Dec 8, 2019 • CPAD Instrumentation Frontier Workshop 2019

Machine Learning based algorithm for reconstructing prompt and displaced muons at Level-1 in CMS detector

Sergo Jindariani [1], Jia Fu Low [2]
on behalf of the CMS collaboration

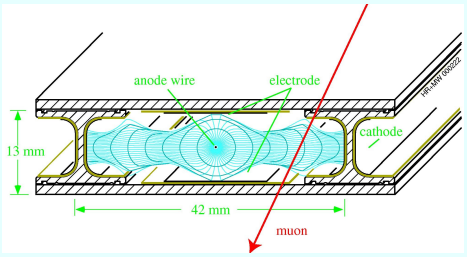
[1] *Fermilab*
[2] *University of Florida*

Overview

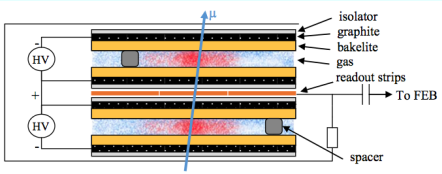
- Introduction
  - CMS muon system & L1 trigger system
  - Endcap Muon Track Finder (EMTF) at L1
  - Phase-2 EMTF objectives
- NN for prompt & displaced muon p_T assignment
- NN implementation in the FPGAs
- Summary

CMS Muon System

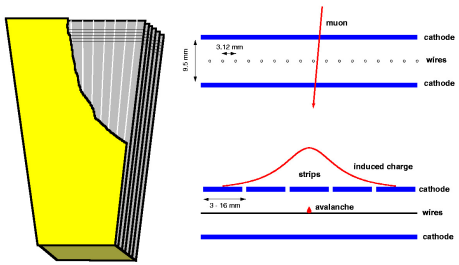
Drift Tubes



Resistive Plate Chambers



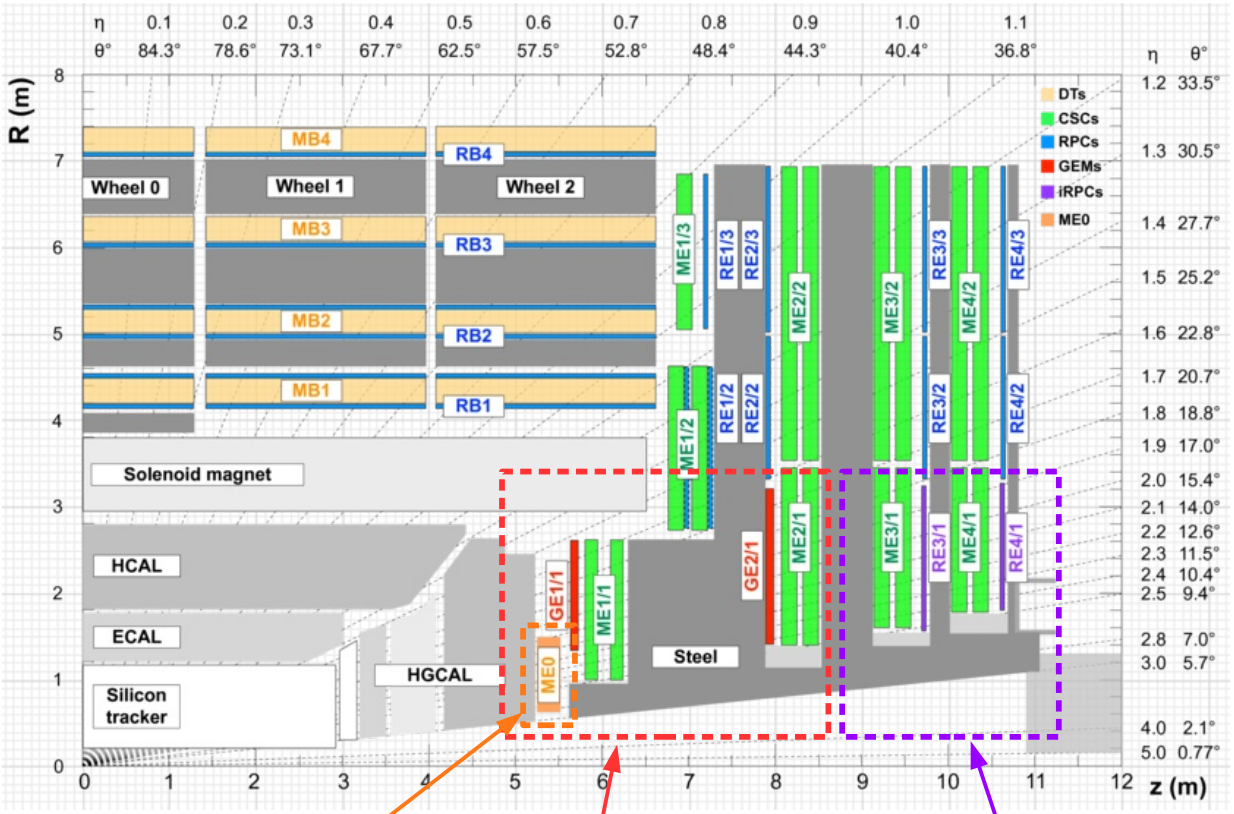
Cathode Strip Chambers



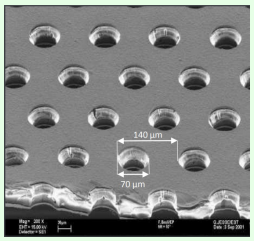
Muons are a crucial signature for many physics processes: Higgs, SUSY, etc. The **CMS Muon System** consists of 1846 muon chambers (DT, CSC, RPC) with ~1M channels to ensure robust trigger and reconstruction of muons.

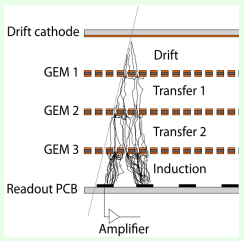
1846 = 250 (DT) + 540 (CSC) + 480 (RPCb) + 576 (RPCe)

Phase-2 Muon Upgrade



ME0 **GEM** (Gas Electron Multiplier) **iRPC** (improved RPC)





To maintain low trigger thresholds for muons at Phase-2 (200 PU), the forward muon system will be enhanced with new muon detectors (GEM, iRPC, ME0).

Some electronics of the existing muon detectors will also be replaced.

Pileup (PU) = in-time, inelastic collisions per bunch crossing.

