Triggering on Displaced Muons with a Fast NN in CMS Endcap Muon Track Finder

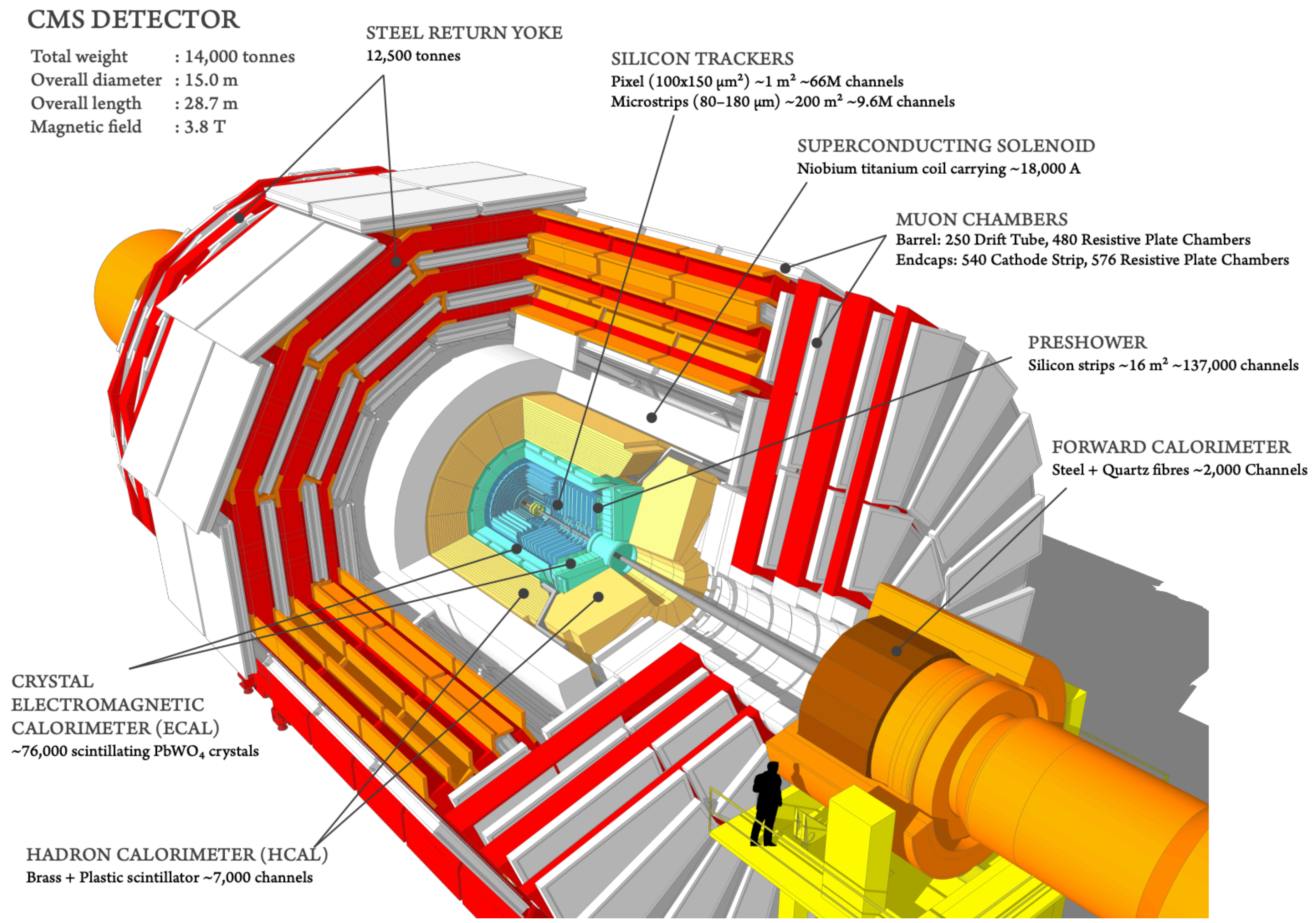
Efe Yiğitbaşı[1], Darin Acosta[1], Osvaldo Miguel Colin[1], Aleksei Greshilov[1], Sergo Jindariani[2], Patrick Kelling[1], Jacobo Konigsberg[3], Jia Fu Low, Alexander Madorsky[3]
on behalf of CMS collaboration

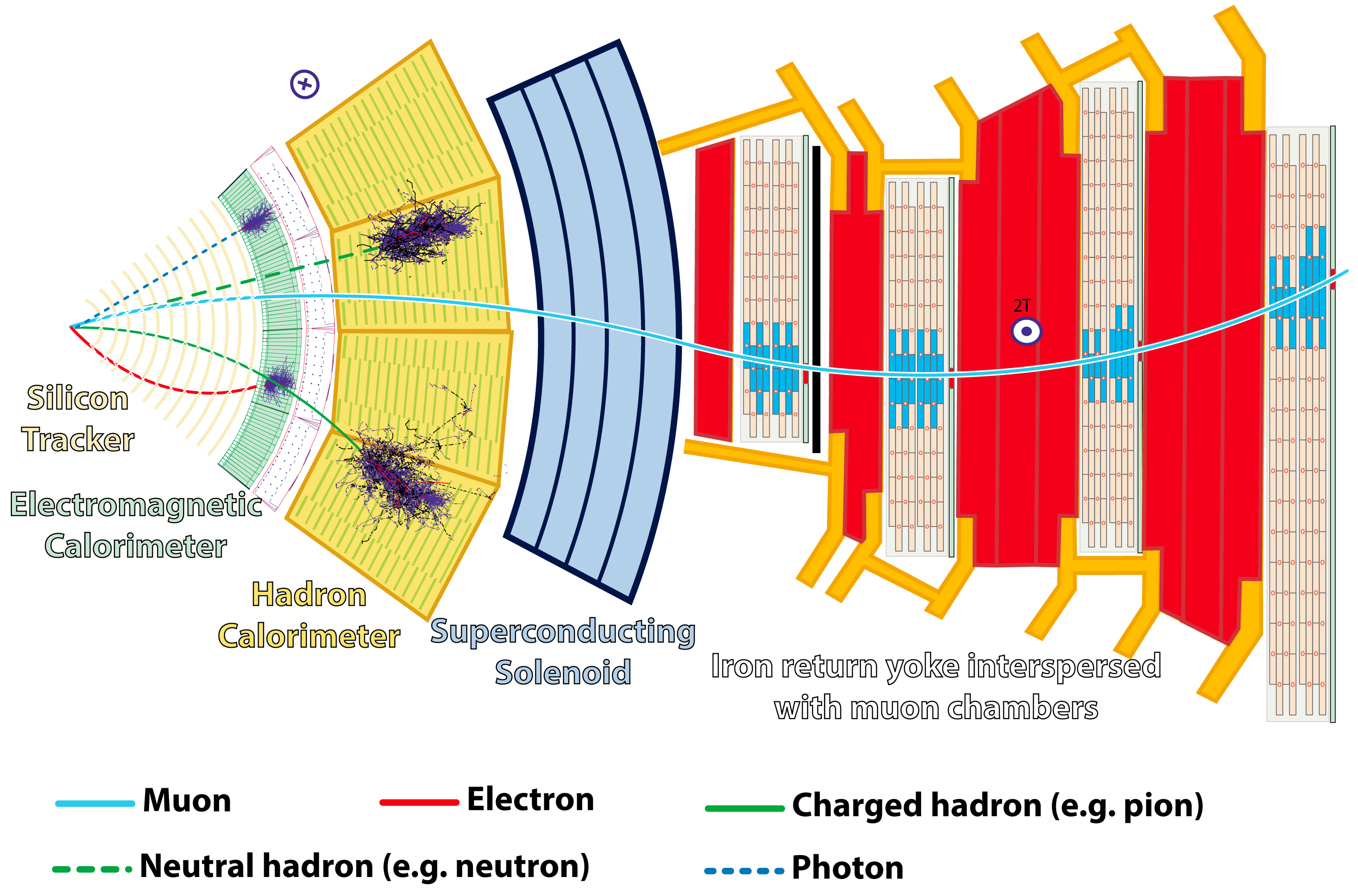
18th June 2025

[1] RICE
[2] Fermilab
[3] UF UNIVERSITY of FLORIDA

CMS Experiment at the LHC

- CMS is one of the two general purpose experiments at the LHC
  - Strong solenoid magnet with 3.8 T magnetic field
  - Multi layer design including silicon tracker, electromagnetic and hadronic calorimeters and muon systems
  - A two level trigger system to select interesting events out of 40 MHz of LHC collisions.
- In Run 3 of the LHC, one of the main areas of interest for CMS experiment is new physics models with unconventional signatures.
  - For example, models with long-lived particles (LLPs)
  - Usually limited by triggers and reconstruction methods before Run 3

