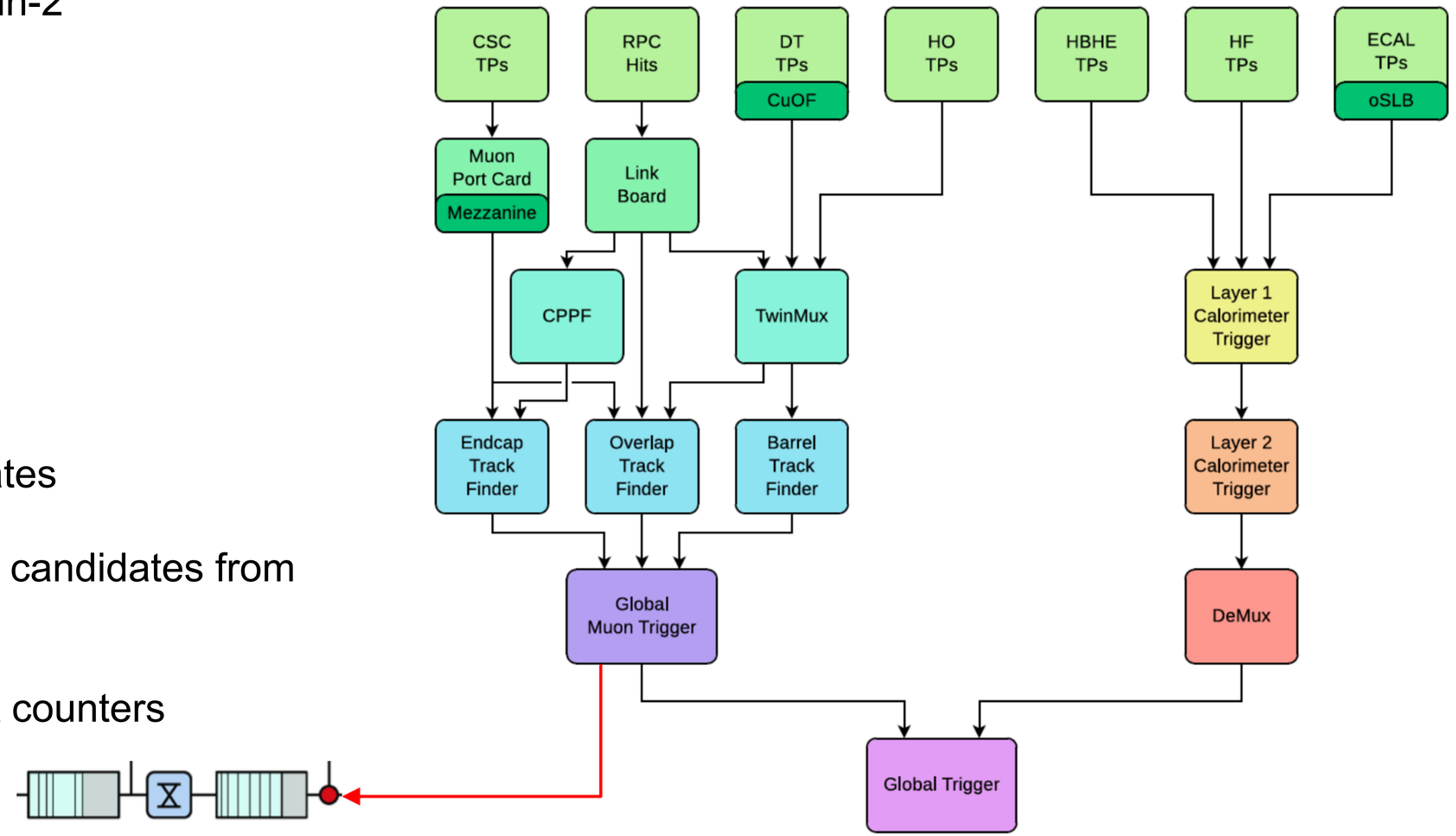
- Physics Case
  - Higgs rare decays
  - Displaced muons
  - Flavour anomalies
  - Bs $\rightarrow \tau\tau$
  - Hadronic physics
  - QCD measurements

- Technical Case
  - Diagnostics of the trigger system at large
  - Detect anomalies in the detector
  - Try out novel Global Trigger algorithms (test in minutes)
  - Cross-check existing Global Trigger algorithms on a BX-by-BX basis
  - Detect and analyse pre-/post-firing and to select cosmic ray muons to be used to test L1 tracking efficiency
  - Provide real time luminosity measurement independent from BRIL and other systems

- The Phase-2 Upgrade of the CMS Level-1 Trigger
- CERN-LHCC-2020-004 ; CMS-TDR-021
- https://cds.cern.ch/record/2714892

# L1 Trigger Scouting Demonstration in Run-2

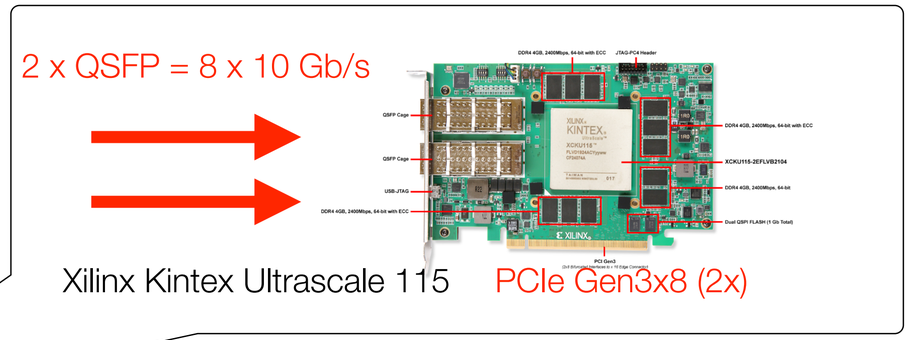
- October - November 2018, Run-2
- Types of runs:
  - 1 week of pp run
  - Large part of HI run
- Capture at 40 MHz
  - Up to 8 final muon candidates
  - Up to 8 intermediate muon candidates from barrel region
  - GMT adds bunch and orbit counters