108 (GEM) + 72 (iRPC) + 36 (ME0)

L1 Muon Trigger

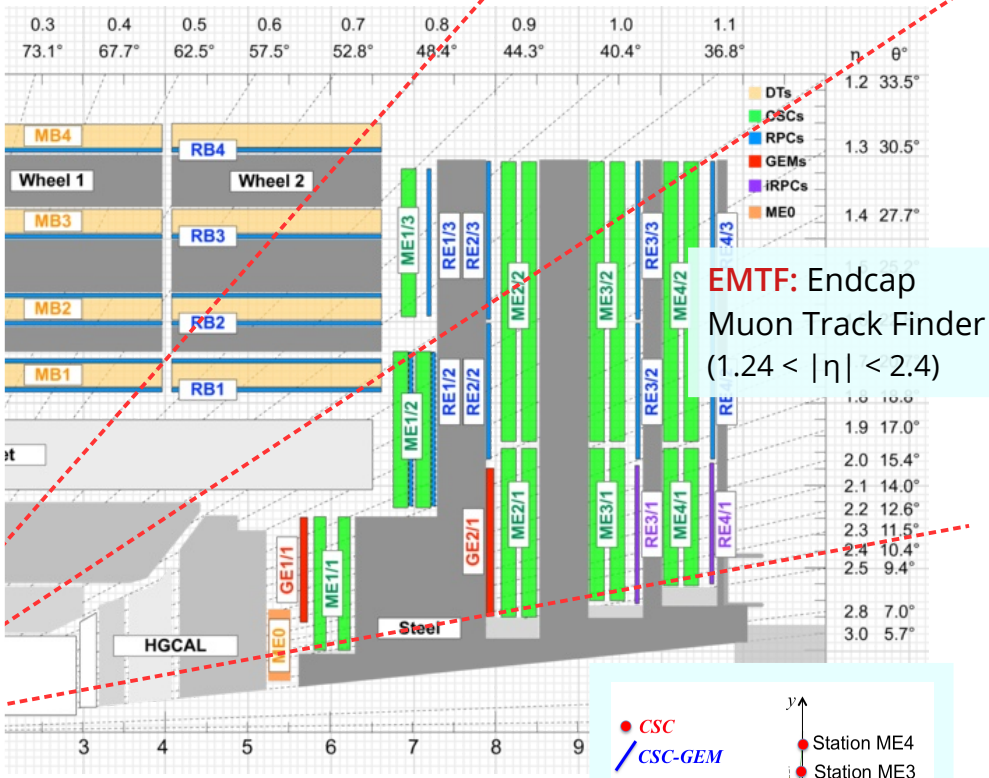
- CMS two-level trigger system:
  - **Level-1 (L1)**: custom electronics (e.g. FPGAs) used to reduce event rate of 40 MHz to 100 kHz within 4 μs latency.
  - **High Level Trigger (HLT)**: large CPU farm used to run software algorithms to further reduce event rate to 1 kHz.
- **L1 muon trigger system**:
  - Muon detectors have dedicated electronics that create the Trigger Primitives (TPs).
  - The TPs are collected and sent to the regional **Track Finders** (barrel, overlap, endcap) which build muons and measure their transverse momenta p_T.
  - The muons are sent to the Global Muon Trigger (GMT), and then to the Global Trigger (GT), which makes the "L1 Accept" trigger decision (typically $p_T > 22$ GeV for muons).
- L1 muon trigger algorithms need to be executed **very quickly and efficiently**.
- For Phase-2, the upgraded L1 trigger will have increased bandwidth (750 kHz) and longer latency (12.5 μs).
  - There will be a "correlator" layer that correlates muons and tracker tracks. Tracker tracks will be available at the L1 for the first time in Phase-2, and will have much better efficiency and p_T resolution. An interesting topic but won't be covered in this talk

Phase-2 objectives

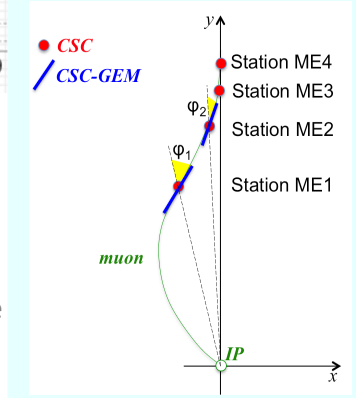
- Challenges:
  - Highly **non-uniform magnetic field** with very little magnetic bending in the very forward region.
  - **Large background** from low p_T muons, punch-throughs, neutrons, etc that could lead to non-linear PU dependence.
- For Phase-2, EMTF has to evolve to
  - Incorporate the **new muon detectors**.
  - Improve **efficiency, redundancy, p_T resolution, timing**.
  - Maintain the same trigger threshold at a reasonable **rate** at 200 PU to remain sensitive to electroweak scale physics.
- Phase-2 algorithm is named **EMTF++**

BMTF: Barrel Muon Track Finder ($0 < |\eta| < 0.83$)

OMTF: Overlap Muon Track Finder ($0.83 < |\eta| < 1.24$)



EMTF: Endcap Muon Track Finder ($1.24 < |\eta| < 2.4$)



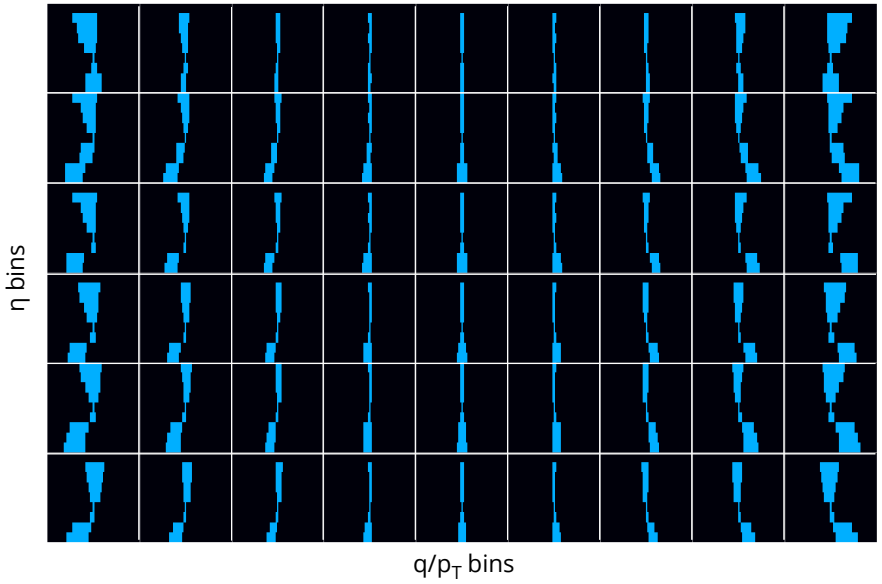
GEM-CSC bending angle improves p_T resolution and reduce trigger rate.