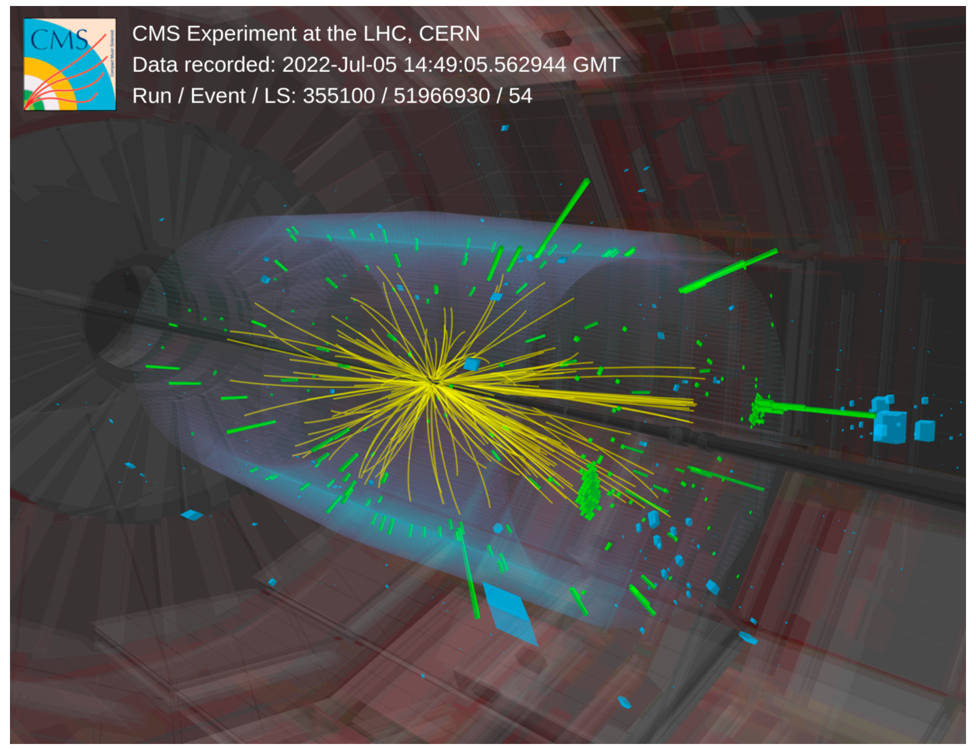




CMS Trigger in Run 3

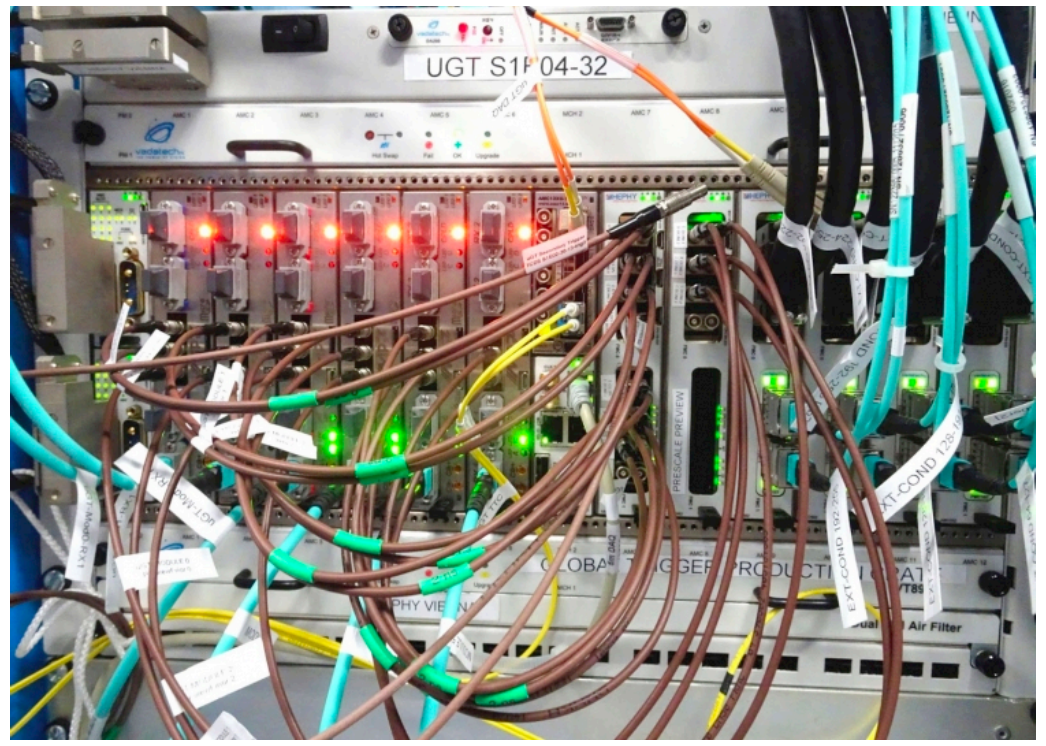
- CMS uses a two level trigger system
- Level-1 Trigger (L1T): Hardware based, using custom electronics (FPGAs)
  - Uses simplified readout, output rate is 110 kHz. **Total latency budget is < 4 μs.**
- High Level Trigger (HLT): Software based, using CPU/GPU farms
  - Uses full event readout with simplified reconstruction, output rate is ~7 kHz for full offline reconstruction and ~25 kHz for trigger-level reconstruction.
- In Run 3, new algorithms were added in L1T & HLT to trigger on rare and unconventional signatures.
  - One area of focus was triggers targeting LLP signatures.
    - Particles from LLP decays can be significantly displaced from the primary interaction point, which can be difficult to capture with traditional triggers

Detector Events
@ 25 ns



40 MHz →

L1T
<4 μs



110 kHz →

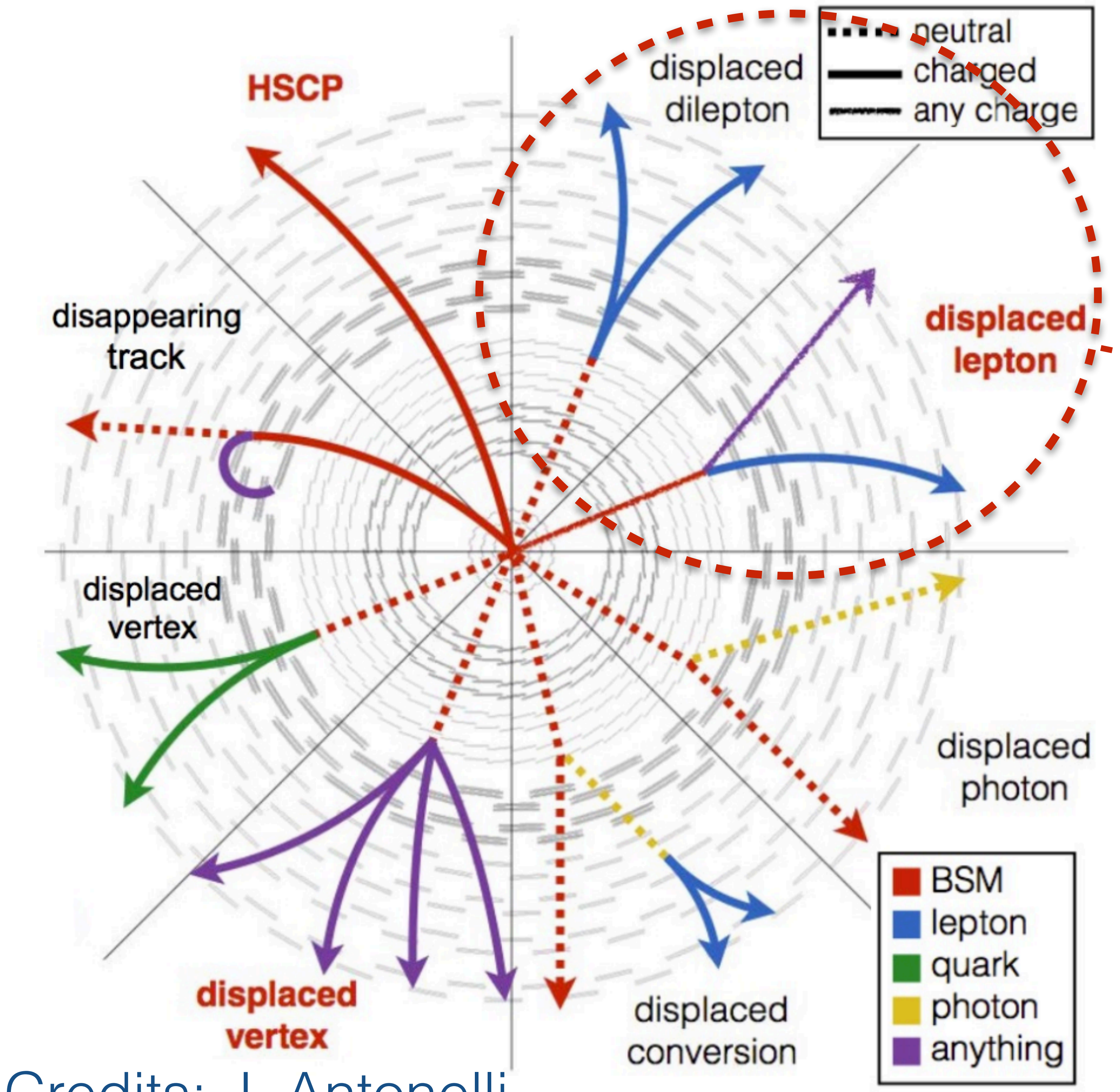
HLT
~500 ms

~7 kHz full reco
~25 kHz trigger-level reco →

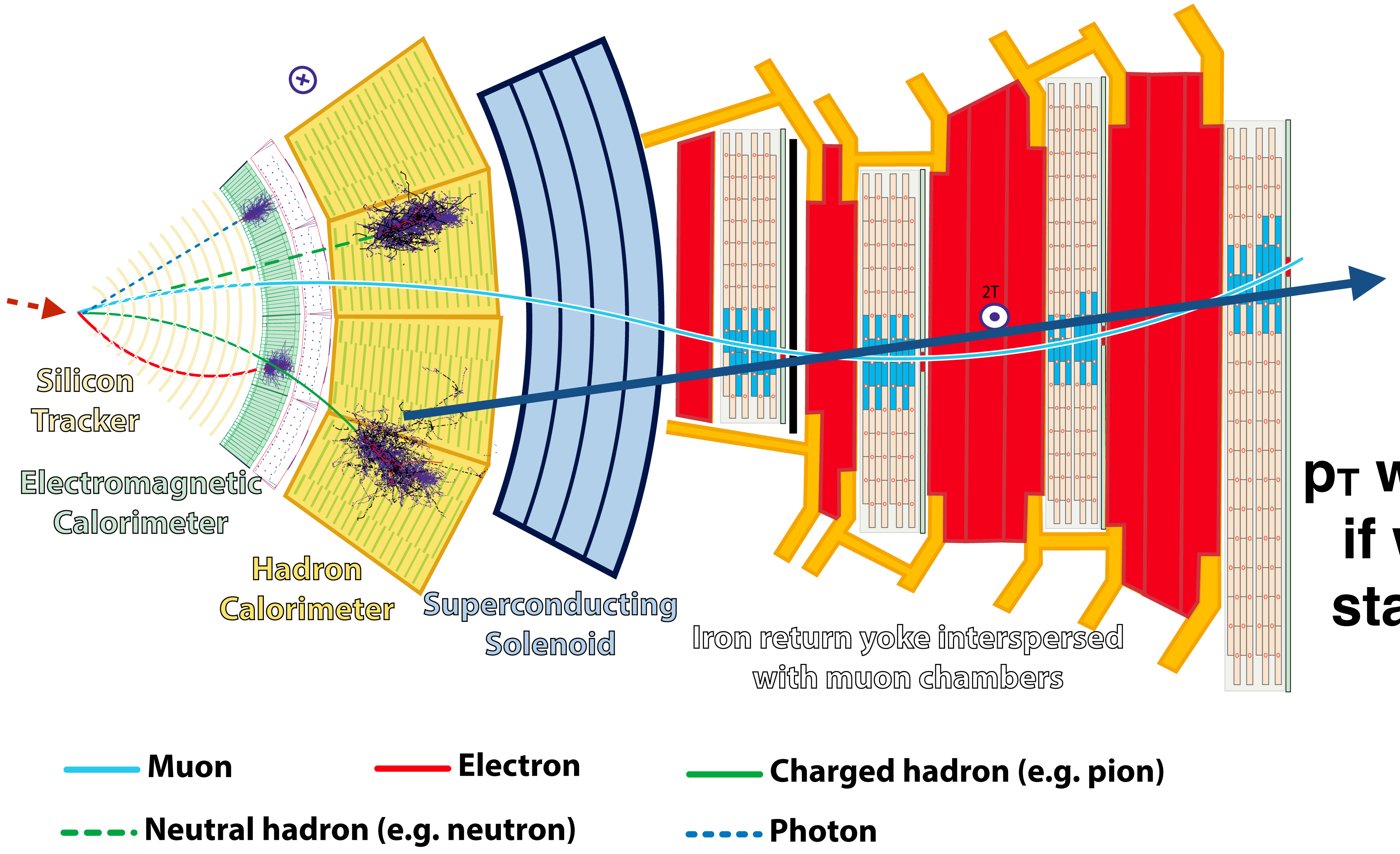
Data storage
for offline analyses

LLPs and Displaced Muons

- Muons from LLP decays originate from displaced vertices.
  - Traditional L1T algorithms are optimized for prompt particles and sometimes even have vetoes to suppress non-collision backgrounds (beam halo, cosmic muons, cavern backgrounds…)
  - In L1 muon trigger systems, the momentum assignment was implicitly or explicitly assuming the interaction point to be another fixed point of the muon track.
    - Leads to an underestimation of p_T for a displaced muon based on the transverse displacement of the muon track (d_{xy})
  - In Run 3, all L1 muon trigger systems added new algorithms for momentum assignment to increase CMS efficiency to displaced muons.