40 MHz Scouting Prototype System

# L1 Trigger Scouting Architecture




8 x optical
10 Gb/s
From GMT

8 GB/s

KCU 1500

Firmware ZS
1/20

DMA

PCIe
Gen3

Soft ZS
1/8

Max 800 MB/s

Ram
disk
mount

Max 100 MB/s

10 GE
NIC

10/40
GE
switch

40 GE
NIC

RAM
Disk

BZIP
1/2

RAID 8
TB

Max 50 MB/s

Infiniband
NIC

Lustre

Dell R720

Dell R720

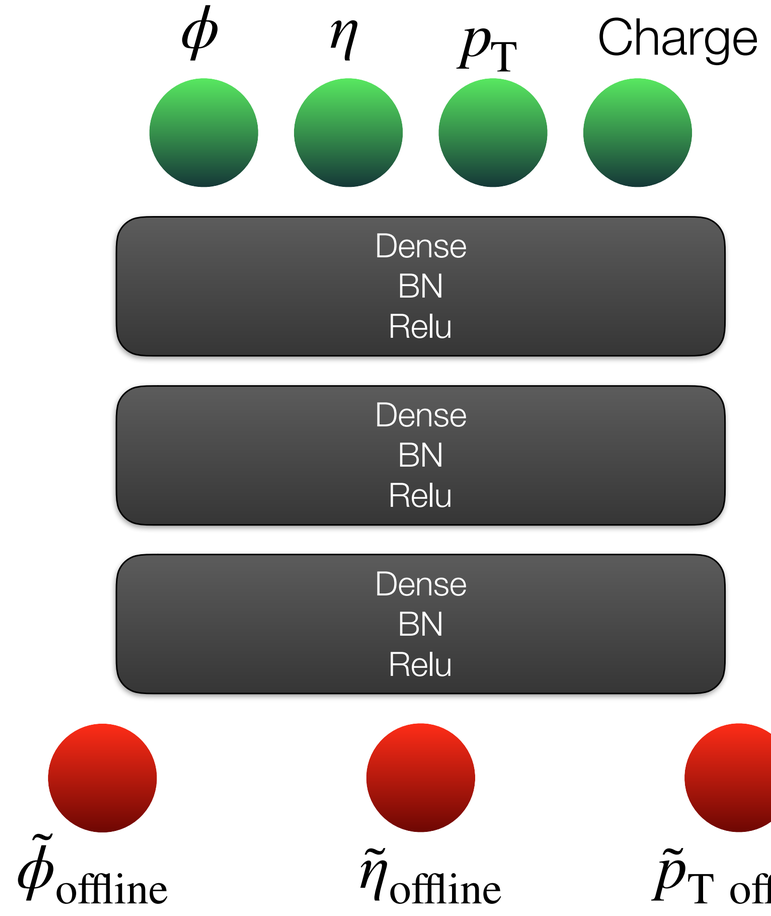
Controller PC
Firmware Update and Monitoring

# Deep Learning in L1 Trigger Scouting in Run-3

- Inputs to the scouting system are **L1 Trigger muon objects**
- Goal - have accurate measurements of muon parameters as soon as possible
- No access to the silicon tracker and the pixel detector data, used to perform the **offline reconstruction**
- **Correct L1 Trigger muon objects** to be as close as possible to the offline reconstructed values
- Train a DL model with the
  - **L1 Trigger muon objects as inputs**
  - **Offline reconstructed data as targets**
- Models trained with ZeroBias datasets, obtained during Run-2
- Models trained on past data will be used to **inference** the corrections to the muon parameters in **real-time**

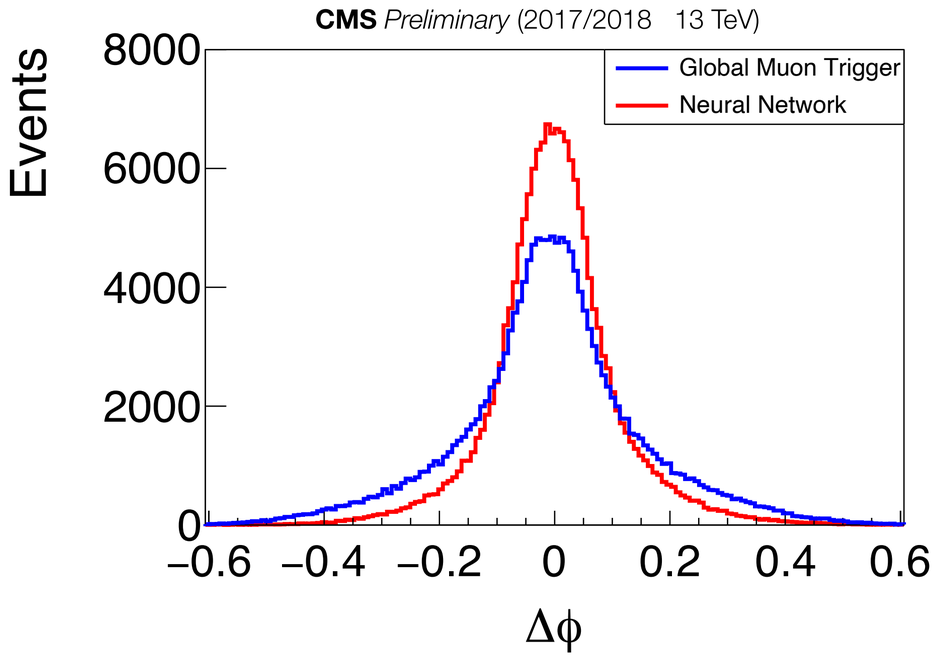
# Deep Learning Models

- Multilayered perceptron to re-fit muon parameters ($\phi$, $\eta$, $p_{\mathrm{T}}$)
- Model inputs - L1 Trigger muon objects
- Model targets - offline reconstructed data
- Improve accuracy of the Global Muon Trigger data
- Development tools
  - Python programming language
  - Keras and Tensorflow as deep learning frameworks
  - Scikit-learn, Matplotlib, Searborn, PyRoot, Pandas, H5py…



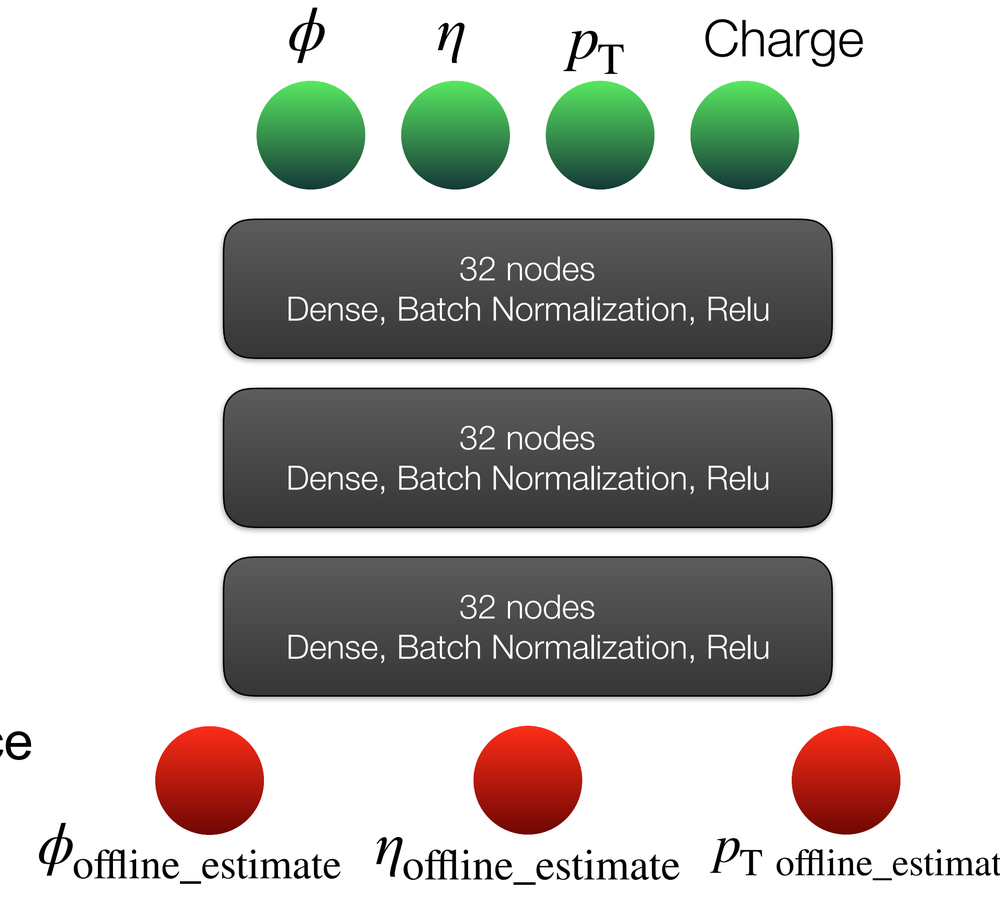
# Deep Learning Metrics

- Goal - minimize the difference between offline reconstructed values and the model predictions
- Baseline - output of the Global Muon Trigger
- Metrics
  - Distribution RMS
  - Mean close to zero
  - Percentage of data in the histogram tails