EMTF++ algorithm

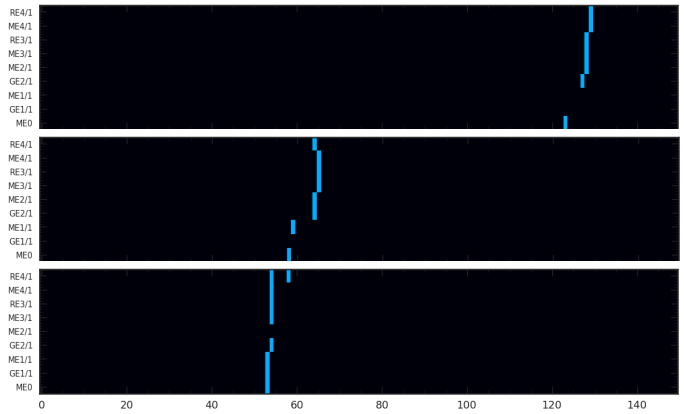
- We receive input trigger primitives from: ■ **CSC,** ■ **RPC,** ■ **GEM,** ■ **iRPC,** ■ **ME0.**
- **Pattern recognition**
  - Use patterns to find stubs in different muon stations that are consistent with muons. Different pattern shapes are used in different p_T bins.
- **Track building**
  - Build track candidates by selecting a unique stub from each muon station. If there are ambiguities, the stubs are ranked by $\Delta\phi$ & $\Delta\theta$ compatibility.
- **p_T assignment**
  - Use the machine learning algorithm (e.g. BDT, NN) to determine the p_T using multiple discriminating variables from the track candidate: $\Delta\phi$'s, $\Delta\theta$'s, bend, η, etc.
  - In Phase-1, the BDT algorithm is used. It is implemented with a large LUT on the FPGA. The input variables are encoded as an 30-bit address, which is used to retrieve the p_T value stored in the LUT. For Phase-2, the LUT address space will be increased to 37-bit.
  - As an alternative, we are investigating **running NN directly on the FPGA** for Phase-2. Focus of this talk

EMTF++ patterns

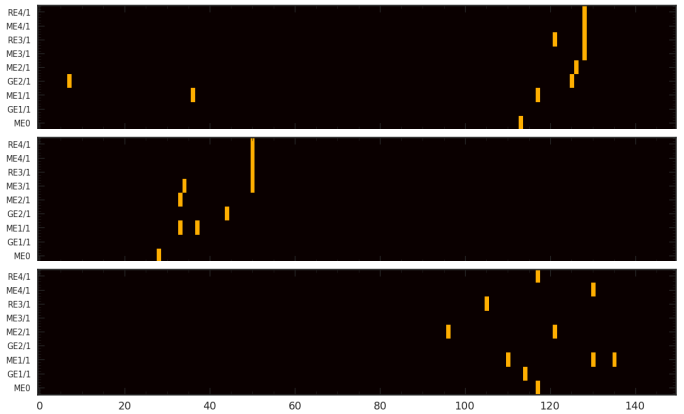
- Muon bending in the Endcap has strong (p_T, η)-dependence.
  - Use 9 bins in q/p_T, 6 bins in η
  - Pattern is the (ϕ,z)-view. From bottom to top: ϕ at innermost station to ϕ at outermost station. ϕ in 0.5° unit.
- Patterns are used to detect if the stubs are consistent with muon bending.



MC particle gun (one muon per event)



MC 200 PU events (pileup only)



60° EMTF sector in the η region of $2<|\eta|<2.16$. Muon moving outward from bottom to top.

p_T assignment with NN

- Extract input from the stubs associated to the track candidates: **ϕ, θ, bend, quality, time**.

	ϕ			θ			bend		
	num of bits	range	resolution	num of bits	range	resolution	num of bits	range	resolution
CSC	**11**	[0,2047]	1/16-strip	**7**	[0,111]	wiregroup	**5**	[-16,15]	1/16-strip
RPC	**5**	[0,31]	strip	**2**	[0,2]	roll	-		
iRPC	**7**	[0,95]	strip	**7 (?)**	[0,82] (?)	position along strip	-		
GEM	**8**	[0,191]	2-strip	**3**	[0,7]	roll	-		
ME0	**10**	[0,767]	1/2-strip	**4**	[0,15]	1/2-roll	**10**	[-512,511]	1/4-strip
DT	**12**	[-2048,2047]	1/4096 rad	**3**	[0,6]	bti group	**10**	[-512,511]	1/512 rad

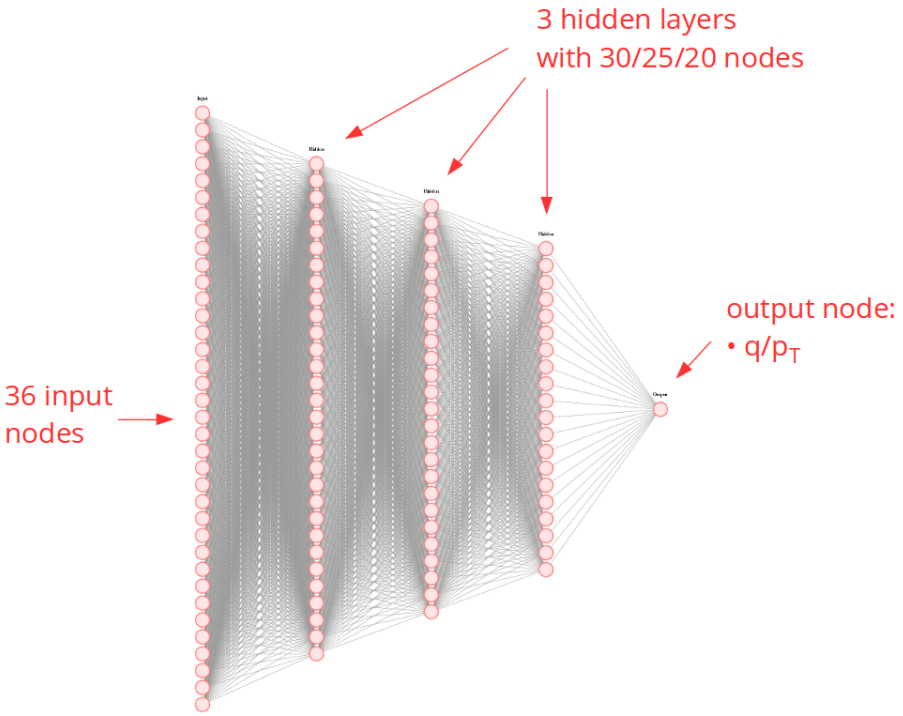
	quality			time					
	num of bits	range	definition	num of bits	range	resolution	num of bits	range	resolution
CSC	**2**	[3,6]	# of layers	-					
RPC	**3**	[1,8]	cluster width	**4**	[0-15]	1/16-BX			
iRPC	**3**	[1,8]	cluster width	**4**	[0-15]	1/16-BX			
GEM	**3**	[1,8]	cluster width	-					
ME0	**4**	[4,6]	# of layers	-					
DT	**3**	[0,6]	quality	-					