Credits: J. Antonelli

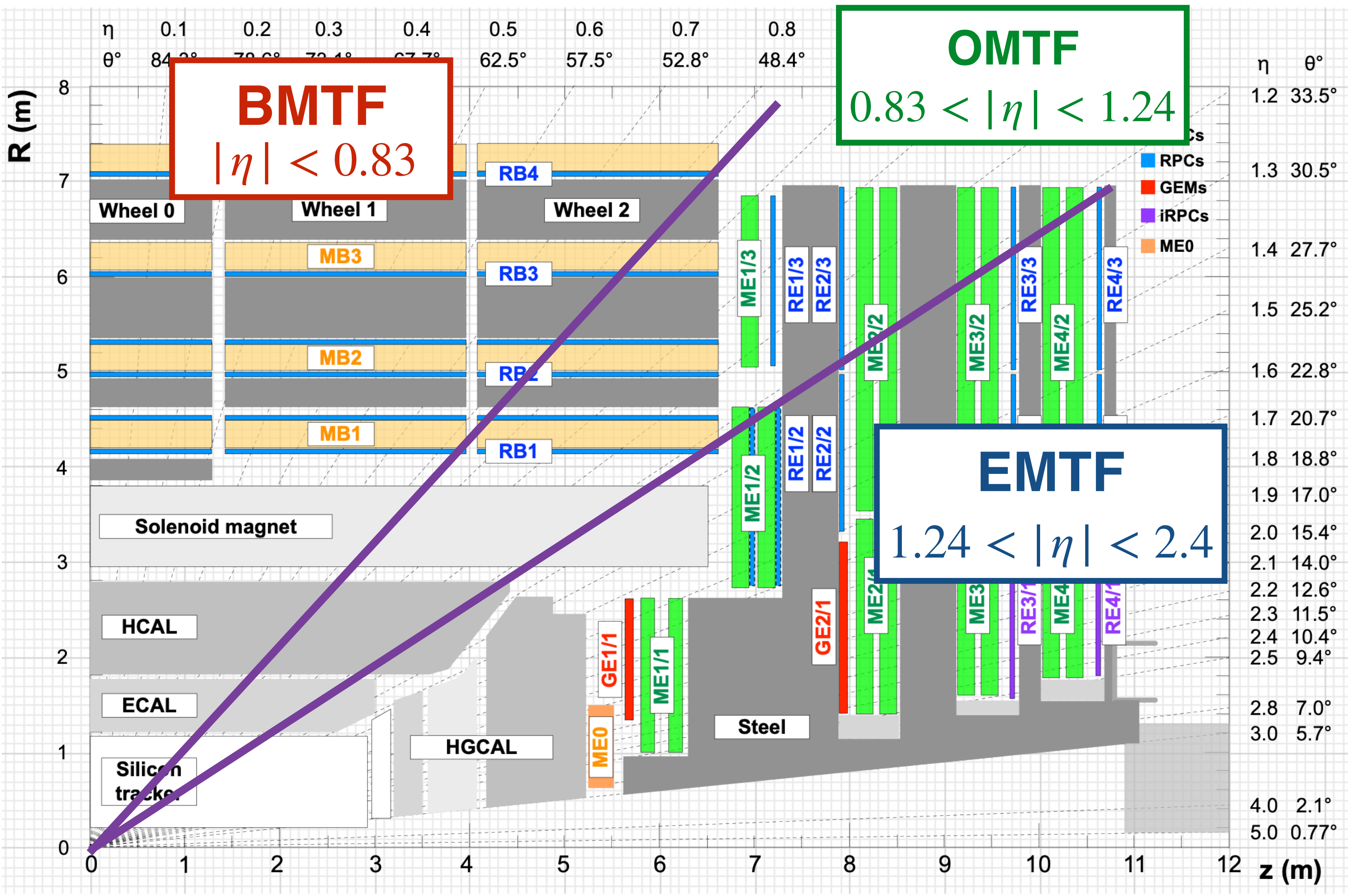


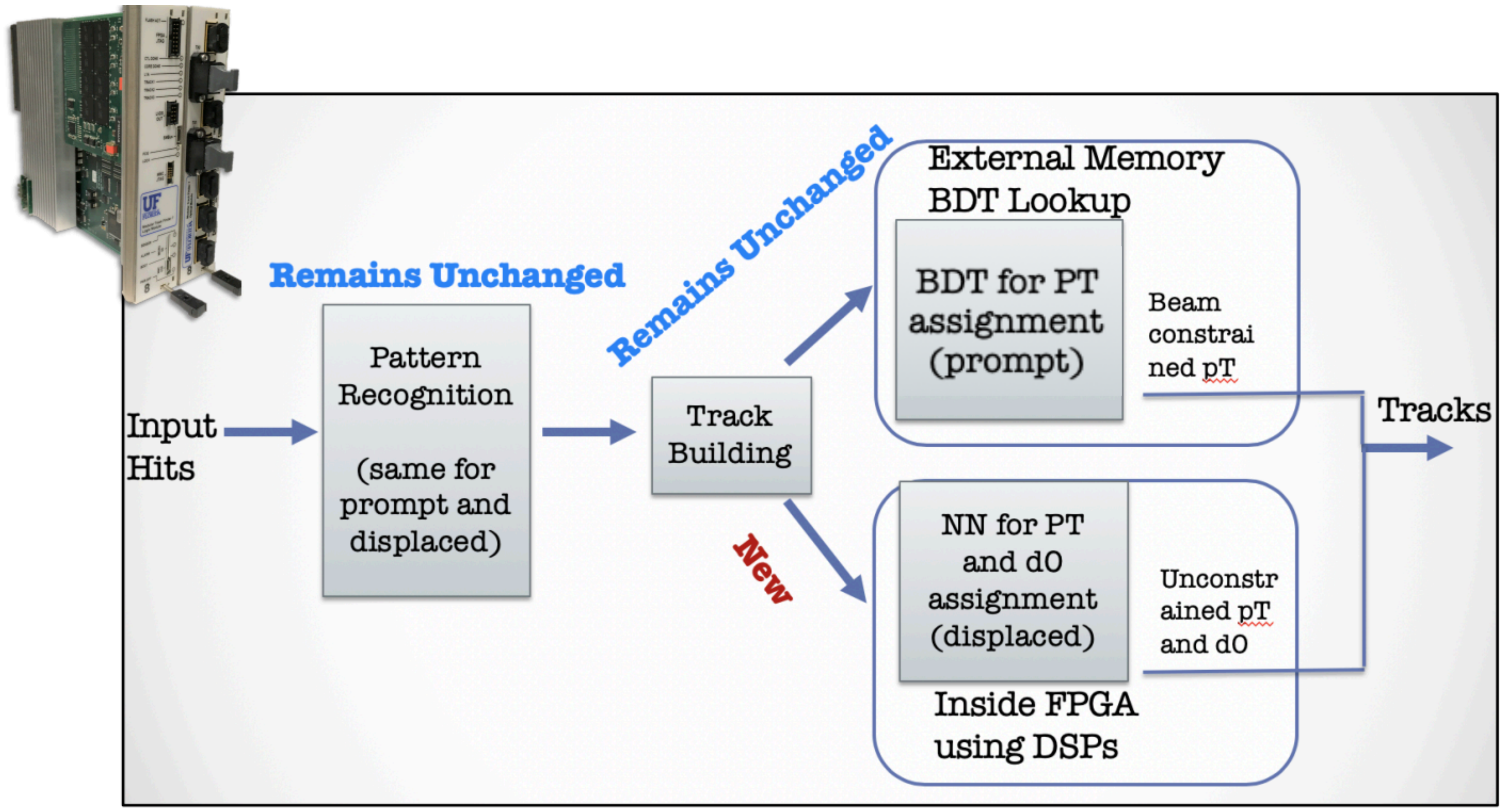
A muon from a displaced vertex

p_T will be underestimated if we assume the track starts at the interaction point

Endcap Muon Track Finder

- CMS L1 muon trigger system uses three different track finders (TF) in different η regions
- In CMS endcaps, Endcap Muon Track Finder (EMTF) builds muon tracks and measures track p_T, η, ϕ etc. in a very short timescale.
  - ~500 ns total latency budget
- CMS endcaps are a challenging environment for triggering
  - Non-uniform magnetic field in the endcaps whose effect gets weaker in the forward direction
  - Different detector technologies with different spatial and timing resolutions
  - Large collision backgrounds which increase with increasing η which can lead to non-linear pileup dependence
- These challenges create an ideal problem for ML based solutions.
  - Since Run 2, EMTF used a BDT based momentum assignment algorithm which was optimized for prompt muons
  - In Run 3, we added a new NN based momentum assignment algorithm which targets displaced muons