# Model Optimization

- Hyperparameters are parameters of a model that can't be trained
  - Number of layers, number of nodes
  - Activation functions
  - Loss function
  - Optimizers
  - Regularization
- Important to assess the affect of various combinations of hyperparameters in order to achieve the optimal performance
- 5-fold cross validation



# Model Optimization Results

| Model Id | Loss function | Optimizer | Regularization | Activation in hidden layers | Activation in output layers | RMS | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | $\Delta\phi$ | $\Delta\eta$ | $\Delta p_{\mathrm{T}}/p_{\mathrm{T}}$ |
| 1 | Logcosh | Adadelta | None | Relu | Linear | **0.119** | **0.031** | 0.167 |
| 2 | Msle | Adam | None | Relu | Linear | 0.145 | 0.032 | 0.193 |
| 3 | Logcosh | Adadelta | None | Softmax | Relu | 0.138 | 0.032 | 0.231 |
| 4 | Logcosh | Adadelta | L2 $\lambda=10^{-7}$ | Relu | Linear | **0.119** | **0.031** | **0.165** |
| 5 | Logcosh | Adadelta | None | Relu | Linear | 0.120 | **0.031** | 0.169 |

$$\Delta\phi = \phi_{\mathrm{predicted}} - \phi_{\mathrm{offline\_reconstructed}}$$
$$\Delta\eta = \eta_{\mathrm{predicted}} - \eta_{\mathrm{offline\_reconstructed}}$$
$$\Delta p_{\mathrm{T}} = p_{\mathrm{T\_predicted}} - p_{\mathrm{T\_offline\_reconstructed}}$$

| | | | |
|---|---|---|---|
| Global Muon Trigger | 0.166 | 0.034 | 0.274 |

# Deep Learning Results, Training Datasets

- **ZeroBias** datasets from Run-2 in 2017 and 2018 used for training end evaluating the deep learning models

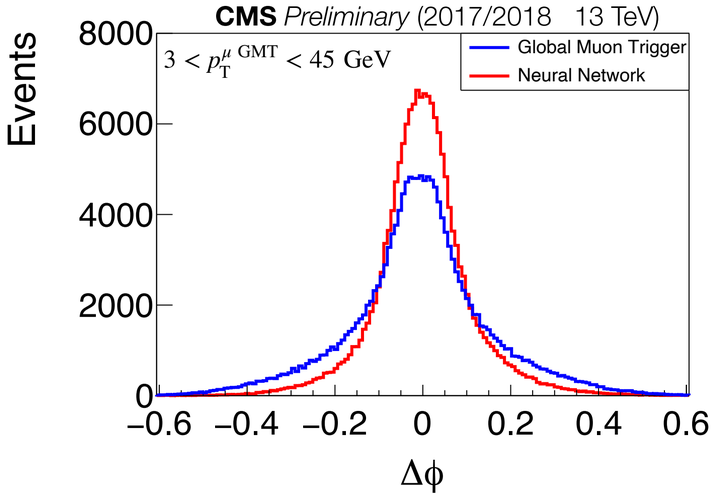
- Selection

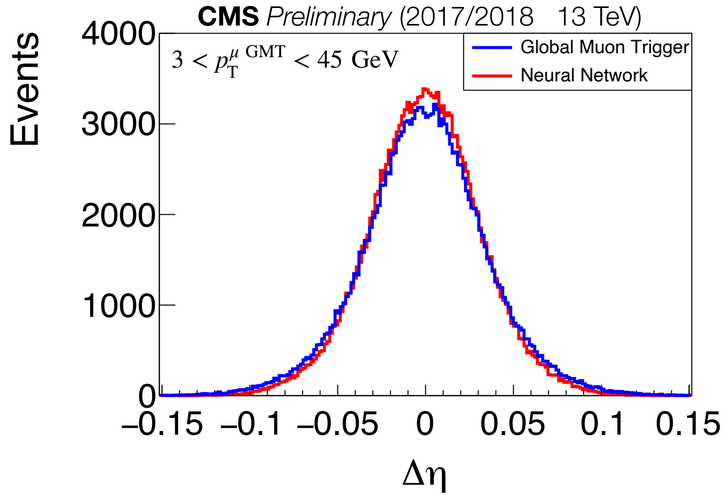
  - L1-Reco matching $\Delta R<0.1$ at 2nd muon station L1 Barrel Muon Track Finder (BMTF) only.

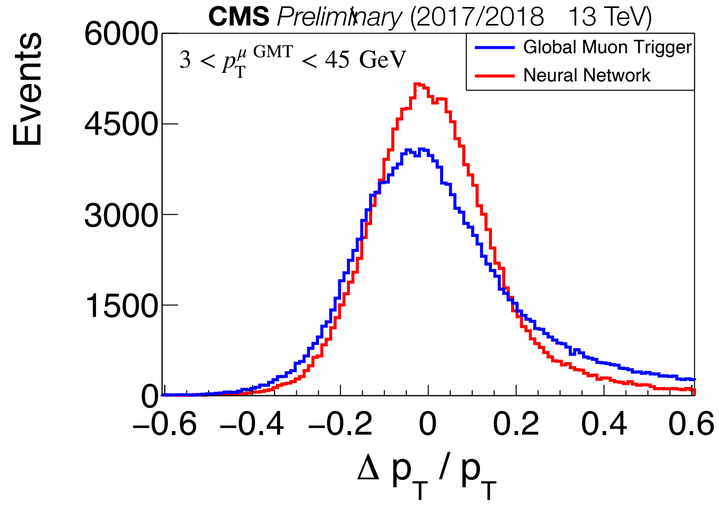
  - **High quality muons** - stubs at each of four stations

  - **$2.5 <$ muon $p_T < 45$ GeV**

# Deep Learning Results, Neural Network vs GMT




Neural Network distribution narrower compared to the GMT

# Deep Learning Results, MuOnia Dataset

- **MuOnia** datasets from Run-2 in 2018 was used for evaluating the deep learning approach in the correction muon pairs