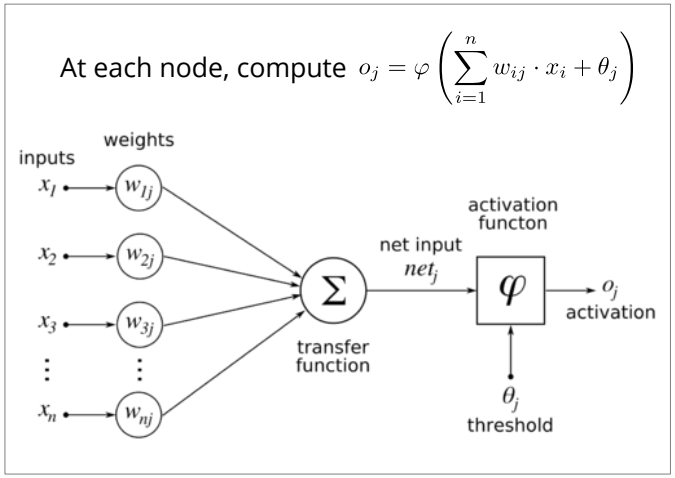
- At the moment, consider **36 features**
  - Note: allocate 12 stations, although a muon can go through at most 8-10 stations depending on η

	ME1/1	ME1/2	ME2	ME3	ME4	RE1	RE2	RE3	RE4	GE1/1	GE2/1	ME0
ϕ	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
θ	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
bend	✔	✔	✔	✔	✔							✔
quality	✔	✔	✔	✔	✔							✔
time												

→ **Regression NN**

p_T assignment with NN





Can be done on the FPGA!

ML framework: K Keras

Loss function: Huber loss [Wikipedia]

Activation function: ReLU

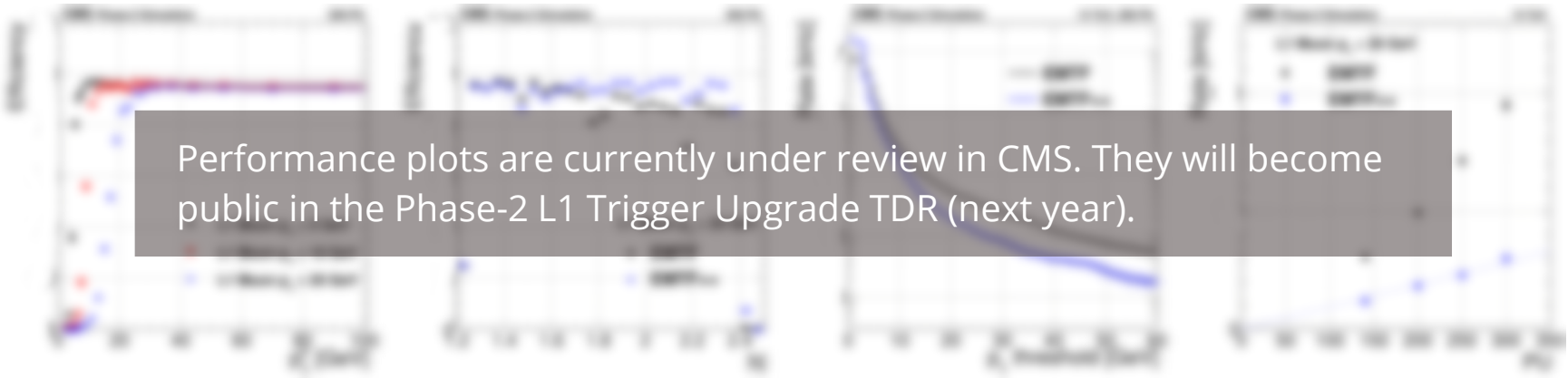
Batch normalization: applied right after the input layer and in each hidden layer

Training dataset: 2M muons

Testing dataset: 1M muons

NN performance

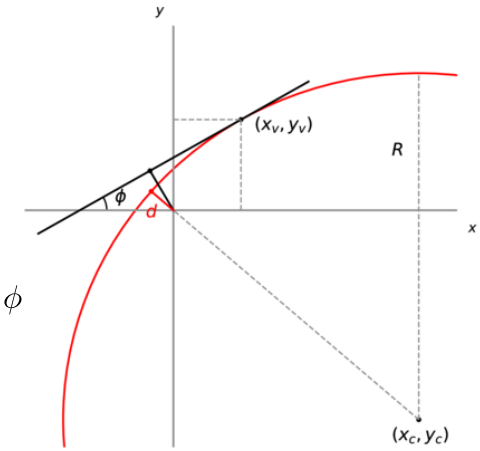
- From simulation studies, NN has been shown to **improve the efficiency and p_T resolution**, and **reduce the trigger rate**
  - NN allows us to easily add info from the additional muon detectors
  - Trigger rate around 20 kHz for L1 p_T > 20 GeV @ 200 PU.
  - Trigger rate linear up to 300 PU!



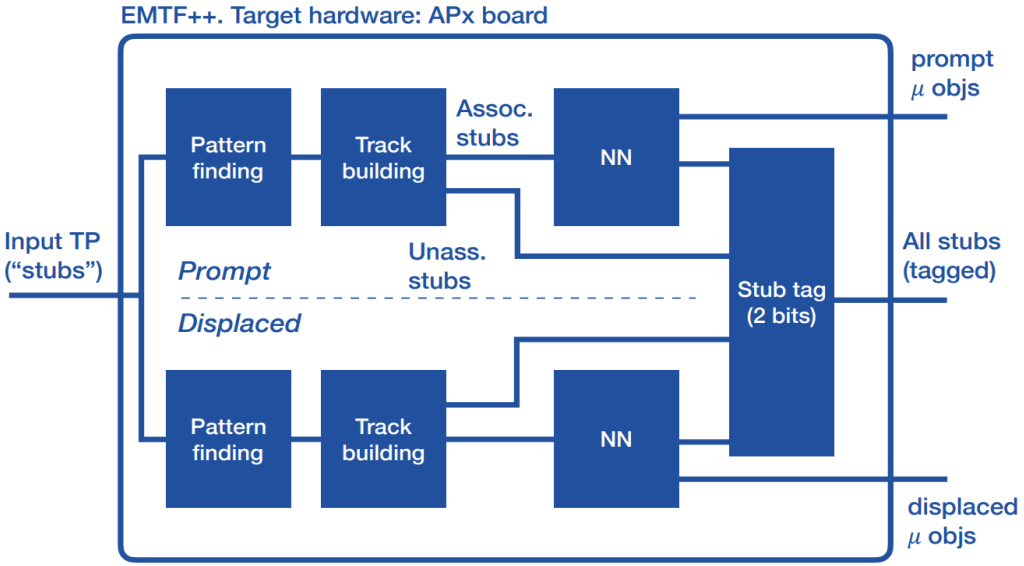
Performance plots are currently under review in CMS. They will become public in the Phase-2 L1 Trigger Upgrade TDR (next year).