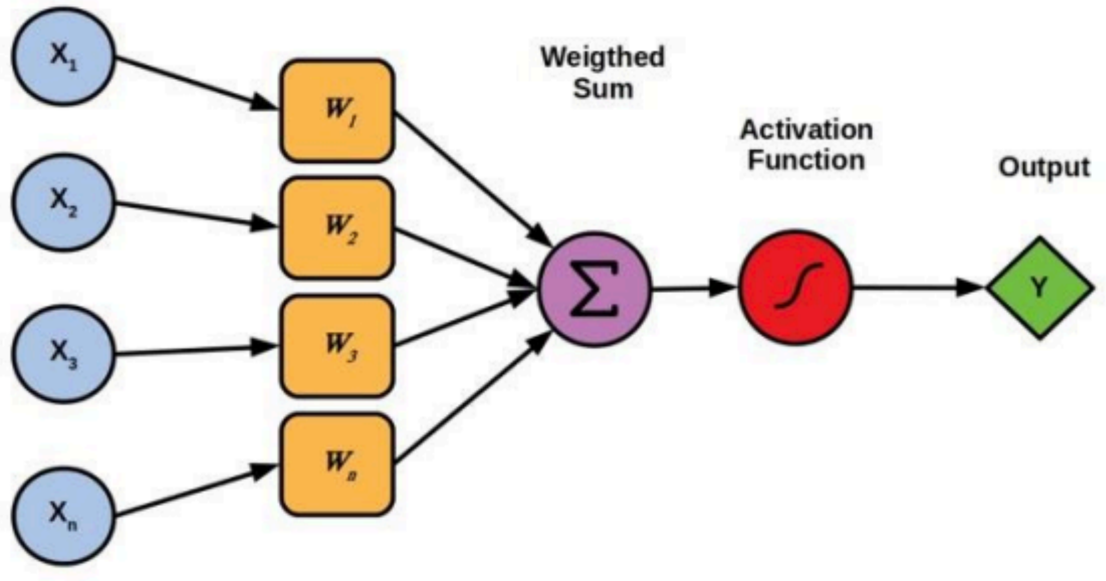


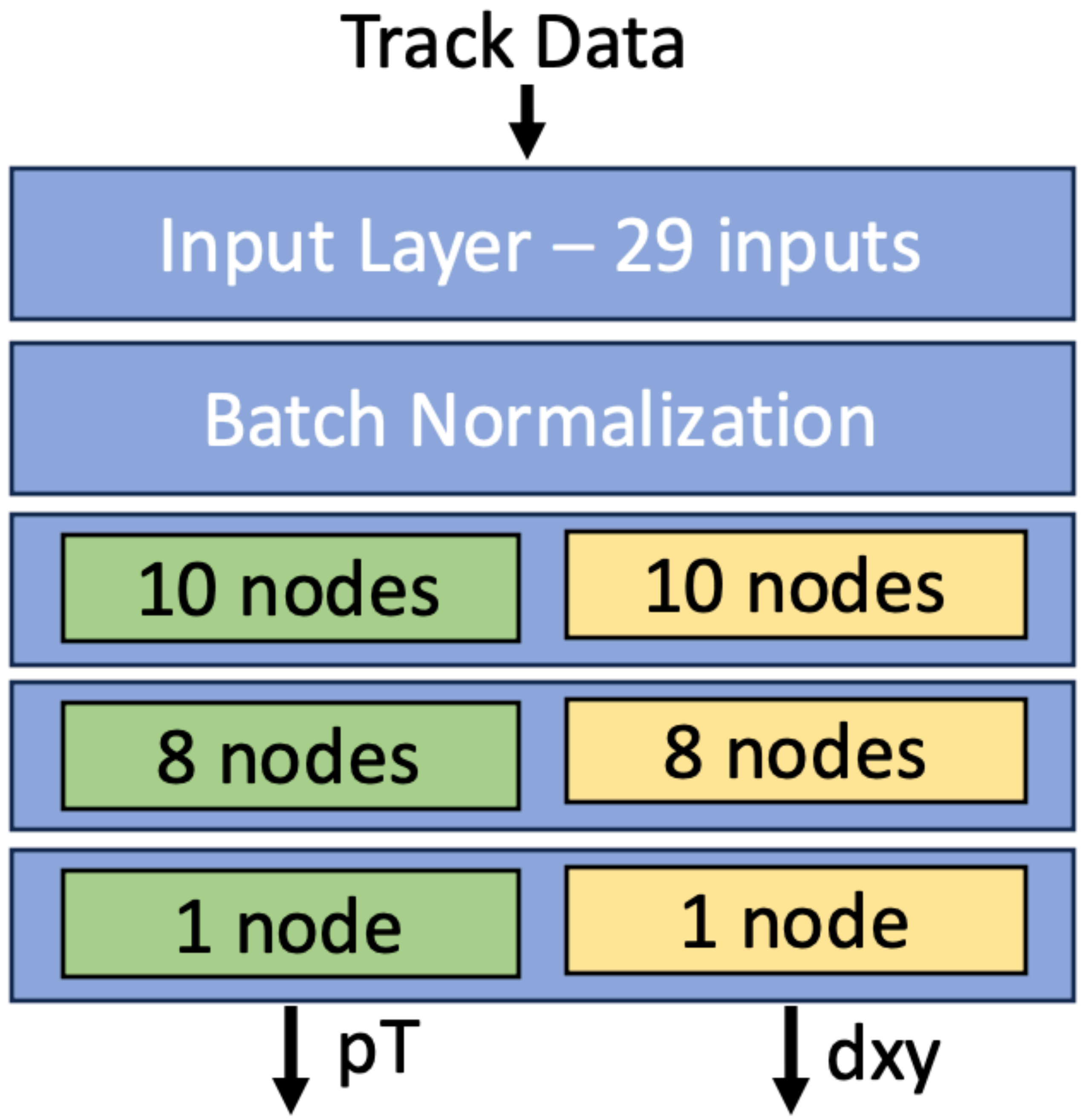


EMTF NN Design

- The goal:
  - Integrate a p_T assignment algorithm for displaced muons that runs in parallel to the BDT based prompt p_T assignment
- Constraints:
  - ~100 ns latency budget to be able to run in parallel
  - Virtex 7 FPGA. Resources used before adding the new algorithm:
    - 74% of FPGA LUTs
    - **2% of DSPs**
    - 76% BRAM
    - 25% Flip Flops
- We needed to restrict LUT usage as much as possible, but have basically no restriction on DSP usage
  - NN implementation in FPGAs can be made to use mostly DSPs, which makes an NN based algorithm the ideal solution for the displaced muon triggers in EMTF.

Single Node of DNN

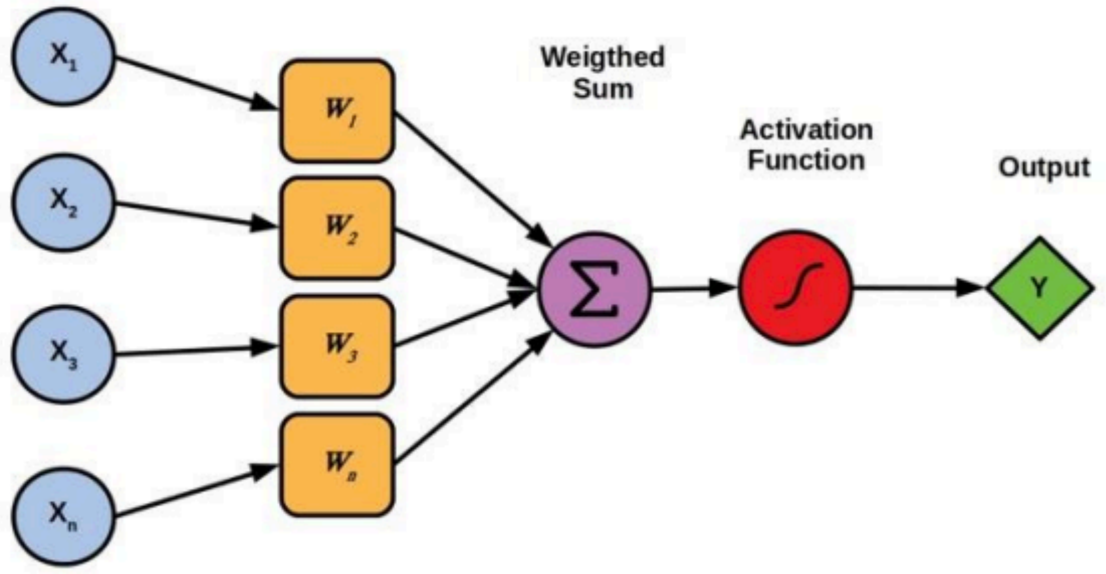


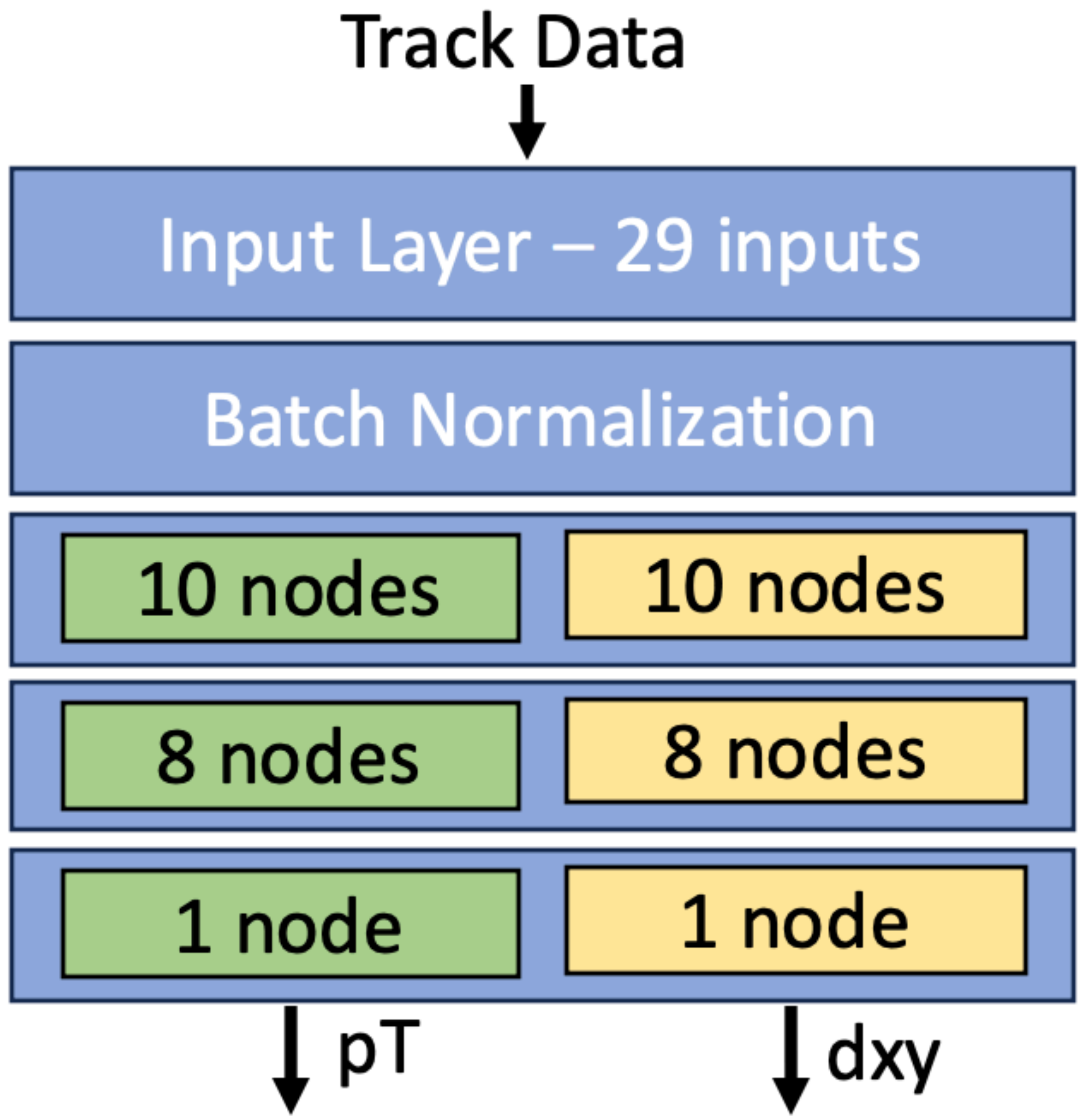


EMTF NN Design

- Keras NN Model:
  - Model has an initial batch normalization layer and 3 dense layers
    - All activations are Rectified Linear Units (ReLU)
  - Two separate NNs, targeting p_T and d_{xy} separately, each using half of the total nodes
  - Trained using logcosh loss functions (similar to Huber loss used by BDT)
  - Normalization is shared between NNs, so we 'stitch' the NNs together before converting to HLS using hls4ml
- Training Data:
  - Created using CMS software emulation of the L1 Trigger, with simulated samples of muons originating from displaced vertices and using the tracks built from these displaced muons
  - Flat in d_{xy} up to 120 cm
  - Flat in $1/p_T$ up to 1 TeV
- After training we quantize the model to fixed point precision using hls4ml [1]