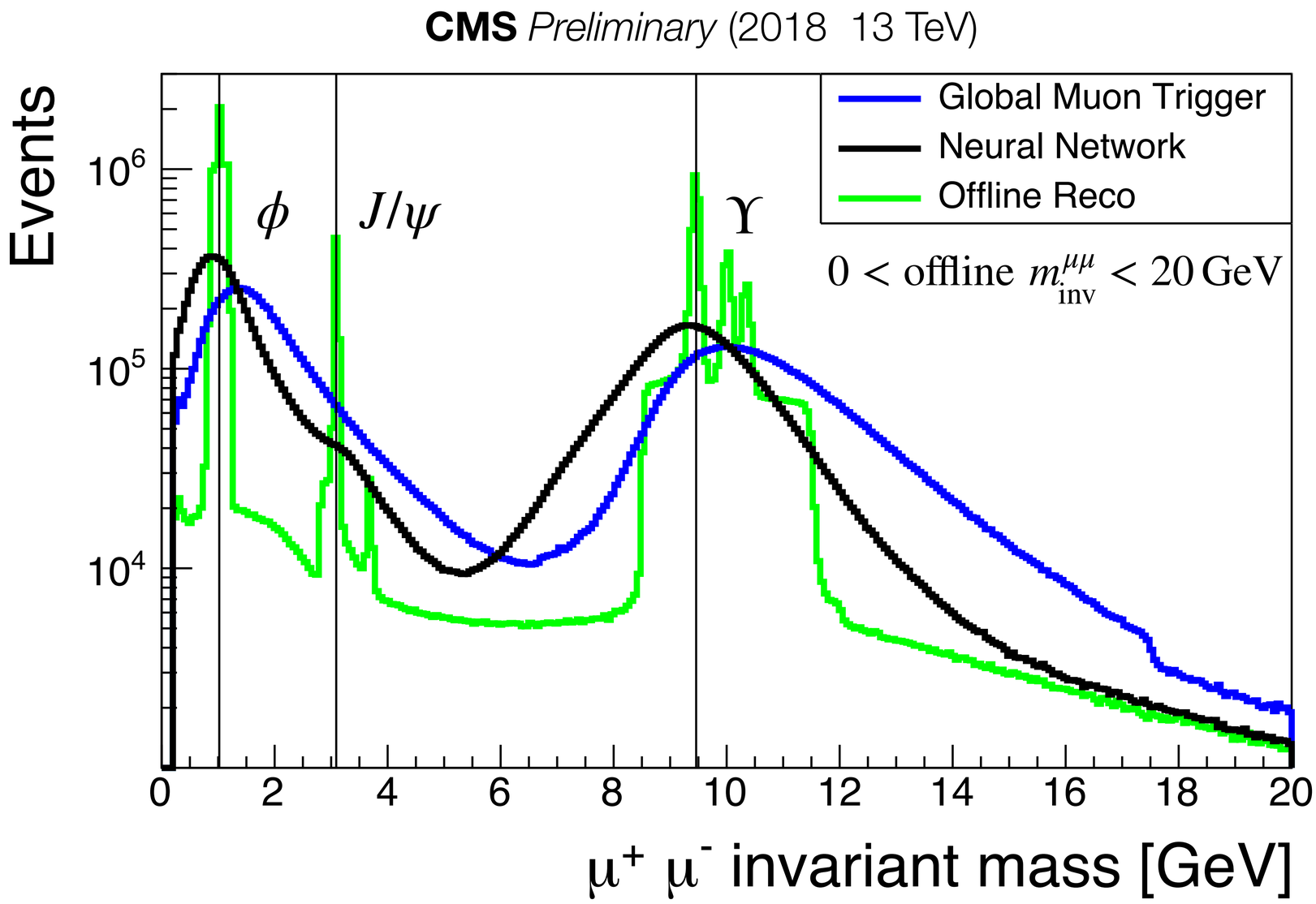
- Selection

  - L1-Reco matching ΔR<0.1 at 2nd muon station L1 Barrel Muon Track Finder (BMTF) only.

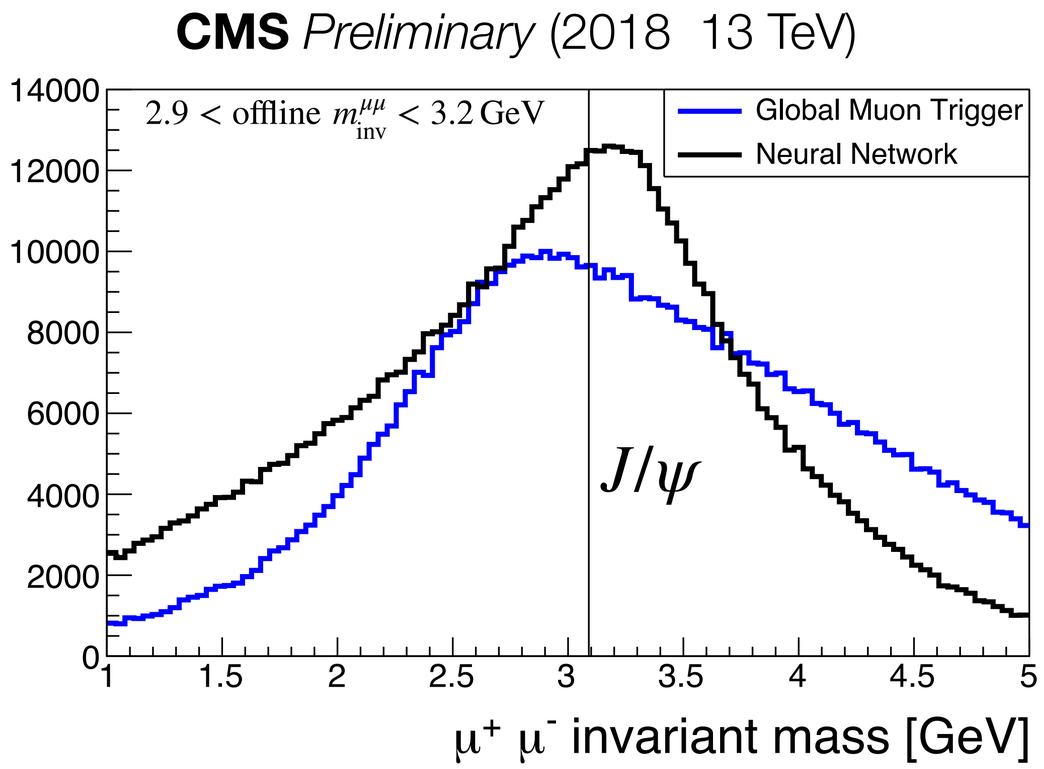
  - **High quality muons** - stubs at each of four stations