Displaced muons

- There is an emerging strong interest in BSM models that involve long-lived particles that can decay into muons (displaced muons).
- The barrel counterpart BMTF has implemented the Kalman Filter (now called KBMTF) which allows them to trigger for both prompt and displaced muons.
  - KBMTF starts the propagation from the outermost station to the innermost. At the end of the propagation, they can decide to add the vertex constraint (prompt) or not (displaced).
  - Prepared for use in Run 3 (starting 2021).
- We wanted to see if we could also trigger for displaced muons in the Endcap using machine learning.
  - Need to find the **vertex-unconstrained p_T** and the **impact parameter d_0**.
  - d_0 is defined as the distance of closest approach of the track w.r.t the positive z-axis. We use the sign convention such that $d_0 = x_v \sin\phi - y_v \cos\phi$ as $p_T \to \infty$.



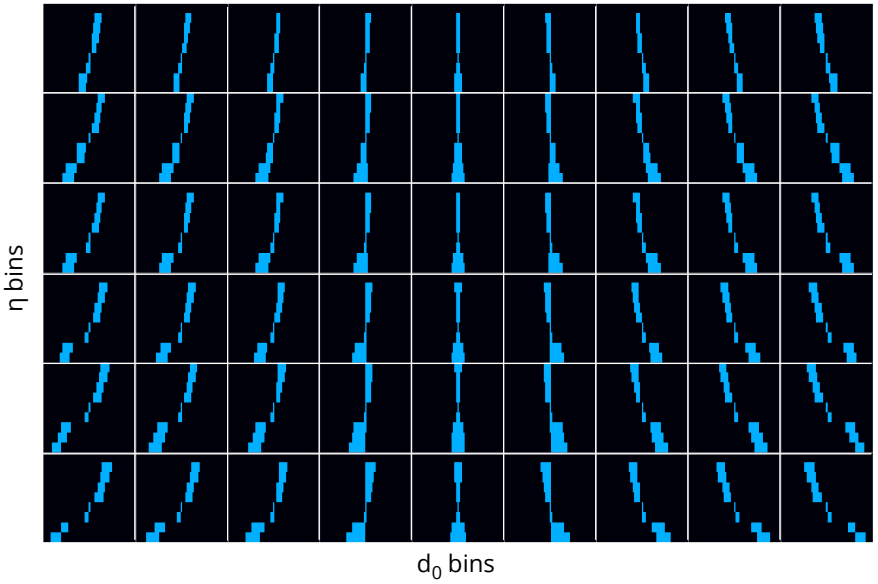
Displaced muons



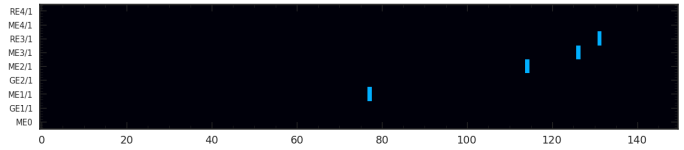
- Current prompt muon algorithm has acceptance for moderately displaced muons (efficiency drops to 0 at d_0 ~ 20 cm)
  - We need to add new displaced patterns to improve the acceptance.
  - And train a new NN for the displaced p_T and d_0 assignments.
- For Phase-2, the simplest plan is to do separate reconstructions for prompt and displaced muons (doubling the num of patterns and NNs)
- For Run-3, due to limited firmware resources, we plan to only add the NN into the current EMTF firmware. More about this later

Displaced EMTF++: patterns

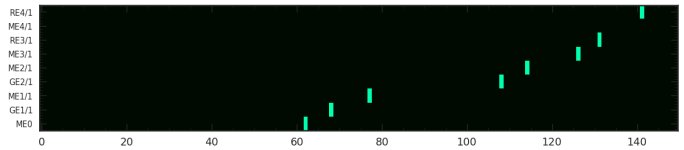
- We generate patterns for high p_T displaced muons
  - Use 9 bins in d_0, 6 bins in η. Require $p_T > 14$ GeV. The d_0 range is -120 to 120 cm.
  - Pattern is the (ϕ,z)-view. From bottom to top: ϕ at innermost station to ϕ at outermost station. ϕ in 0.5° unit.
- New patterns improve acceptance to about 90% in the low η region, but worse for large d_0 and high η muons.
  - Some inefficiency due to TP reconstruction due to large incidence angle
- These patterns are useful for Phase-2. But for Run-3, we still need to figure out how to modify the patterns in the current EMTF firmware.



Trigger primitives for a highly displaced muon



GEANT hits from the same muon

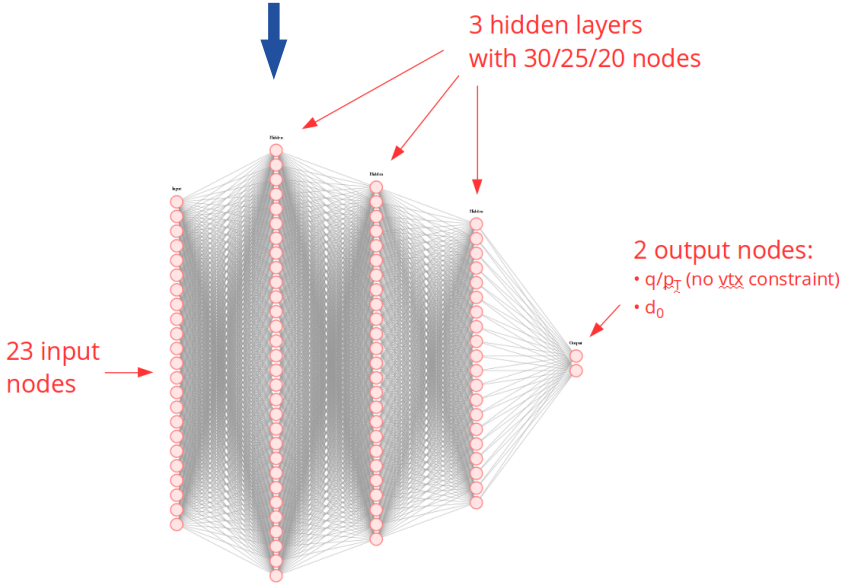


60° EMTF sector in the η region of $2<|\eta|<2.16$. Muon moving outward from bottom to top.

Displaced EMTF++: NN

- In order to deploy during Run-3, displaced NN is trained with only inputs that are available in the current EMTF firmware.
  - Reduced to **23 features** (CSC/RPC only)
    - CSC ϕ and bend are the not the improved version as used for Phase-2 prompt NN.
  - 4 stations → 6 possible pairs: **1-2, 1-3, 1-4, 2-3, 2-4, 3-4**
    - RPC is subbed in if CSC is not found in a given station/chamber.