Single Node of DNN

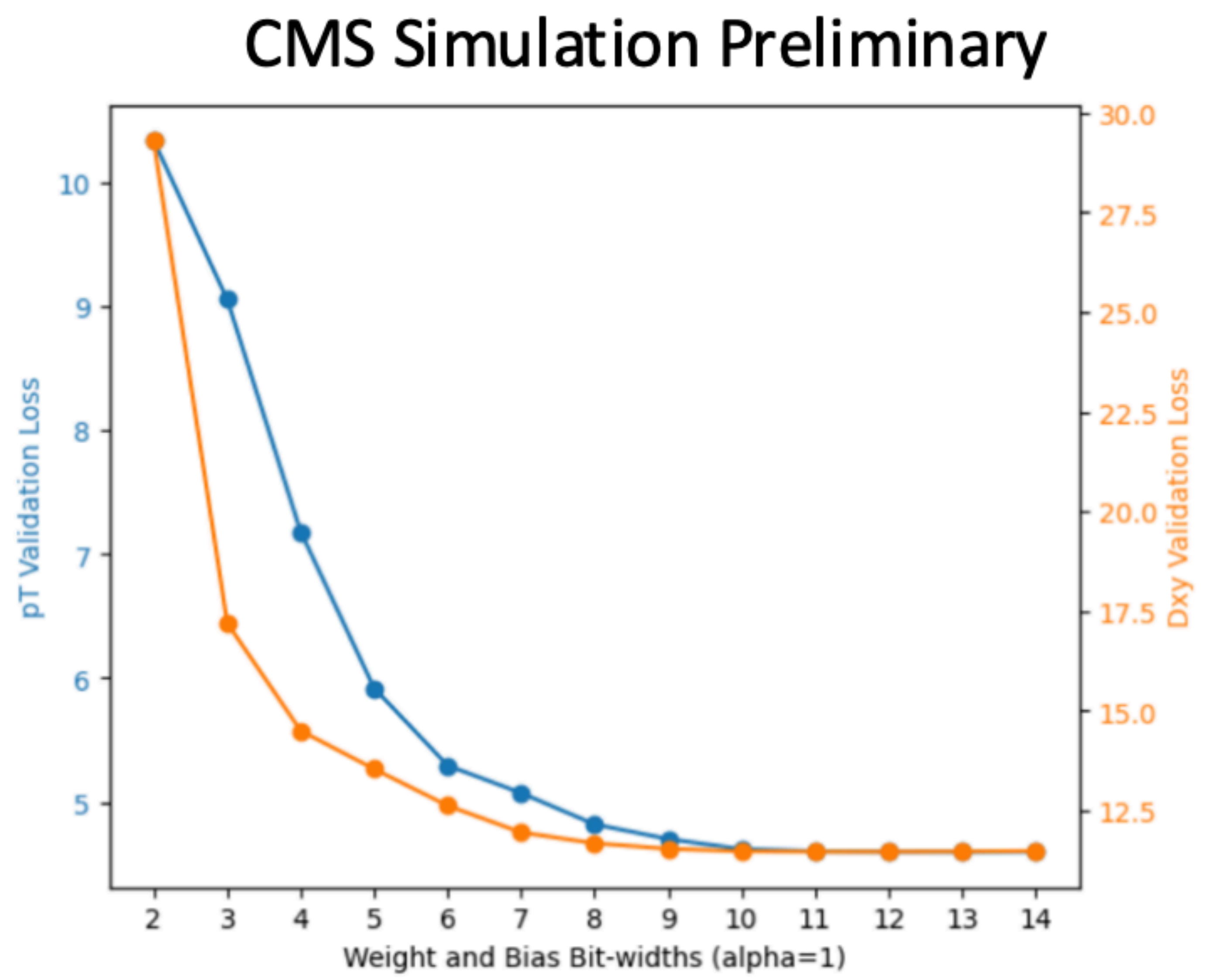


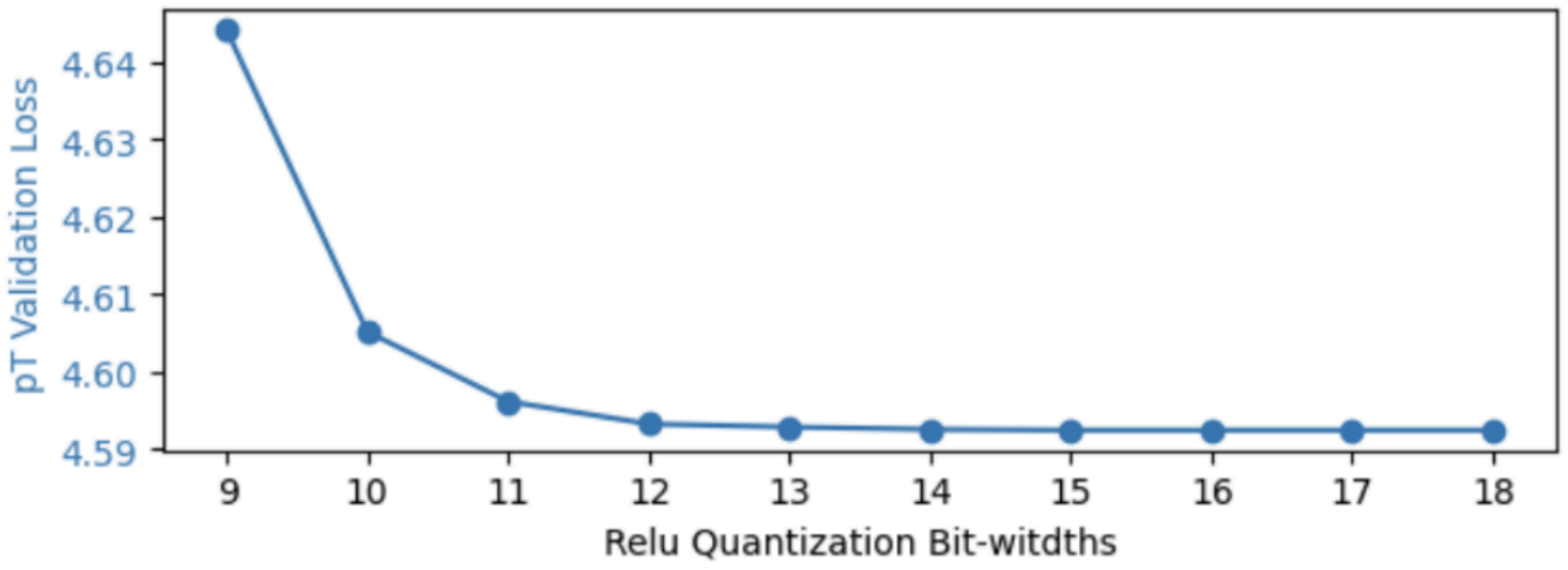


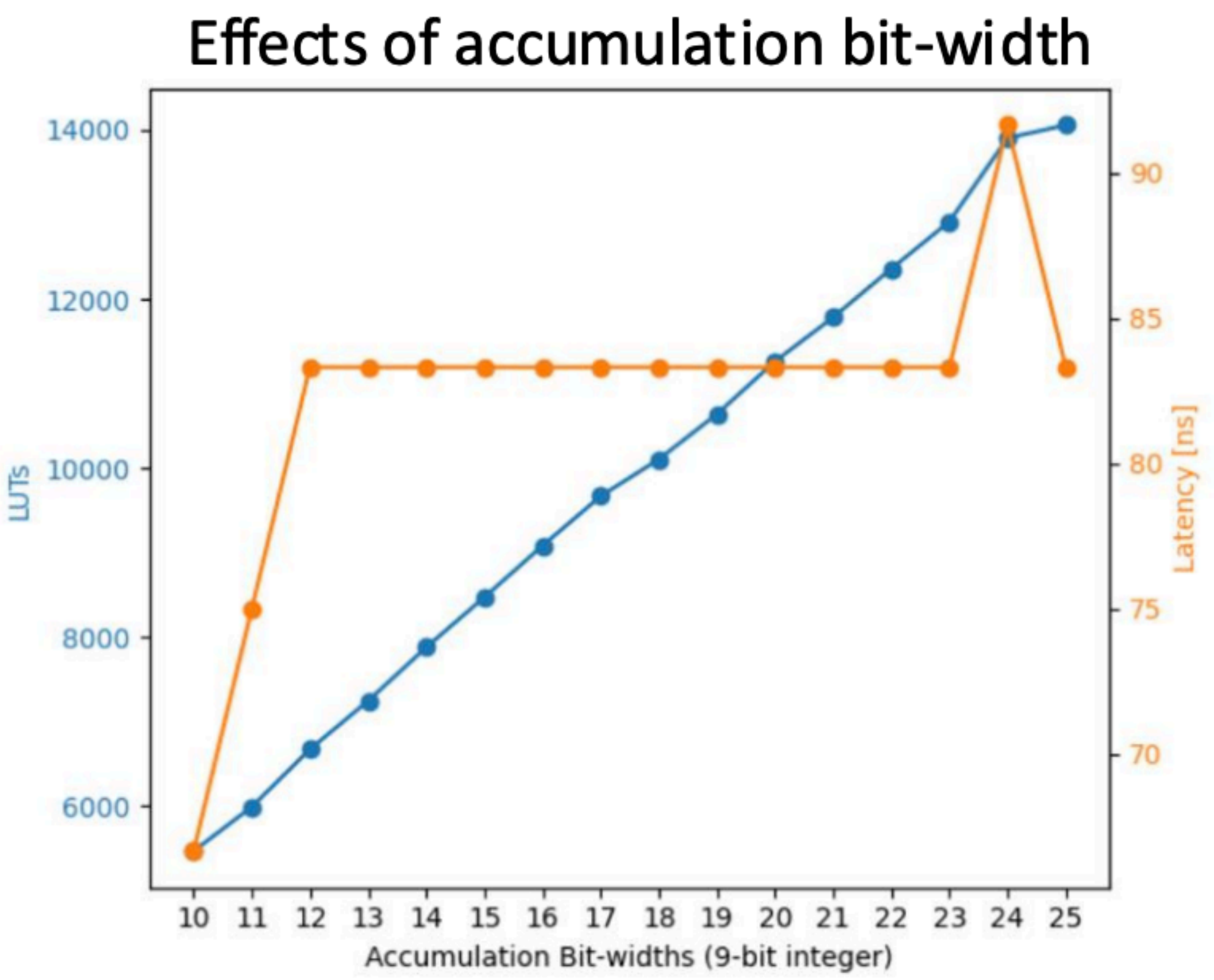
[1] https://arxiv.org/abs/1804.06913

EMTF NN Implementation

- It was very challenging to implement an NN with required performance while fitting within the FPGA resource budgets.
  - Many iterations through the years which finally converged in 2023 with the help of a few optimizations.
- Decided to implement a wrapper around the NN model with set latency which also handles input/output conversions.
  - About half of our inputs are 1-bit values. With a wrapper and without an initial batch normalization layer, our first dense layer multiplications become (weight * 1-bit number). Leads to lots of resource savings and lowers latency.
- Quantizing the model was tricky.
  - We cannot quantize to very low precisions for weights and ReLU Activations without degrading performance. DSP usage goes down, but this is not a number we are worried about. By tuning our accumulation bit-widths we can achieve large LUT savings.