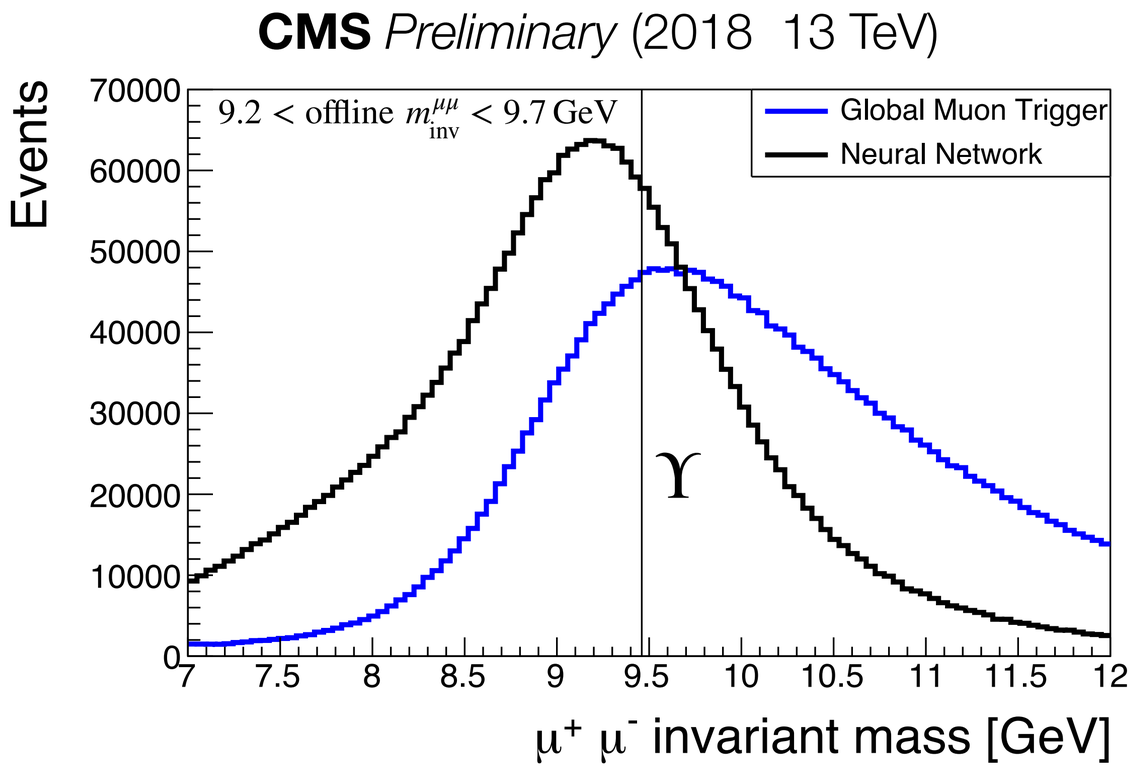
  - **2.5 < muon $p_T$ < 45 GeV**

# Deep Learning Results, MuOnia Dataset



# Deep Learning Results, MuOnia Dataset



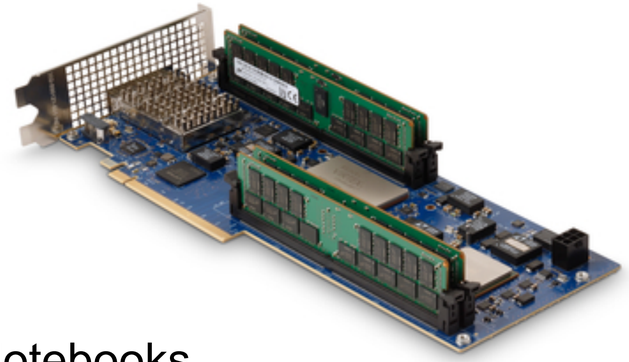

- Selection on the offline reconstructed invariant mass
- The ranges represent **close surroundings** of the the particle resonances, **J/ψ meson and ϒ meson**
- The plots show distributions of the invariant mass produced by the Global Muon Trigger and Neural Network

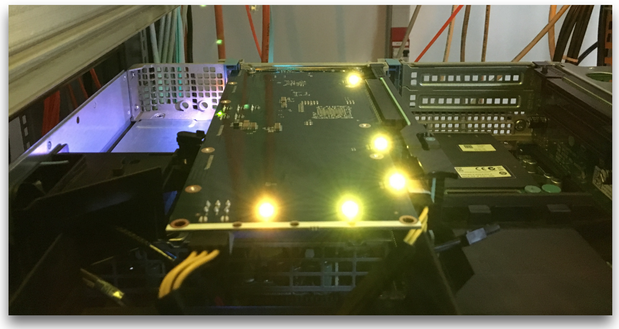
# Deep Learning Hardware

- Deep learning models are trained on PCs with GPUs
- To maximise throughput and minimise latency for inference, it is advantageous to implement deep learning models in FPGAs for L1 scouting
- One way - write VHDL code
- Simpler way - use deep learning compilers (offered by various vendors)

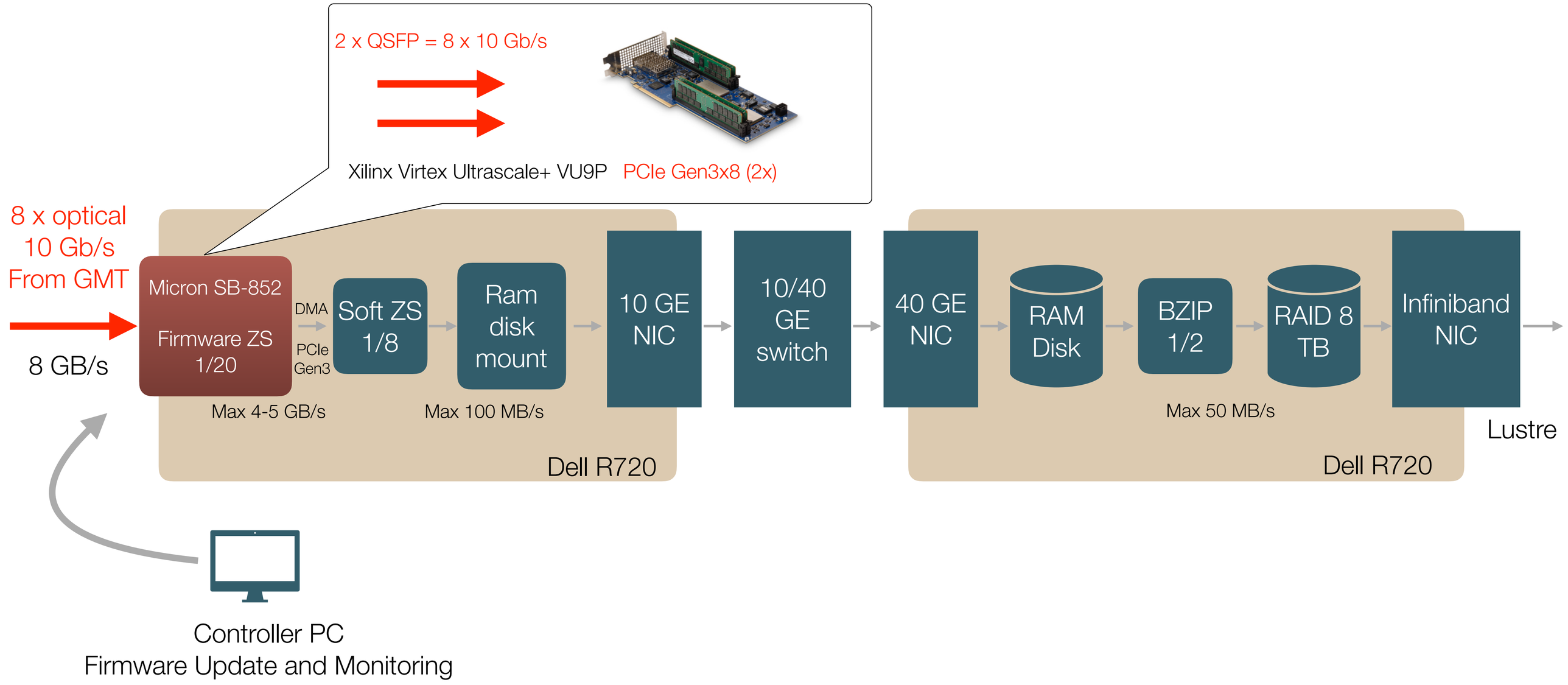
# Micron Deep Learning Accelerator

- Hardware FPGA-based boards
  - Expected high bandwidth, low latency, high performance memory
- Software
  - Board configured from user's Python code
  - No need for coding neural networks in VHDL
  - Execute deep learning models on the board from Python scripts and Jupyter Notebooks during development phase
- Advantages
  - Good long term solution, easy to modify and deploy models
  - Reduced engineering workload, allows more time for scientific research