Δϕ						Δθ						Bend				is RPC (1 bit)				Ring (1 bit)				F/R (1 bit)				Track
1-2	1-3	1-4	2-3	2-4	3-4	1-2	1-3	1-4	2-3	2-4	3-4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	θ
✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔				✔				✔



2 regression outputs:
- q/p_T (without vertex constraint)
- d_0

Loss function is the combination:

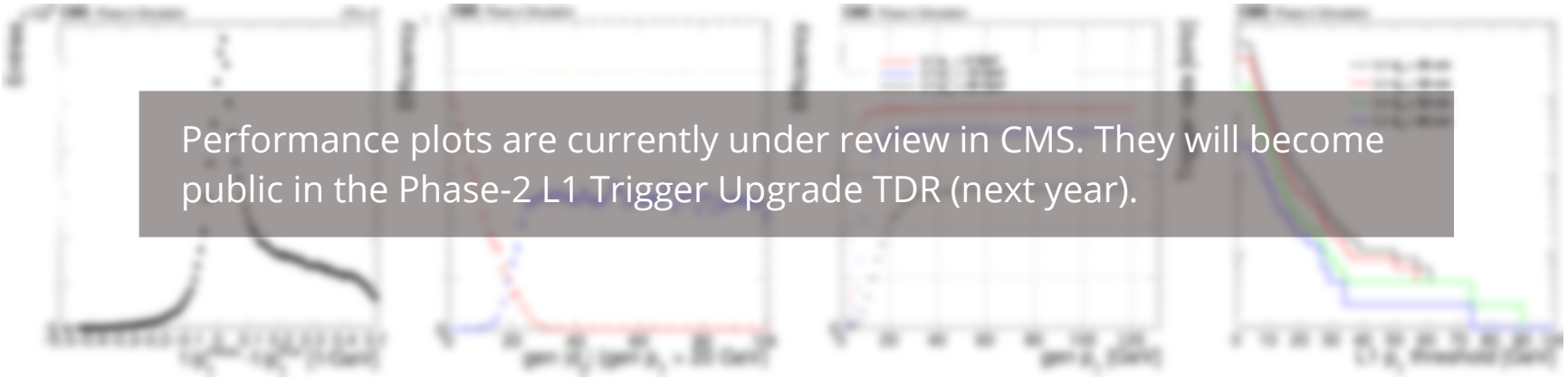
$$\ell_{\text{total}} = \ell_{q/p_{\text{T}}} + \alpha \cdot \ell_{d_0}$$

where α is an adjustable scale factor

Training muons: $p_T > 2$ GeV, $|d_0| < 120$ cm

Displaced EMTF++: NN performance

- We decided to use **L1 p_T > 20 GeV** and **L1 d_0 > 20 cm** as the working point in the Endcap.
  - This gives us 40-50% efficiency for displaced muons with d_0 from 30 to 100 cm, when averaged over η from 1.2 to 2.4 (there is a strong η dependence).
  - Trigger rate of O(10) kHz at 200 PU.



Performance plots are currently under review in CMS. They will become public in the Phase-2 L1 Trigger Upgrade TDR (next year).

- **What we have learned:**

Still work in progress

  - p_T resolution for displaced (60%) is much worse compared to prompt (20%) due to losing the vertex constraint.
  - Large tail in the p_T resolution where p_T is under-measured for high-p_T muons. This leads to low efficiency at the plateau even with L1 p_T > 20 GeV cut.
  - d_0 resolution is quite good — about 5 cm.
  - Performance has strong η dependence. Efficiency for large d_0 muons at high η is basically zero.

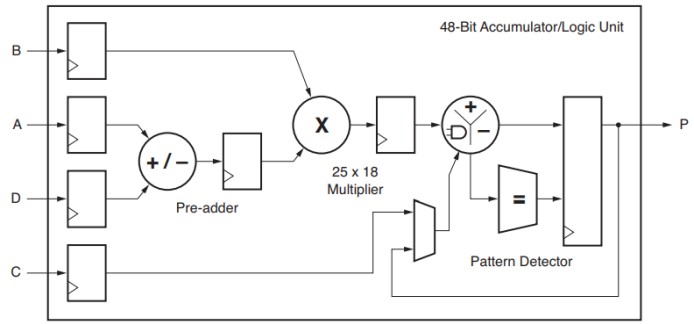
Firmware implementation

- hls4ml

See talk by Sergo

  - A toolkit to implement fast neural network inferences in FPGAs using Vivado High-Level Synthesis (HLS).
  - Can convert NN models from popular ML libraries (Keras, Tensorflow, etc) into VHDL codes, which can be used to generate the firmware.
    - Optimize usage of DSPs in the modern FPGAs for multiply-accumulate operations
  - See arxiv:1804.06913