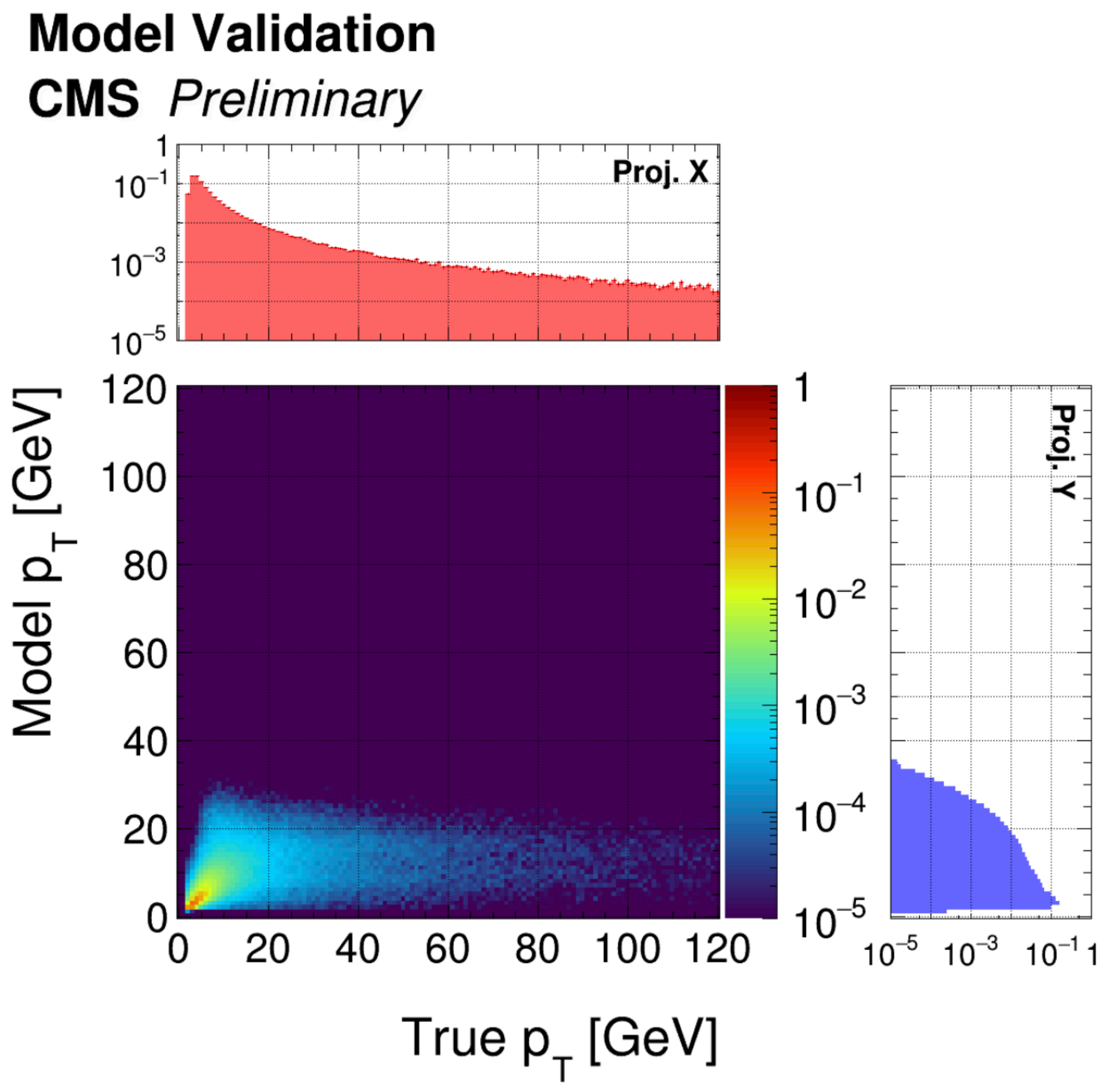


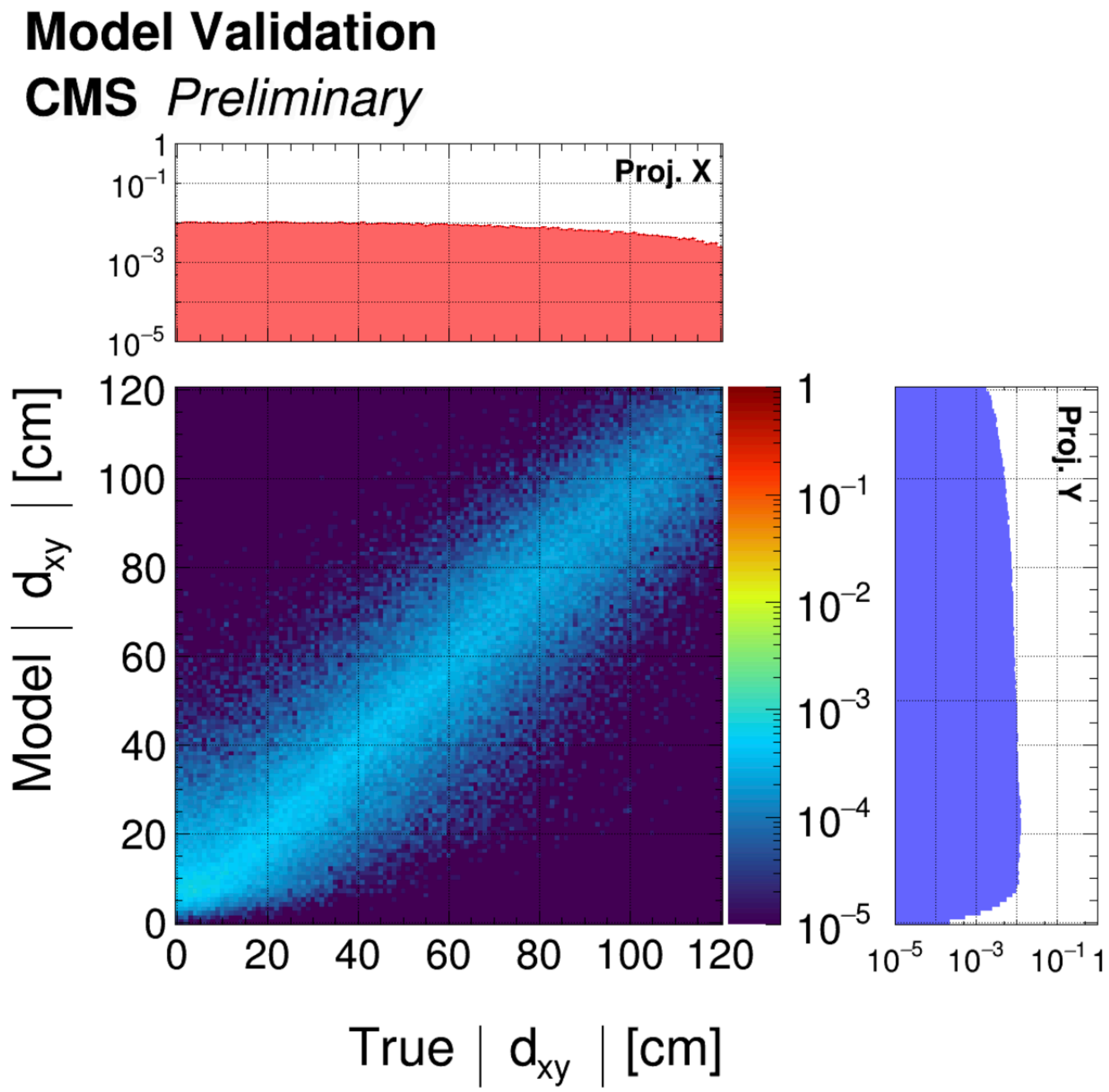




EMTF NN Implementation

- Final Latency: 10 clocks (83 ns)
- Synthesis Report in HLS (NN only)
  - LUTs: 14k (3.2%)
  - DSP: 767 (21.3%)
- Final Vivado Implementation resources with NN
  - LUTs: 76.3% (from 74%)
  - DSPs: 14.4% (from 2%)
- Model validation shows good p_T & d_{xy} regression



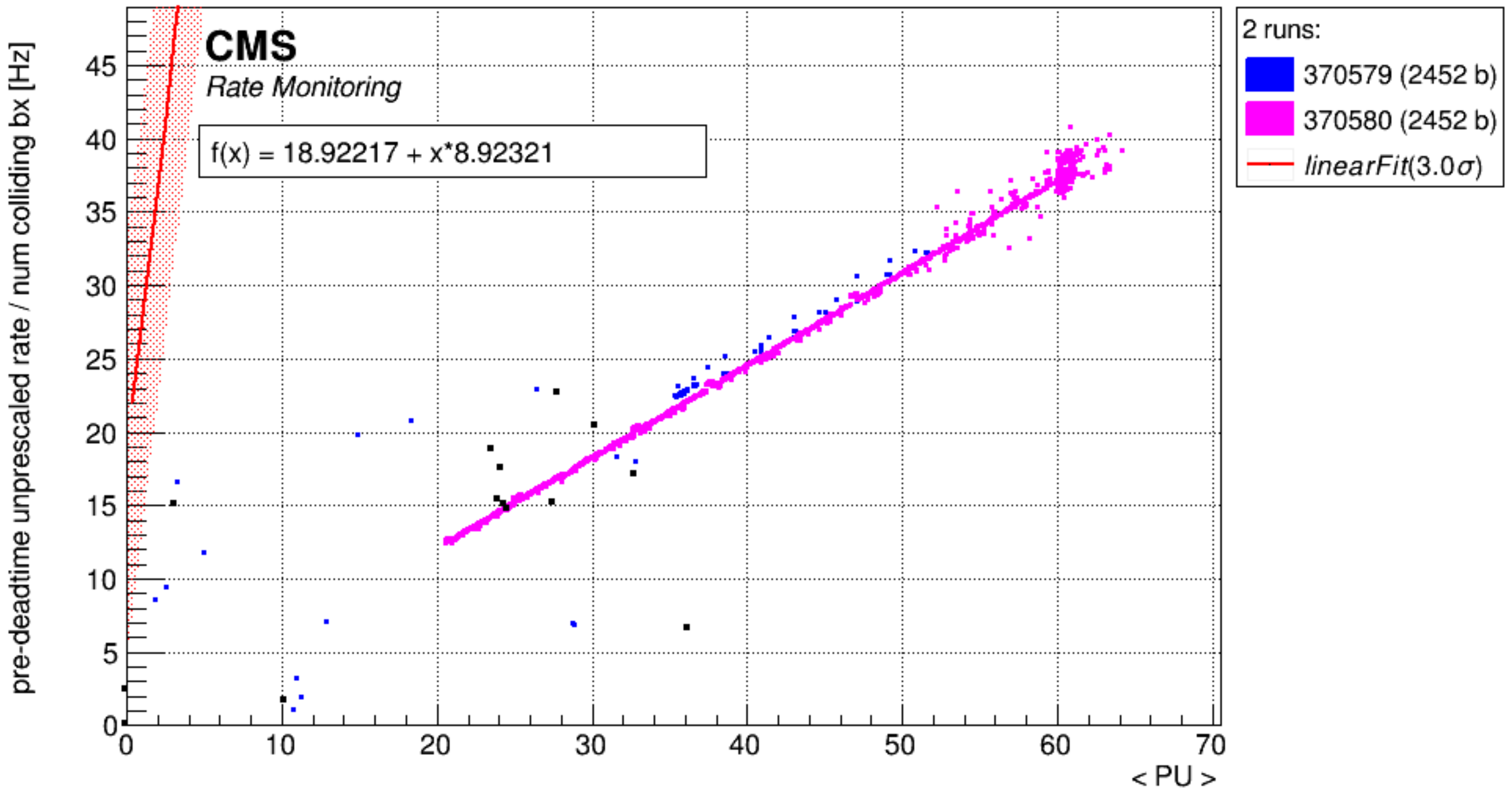


EMTF NN Deployment

- This model has been deployed and running since June 2023 and has been used for triggering from the start of the 2024 run
  - This was the first NN running in CMS L1T FPGAs for data taking
- New triggers using the NN information were implemented in 2024 to extend CMS displaced muon trigger coverage to |η| < 2.0

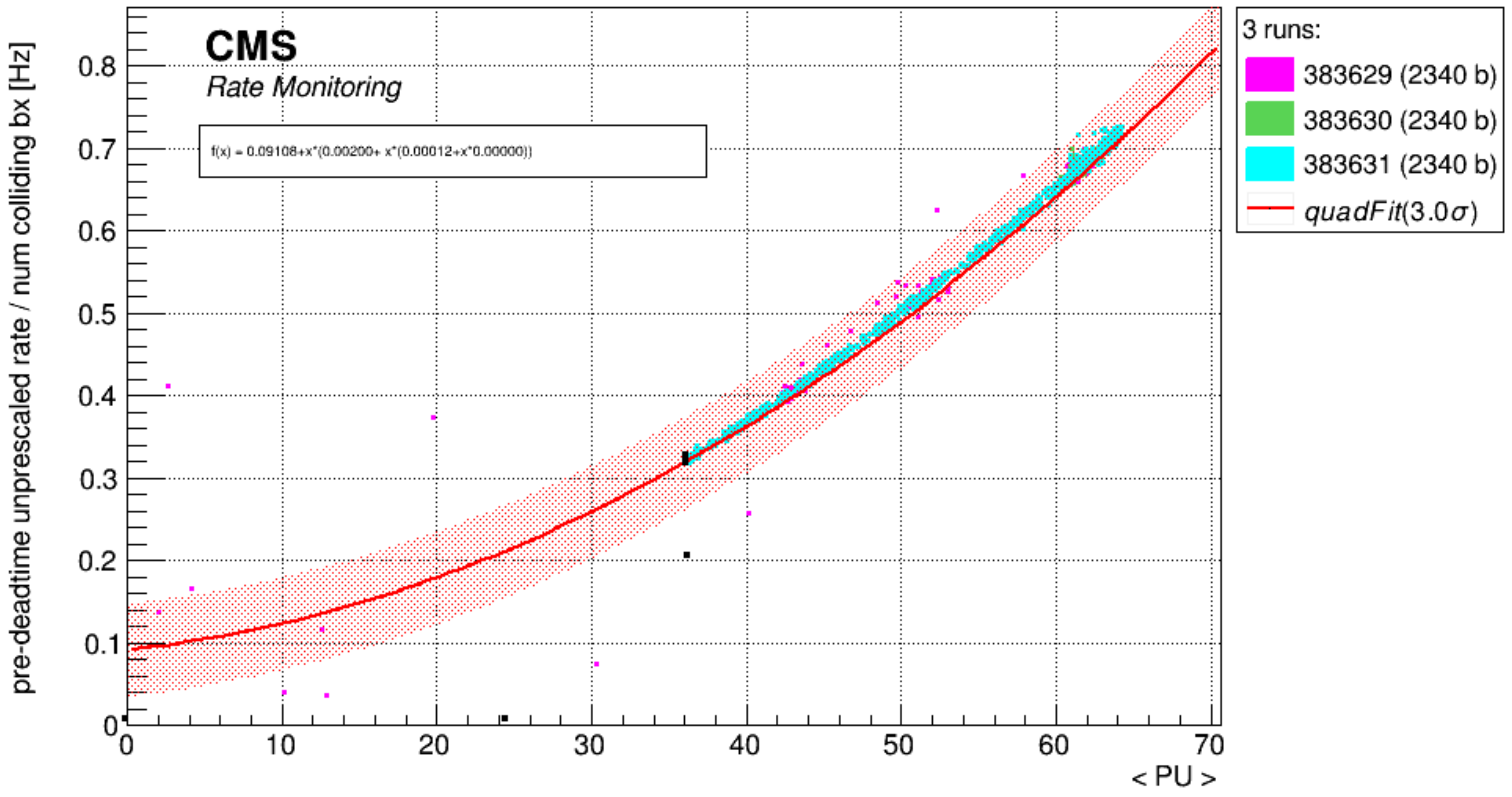
Rate vs PU plot from July 2023
Trigger shown here is for monitoring and commissioning