# L1 Trigger Scouting Architecture with Micron DLA



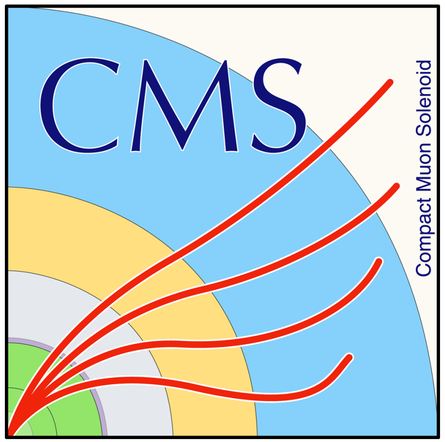
# Summary

- Achieved
  - Trained neural networks using the ZeroBias datasets from 2017 and 2018
  - Demonstrated improvement on resolution on Global Muon Trigger muons
  - Demonstrated improvement on resolution of the invariant mass
  - Compiled and ran developed models for inference using the Micron Deep Learning Accelerator
- Next steps
  - Integrate Micron hardware within the scouting infrastructure, for the demonstrator
  - Run scouting demonstrator in Run-3 (2021)



Many thanks to:

- CMS DAQ group for great work on scouting implementation
- Micron Technology, for the financial support and innovative work on machine learning accelerators
- CERN Openlab for organising the collaboration
- CTD Committee for the chance to present at the conference



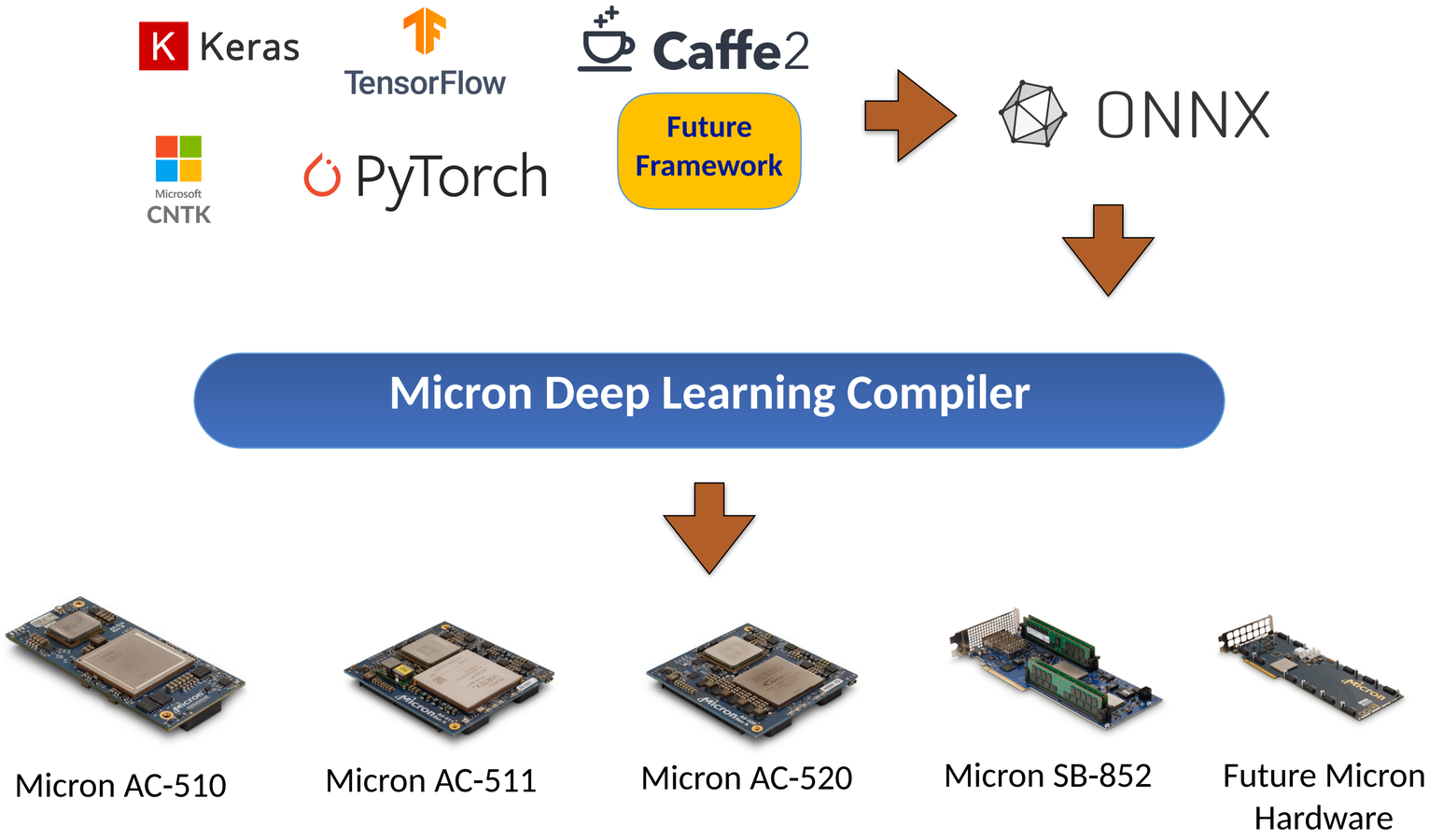

Glad to answer your questions!

# References


- H. Sakulin, “40 MHz Level-1 Trigger Scouting for the Compact Muon Solenoid Experiment”, Presented at 24th International Conference on Computing in High-Energy and Nuclear Physics, Adelaide, Australia, 4-8 November 2019
- Duarte, J.M., “Fast Reconstruction and Data Scouting”, Presented at Connecting the Dots 2018 4th International Workshop, Seattle, USA, 20-22 March 2018
- The CMS Collaboration, “The Phase-2 Upgrade of the CMS Level-1 Trigger”, CERN/LHCC (to be published).
- The CMS Collaboration, “Scales for inputs to μGT (φ, η, pt/Et), and others”, http://globaltrigger.hephy.at/files/upgrade/ugt/scales_inputs_2_ugt_2017Aug14.pdf
- Micron Technology, Inc., www.micron.com/products/advanced-solutions/advanced-computing-solutions/hpc-single-board-accelerators/sb-852
- A. X. M. Chang, A. Zaidy, M. Vitez, L. Burzawa, Eugenio Culurciello, “Deep neural networks compiler for a trace-based accelerator”, Journal of Systems Architecture 102, 101659, (2020), 372 doi:10.1016/j.sysarc.2019.101659.