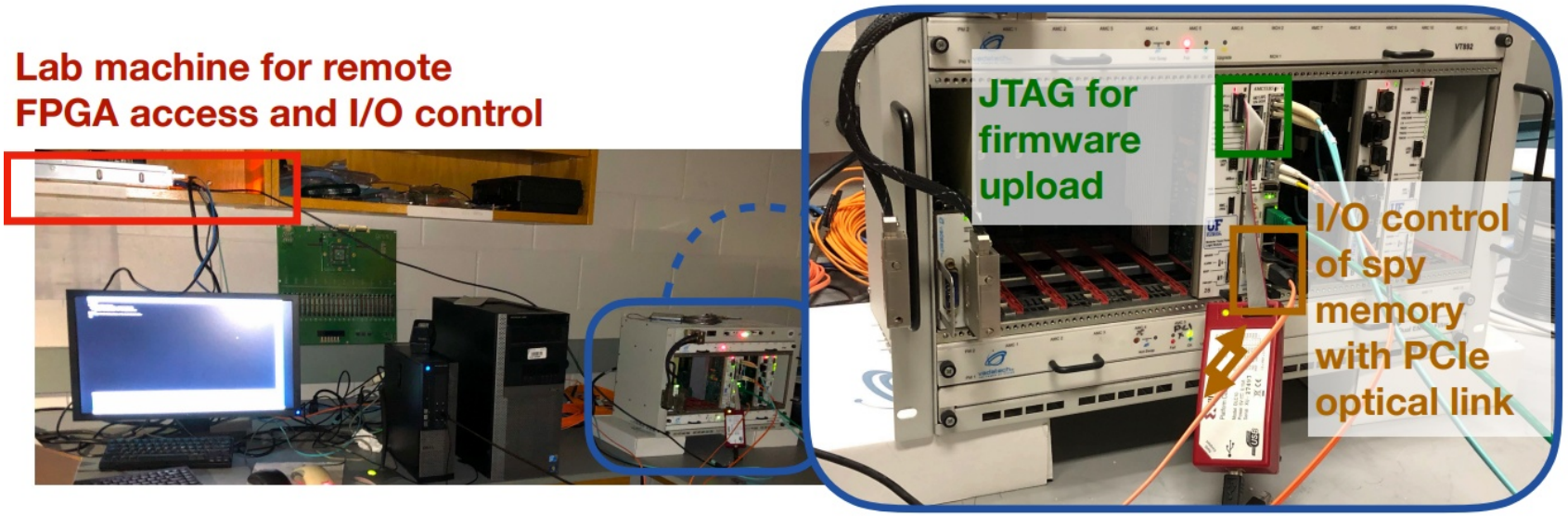
Basic DSP48E1 Slice functionality



Device	# of DSPs
Kintex-7 325T	840
Virtex-7 690T	3600
Kintex UltraScale KU115	5500
Virtex UltraScale+ VU9P	6800

The hardware: MTF7

- The current EMTF firmware runs on **MTF7 board** [1,2]
  - Virtex-7 690T-2 FPGA which has 3,600 DSP's.
  - Largest logic resource usage: LUT and BRAM. Only 1.2% DSP's are used.
- MTF7 setup at UF for testing firmware:
  - PCIe communication for large-scale tests and debug.



[1] CMS Collaboration, "CMS Technical Design Report for the Level-1 Trigger Upgrade", CMS-TDR-012
[2] D. Acosta et al., "The CMS Modular Track Finder boards, MTF6 and MTF7", Journal of Instrumentation 8 (2013) C12034

NN resource usage

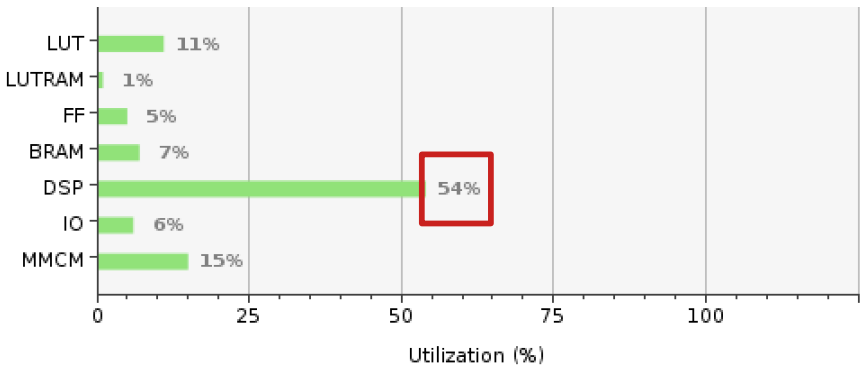
HLS estimates

Name	BRAM_18K	DSP48E	FF	LUT
DSP	-	-	-	-
Expression	-	-	0	6
FIFO	-	-	-	-
Instance	39	2017	220117	83726
Memory	-	-	-	-
Multiplexer	-	-	-	1404
Register	0	-	3486	32
Total	39	2017	223603	85168
Available	2940	3600	866400	433200
Utilization (%)	1	56	25	19

```
+ Timing (ns):
    * Summary:
    +--------+-------+----------+------------+
    |  Clock | Target| Estimated| Uncertainty|
    +--------+-------+----------+------------+
    |ap_clk  |   4.00|     3.492|        0.50|
    +--------+-------+----------+------------+

+ Latency (clock cycles):
    * Summary:
    +-----+-----+-----+-----+----------+
    |  Latency  |  Interval | Pipeline |
    | min | max | min | max |   Type   |
    +-----+-----+-----+-----+----------+
    |   48|   48|    1|    1| function |
    +-----+-----+-----+-----+----------+
```

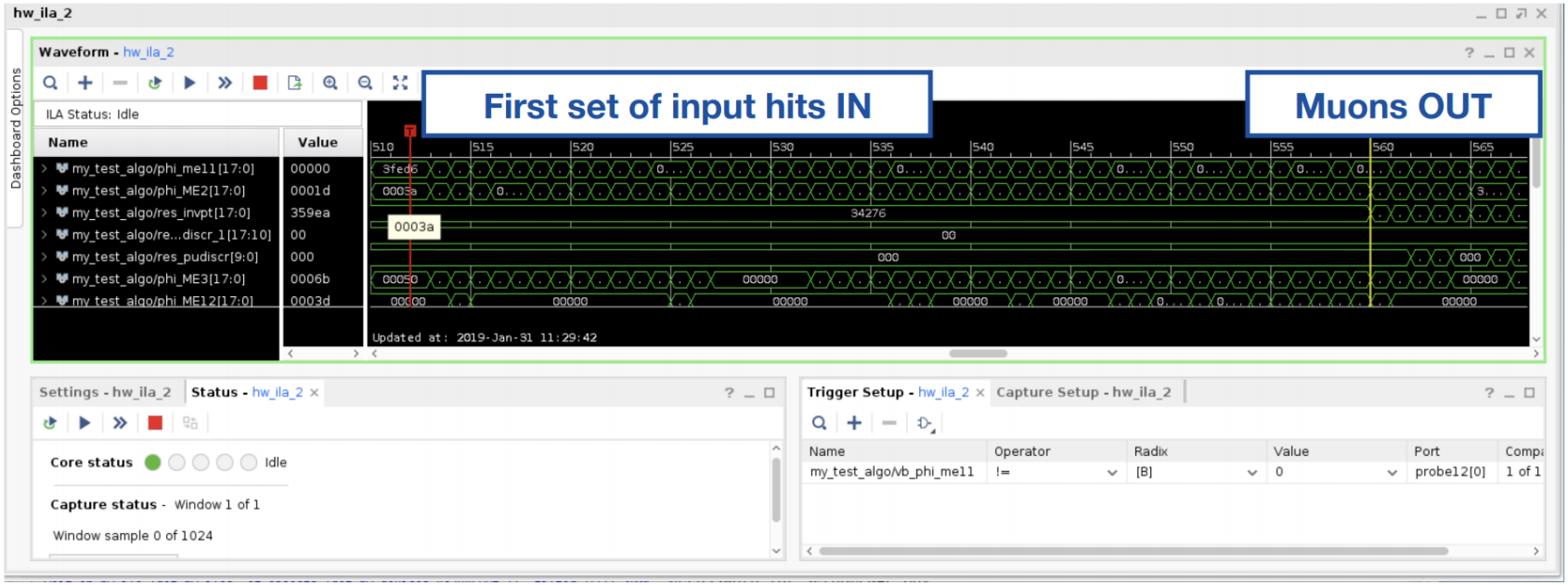
Implementation



HLS resources estimate is accurate for DSP, conservative for FF and LUT

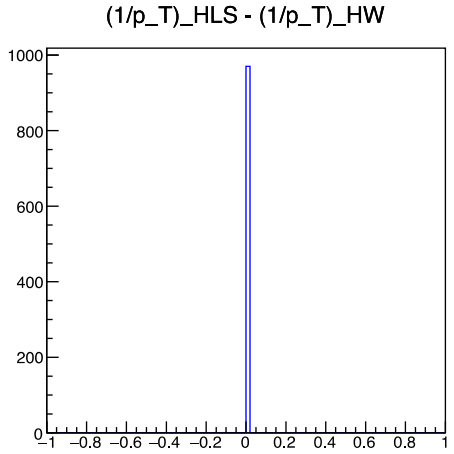
Latency estimate of 48 clk

Running NN on hardware



Running at 200 MHz. Verified latency of 48 clk as given by HLS estimate. So, it takes 240 ns to get the p_T & d_0 of the first muon.