L1_SingleMu0_Upt10_EMTF



Rate vs PU plot from July 2024
Showing one of the new triggers implemented in 2024

L1_DoubleMu0_Upt6_SQ_er2p0



Expected Performance

- The expected impact of the NN algorithm is at large muon d_{xy}
  - At large d_{xy} prompt p_T assignment underestimates the displaced muon p_T
  - The expected performance based on generator level studies shows ~80% trigger efficiency up to 100 cm of muon d_{xy} at low $|\eta|$ whereas the prompt algorithm efficiency drops very quickly after ~10 cm.
    - The most forward region with $|\eta| > 2.1$ is very challenging for achieving good efficiency while maintaining low trigger rate
- We are still working on getting public performance plots based on 2024 data which should be finalized soon
- These improvements have substantial impact on CMS searches with displaced muons in the final state in Run 3.

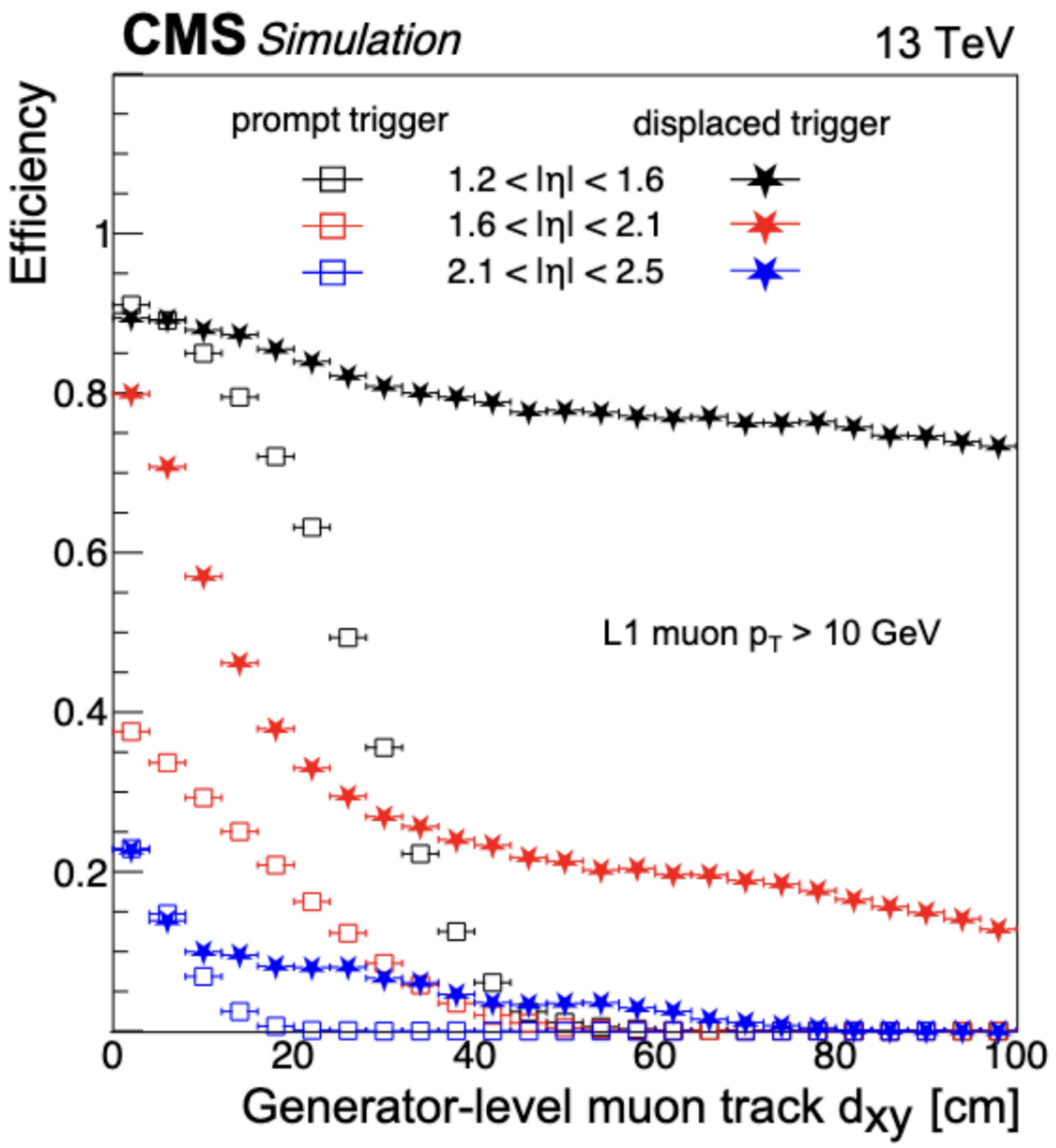
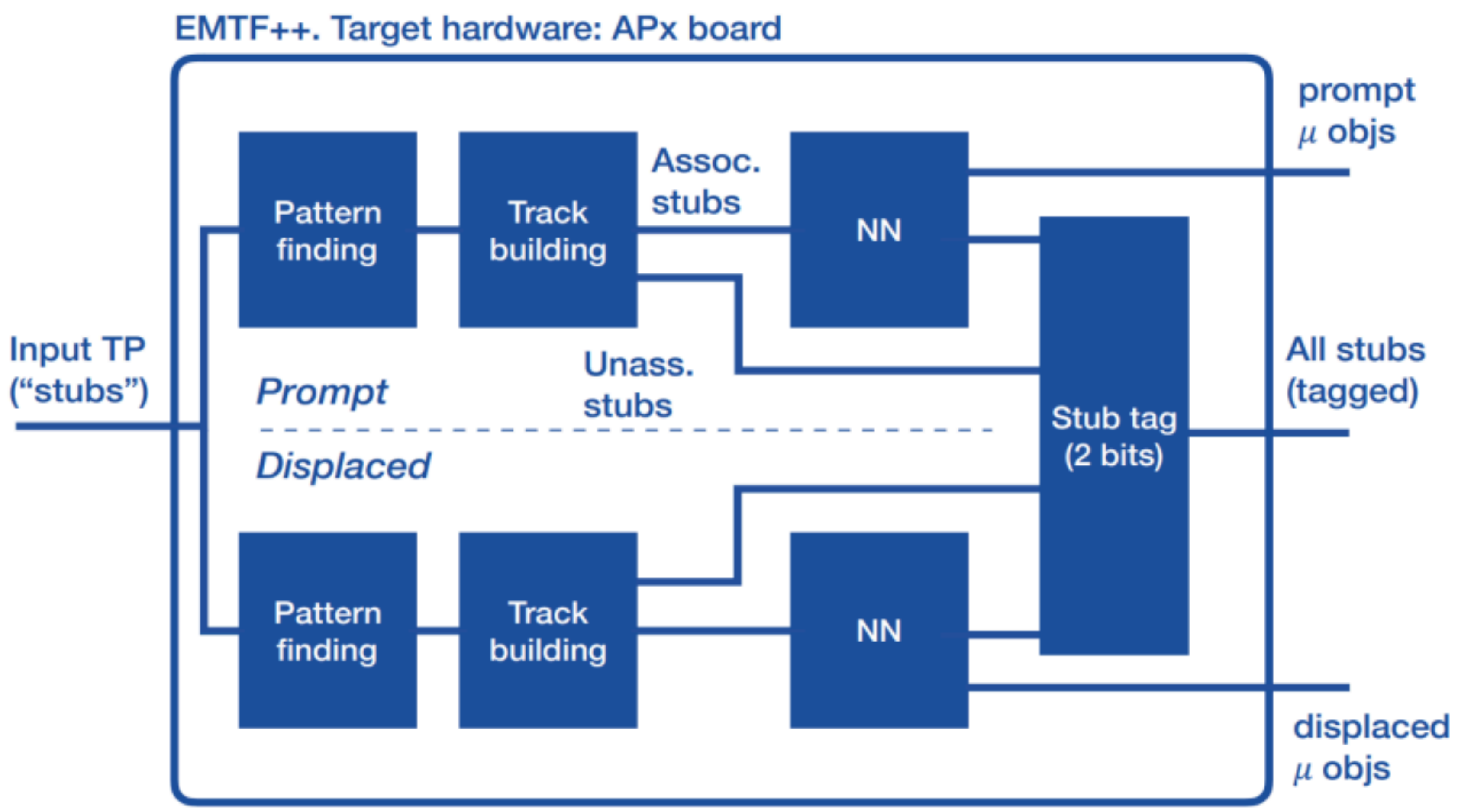
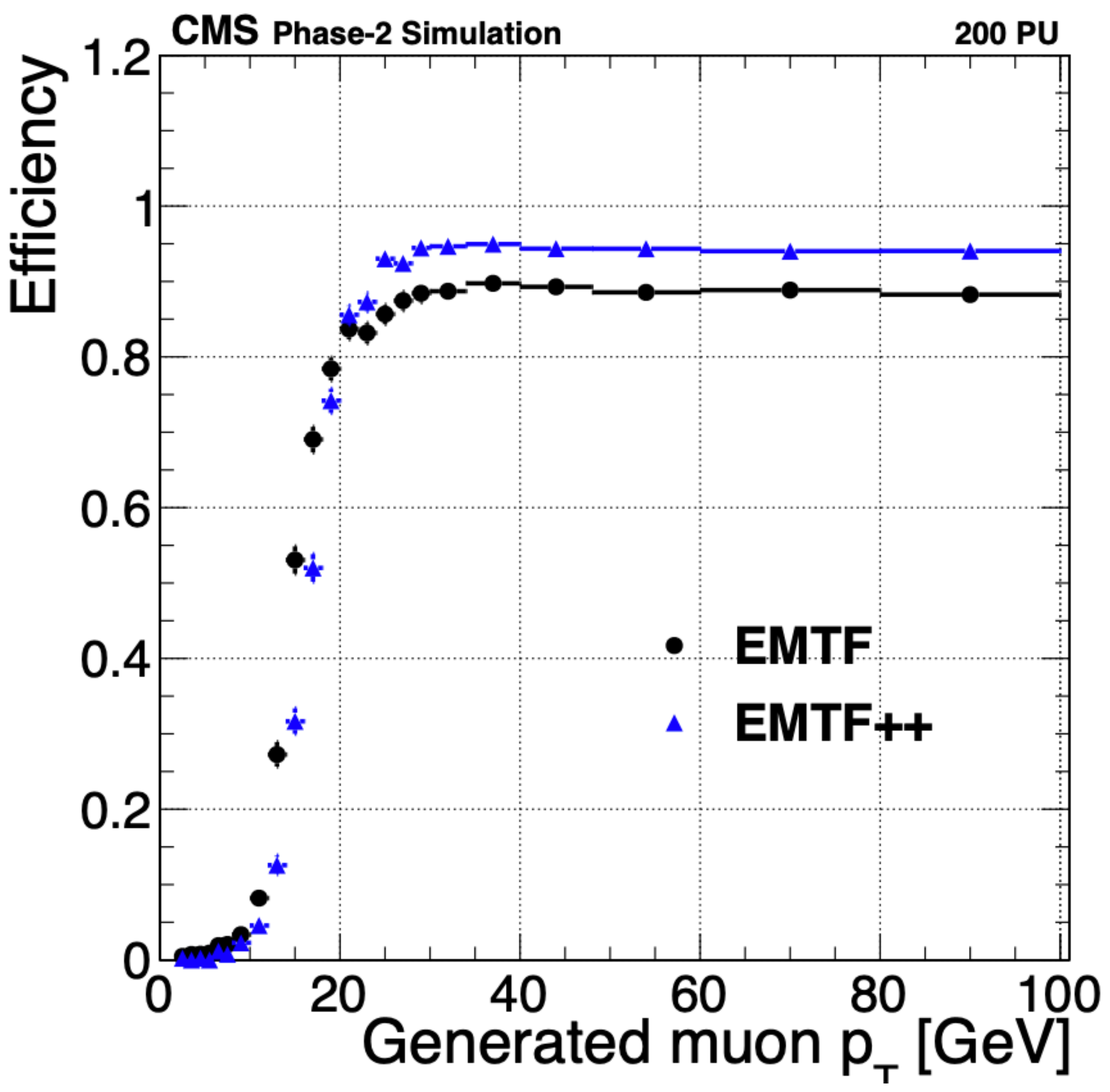