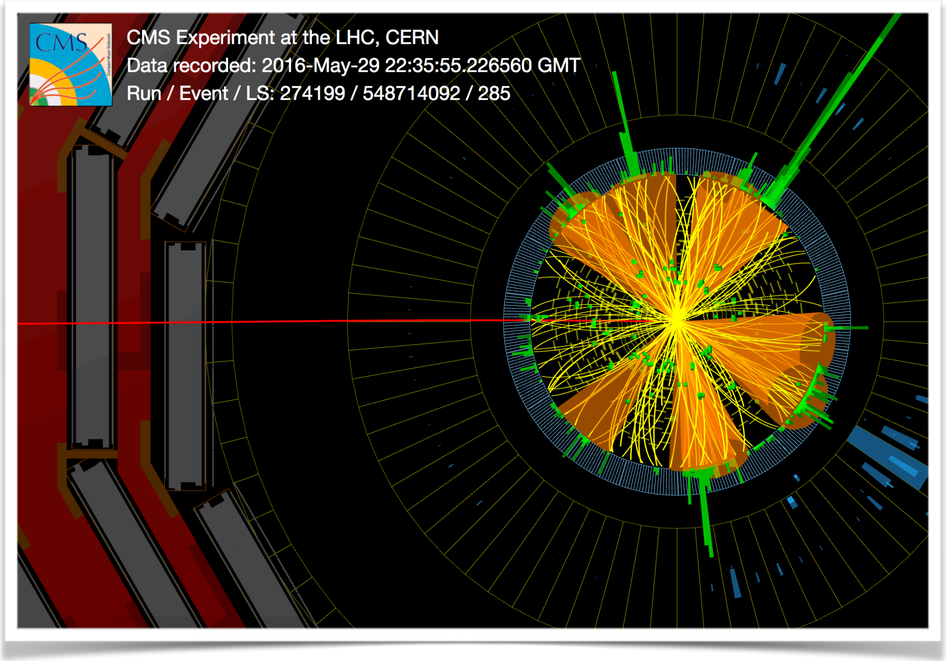
# BACKUP

# Micron Deep Learning Accelerator



# Machine/Deep Learning in HEP

- Machine learning being exploited across particle physics
  - Reconstruction of collision events
  - Particle identification
  - Clustering
  - Energy regression




- Can train a machine/deep learning model to achieve performance of complex algorithms in much lower latency

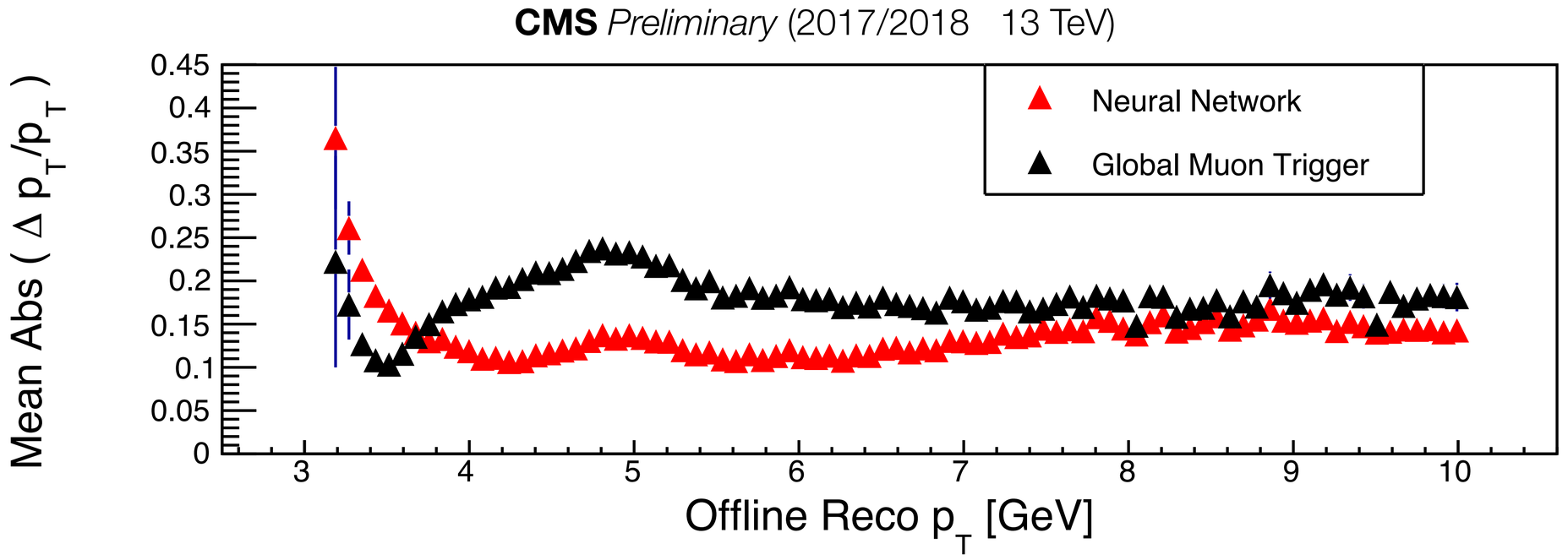
# Model Optimization Detailed Results

| Id | Loss fnc. | Opt. | Regular. | Act.hl. | Act.ol |
|---|---|---|---|---|---|
| 0 | mse | Adam | None | Relu | Linear |
| 1 | msle | Adam | None | Relu | Linear |
| 2 | logcosh | Adam | None | Relu | Linear |
| 3 | logcosh | SGD | None | Relu | Linear |
| 4 | logcosh | Adadelta | None | Relu | Linear |
| 5 | logcosh | Adadelta | L1 $\lambda$=$10^{-5}$ | Relu | Linear |
| 6 | logcosh | Adadelta | L2 $\lambda$=$10^{-5}$ | Relu | Linear |
| 7 | logcosh | Adam | L2 $\lambda$=$10^{-5}$ | Relu | Linear |
| 8 | logcosh | Adadelta | L2 $\lambda$=$10^{-7}$ | Relu | Linear |
| 9 | logcosh | Adadelta | L2 $\lambda$=$10^{-9}$ | Relu | Linear |
| 10 | logcosh | Adadelta | None | Softmax | Linear |
| 11 | logcosh | Adadelta | None | Softmax | Relu |
| 12 | logcosh | Adadelta | None | Relu | Relu |
| 13 | logcosh | Adagrad | None | Relu | Linear |
| 14 | logcosh | Adam | L2 $\lambda$=$10^{-7}$ | Relu | Linear |
| 15 | logcosh | Adam | None | Softmax | Linear |

# Model Optimization Detailed Results

| | RMS | | | Data in the core [%] | | | Data in the tail [%] | | |
|---|---|---|---|---|---|---|---|---|---|
| Id | $\Delta\phi$ | $\Delta\eta$ | $\Delta p_{\mathrm{T}}/p_{\mathrm{Tr}}$ | $\Delta\phi$ | $\Delta\eta$ | $\Delta p_{\mathrm{T}}/p_{\mathrm{Tr}}$ | $\Delta\phi$ | $\Delta\eta$ | $\Delta p_{\mathrm{T}}/p_{\mathrm{Tr}}$ |
| ext | 0.166 | 0.034 | 0.274 | 14.82 | 4.86 | 10.54 | 41.25 | 15.14 | 44.84 |
| 0 | 0.123 | 0.031 | 0.170 | 19.28 | 5.30 | 14.70 | 27.85 | 11.52 | 27.40 |
| 1 | 0.145 | 0.032 | 0.193 | 16.95 | 4.93 | 11.56 | 37.19 | 15.52 | 36.82 |
| 2 | 0.120 | 0.031 | 0.169 | 20.17 | 5.28 | 14.20 | 27.07 | 11.55 | 28.29 |
| 3 | 0.123 | 0.032 | 0.170 | 18.80 | 5.28 | 13.99 | 28.77 | 11.82 | 28.78 |
| 4 | 0.119 | 0.031 | 0.167 | 19.75 | 5.29 | 14.76 | 26.79 | 11.45 | 26.68 |
| 5 | 0.122 | 0.032 | 0.184 | 19.62 | 5.26 | 13.87 | 27.38 | 11.98 | 29.75 |
| 6 | 0.121 | 0.032 | 0.170 | 19.24 | 5.26 | 13.45 | 28.48 | 11.81 | 29.85 |
| 7 | 0.121 | 0.032 | 0.171 | 19.30 | 5.29 | 13.91 | 27.70 | 11.79 | 28.76 |
| 8 | 0.119 | 0.031 | 0.165 | 19.37 | 5.29 | 14.85 | 26.90 | 11.58 | 26.62 |
| 9 | 0.117 | 0.031 | 0.167 | 19.56 | 5.27 | 14.70 | 26.73 | 11.54 | 26.84 |
| 10 | 0.120 | 0.032 | 0.168 | 19.73 | 5.21 | 14.61 | 26.91 | 12.06 | 27.12 |
| 11 | 0.138 | 0.032 | 0.231 | 17.04 | 4.97 | 8.46 | 34.11 | 13.77 | 49.23 |
| 12 | 0.136 | 0.032 | 0.227 | 16.91 | 5.04 | 8.45 | 34.05 | 13.72 | 49.57 |
| 13 | 0.119 | 0.031 | 0.169 | 19.52 | 5.30 | 14.73 | 27.23 | 11.66 | 26.86 |
| 14 | 0.118 | 0.031 | 0.167 | 20.16 | 5.34 | 14.61 | 26.85 | 11.50 | 27.04 |
| 15 | 0.119 | 0.031 | 0.167 | 19.61 | 5.24 | 14.76 | 26.89 | 11.54 | 26.48 |