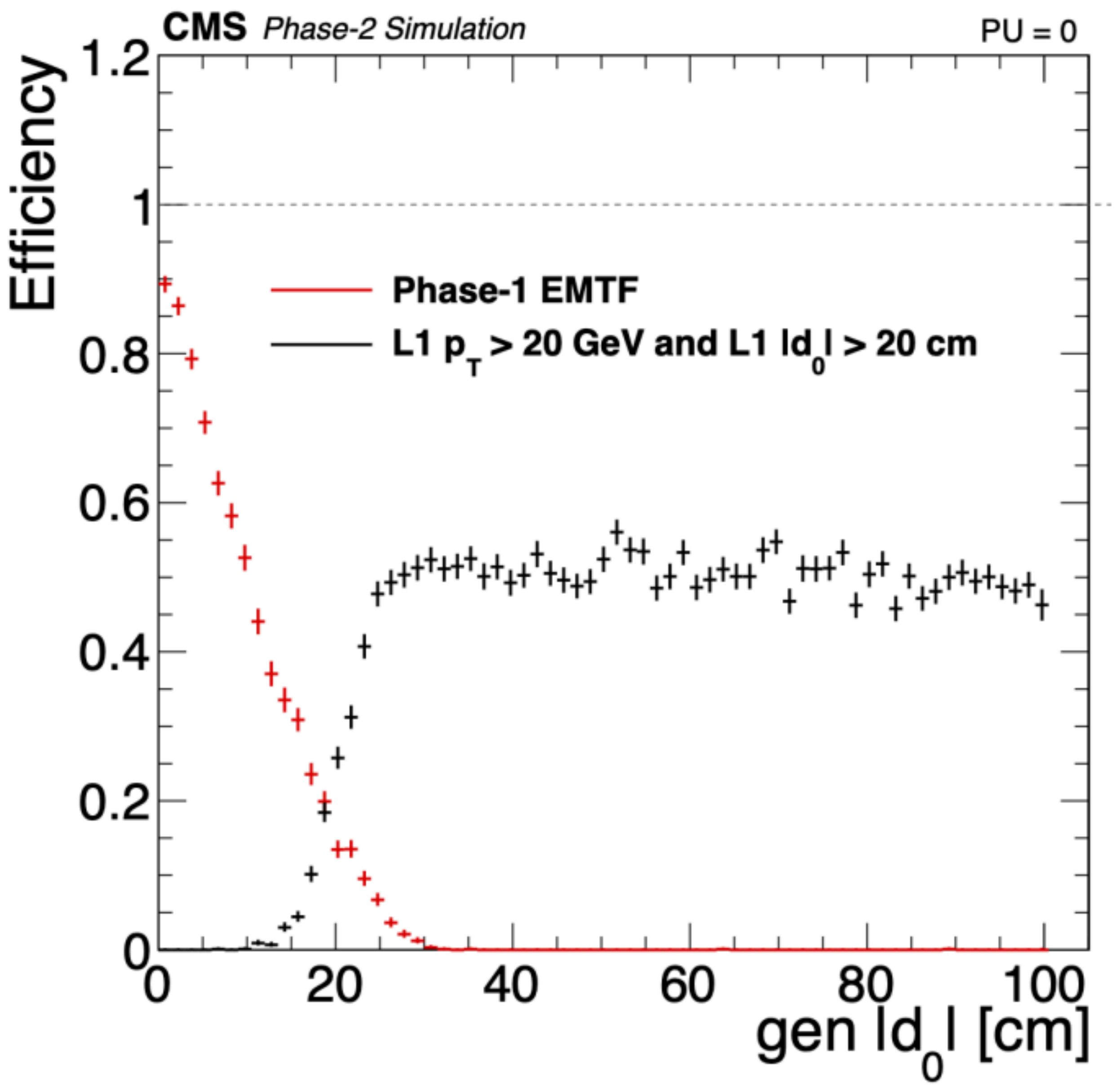
Figure taken from CMS Run 3 Detector Paper

EMTF NN Phase 2

- For the HL-LHC, CMS L1T will go through substantial changes.
  - Higher output rate (up to 750 kHz) and bigger latency budget (< 12.5 µs)
- In addition, the upgraded FPGAs will be much more powerful than what we have in Run 3
- The upgrade of EMTF algorithm (EMTF++), will have two NNs running in parallel: one for prompt muons and one for displaced muons.
  - With the new resources available, we will be able to implement much more elaborate NN based algorithms to improve upon the performance of the Run 3 trigger







Figures taken from CMS L1T Phase-2 TDR

Conclusions

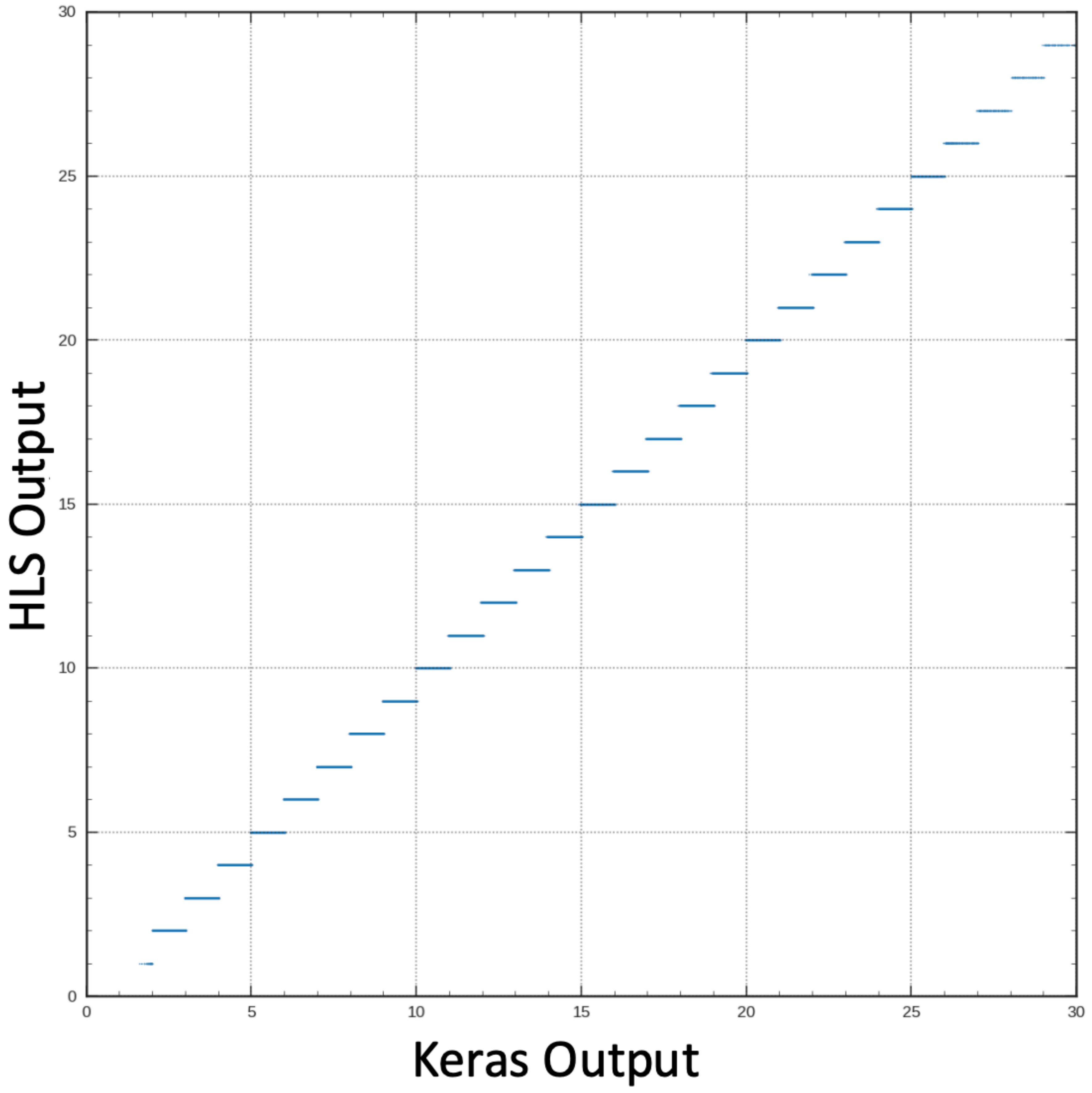
- EMTF is a good example of how ML algorithms can help CMS L1T system since Run 2.
  - Challenging conditions and unconventional signatures can benefit greatly from ML solutions
- We managed to deploy the first NN running in CMS L1T FPGAs for data taking, even with tight latency constraints and a nearly full chip.
  - Many lessons were learned on how to optimize NN models to fit into tight constraints while retaining performance
- ML based trigger algorithms deployed in Run 3 also provide valuable experience for CMS L1T at HL-LHC
  - There will be many more AI/ML algorithms in L1T for HL-LHC
    - Improving CMS capabilities for triggering on LLPs, anomaly detection, object reconstruction and jet tagging at L1T…

BACKUP

Keras vs HLS Outputs

HLS vs Keras Output – pT

CMS Simulation Preliminary



HLS vs Keras Output – Dxy

CMS Simulation Preliminary