# Deep Learning Results, Neural Network vs GMT




On the larger part of the offline reconstructed $p_T$ range, the neural network produces improved results over Global Muon Trigger outputs; better by a factor of two at certain ranges

# Scouting Timeline

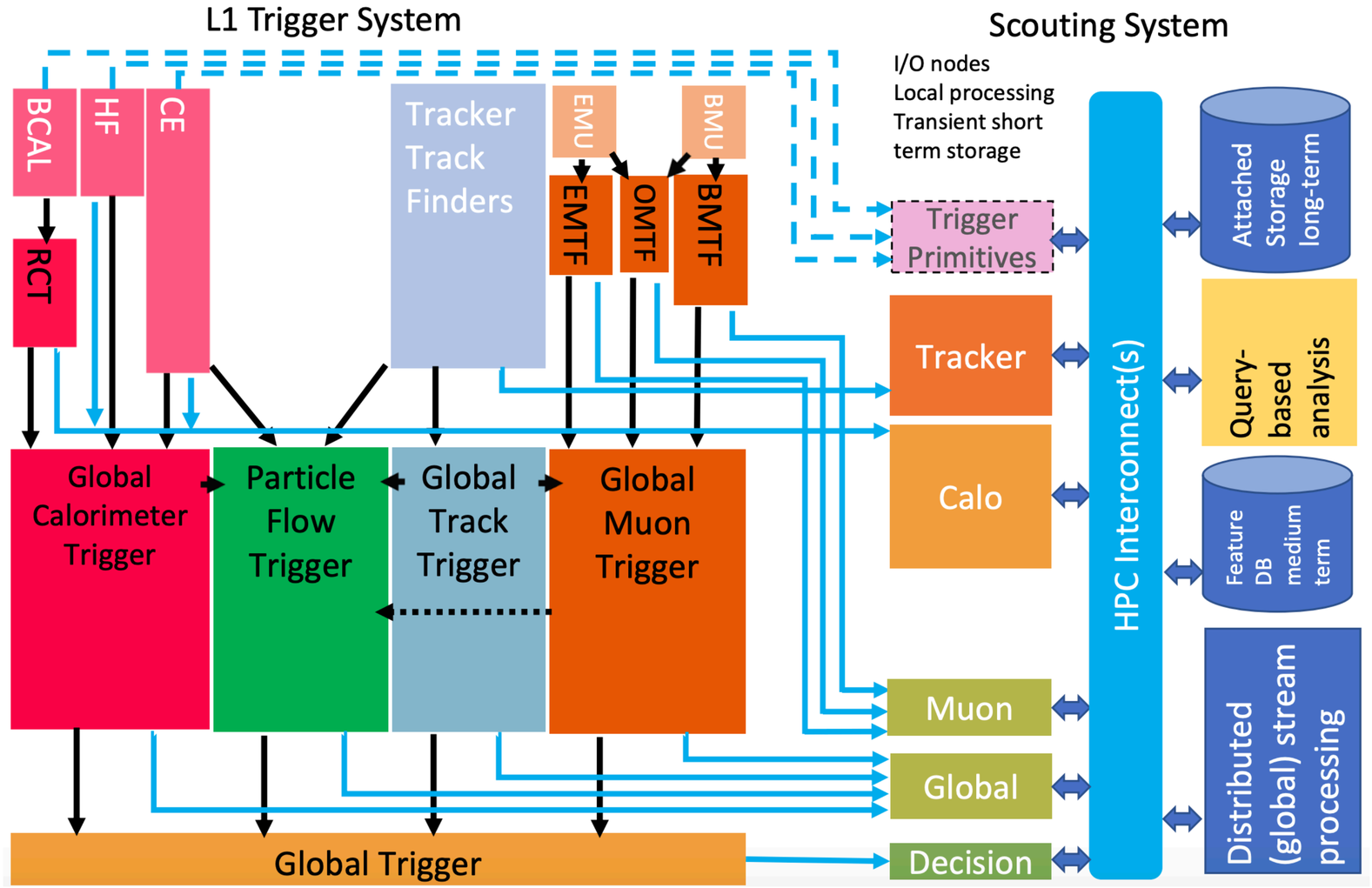
| Year | Type | Areas |
|---|---|---|
| 2011 | HLT | |
| 2018 | L1 | GMT |
| 2021 | L1 | GMT, CL2, KBMTF |
| 2027 | L1 | CMS Phase 2 Trigger Objects |

GMT - Global Muon Trigger

CL2 - Calorimeter Trigger Layer 2

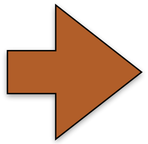
KBMTF - Kalman Filter Barrel Muon Track Finder

# L1 Trigger Scouting Phase-2 Infrastructure



# L1 Trigger Muon Objects

GMT



| Muon Object | | | |
|---|---|---|---|
| Parameter | Range | Step | Bits |
| *φ (extrapolated)* | 2π | 2π/576~0.011 | 10 |
| *η (extrapolated)* | -2.45 .. 2.45 | 0.0870/8=0.010875 | 8+1=9 |
| $P_T$ | 0 .. 255 GeV | 0.5 | 9 |
| Charge Valid | | | 1 |
| Charge Sign | | | 1 |
| Quality | | | 4 |
| Iso | | | 2 |
| Index Bits | | | 7 |
| *φ (raw, from trackfinder)* | 2π | 2π/576~0.011 | 10 |
| *η (raw, from trackfinder)* | -2.45 .. 2.45 | 0.0870/8=0.010875 | 8+1=9 |
| Reserved | | | 2 |
| **TOTAL** | | | **64** |

# Micron DLA Performance

| | Throughput [inferences/s] |
|---|---|
| AC-510, SB-852 - 1 cluster | > $1.3 * 10^6$ |
| AC-510 - 2 clusters | > $2.7 * 10^6$ |
| Required | $1 * 10^6$ |