Validated HLS outputs with HW outputs for 1k muons.

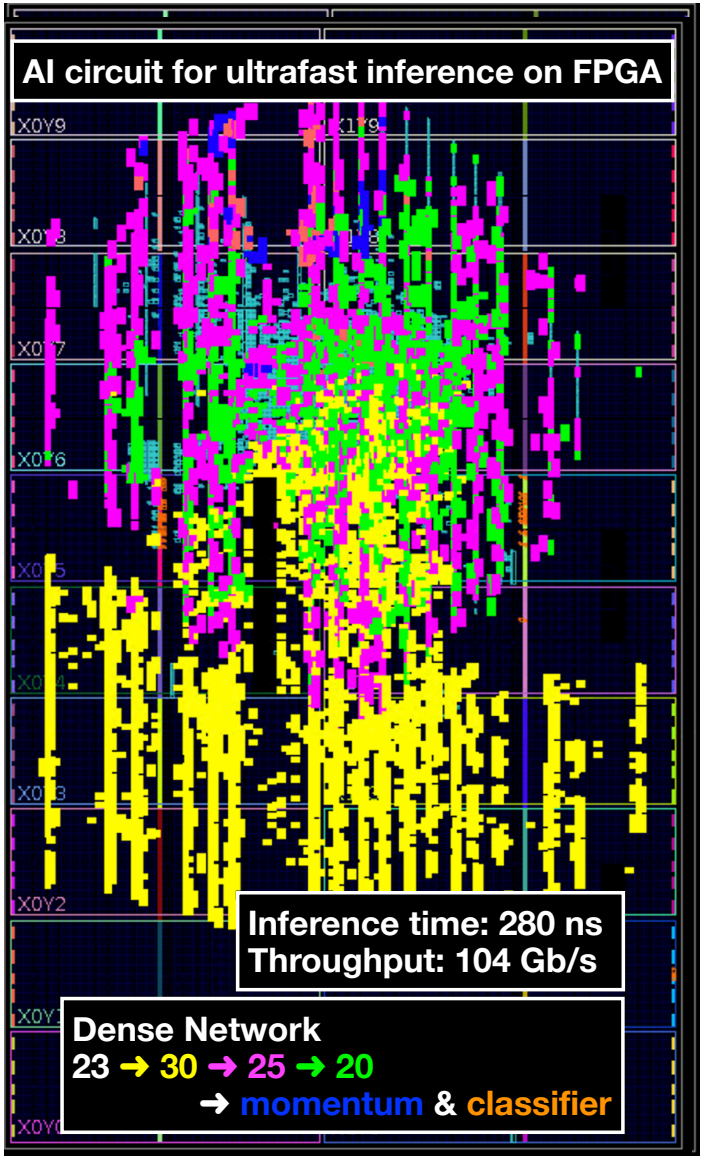


NN resource usage

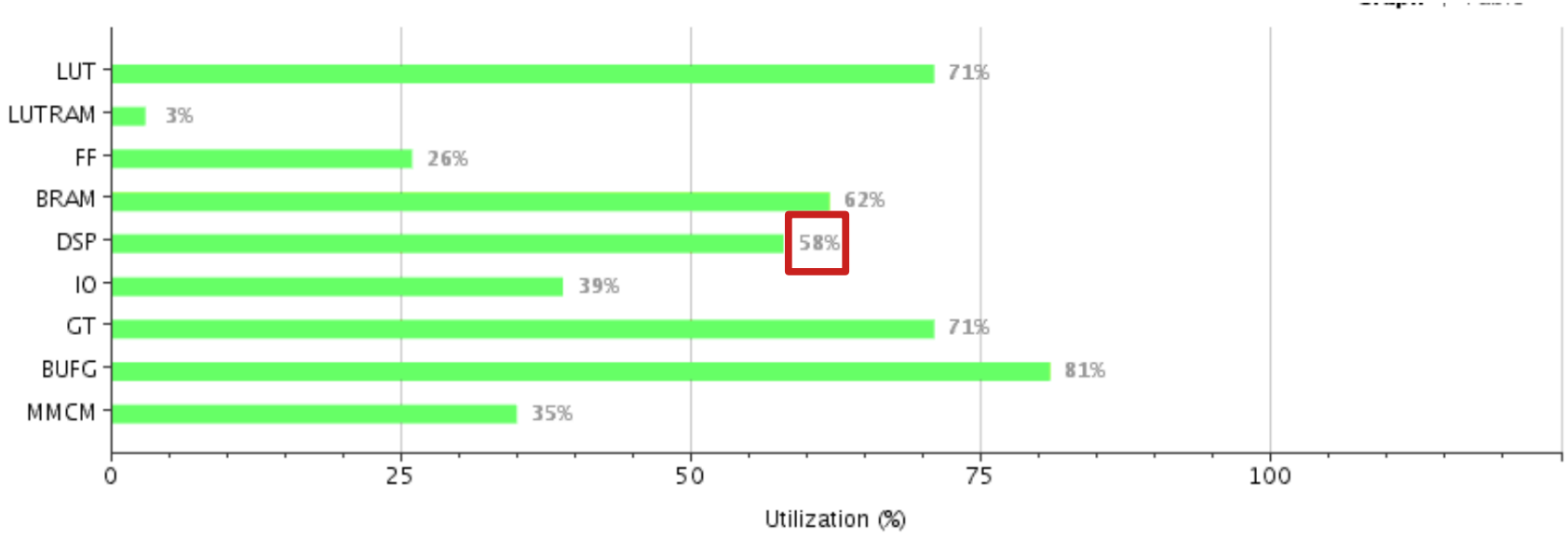
The silicon neural network

- hidden layer #1
- hidden layer #2
- hidden layer #3

(recently updated firmware with 250 MHz freq and 70 clk latency)



Adding NN into the EMTF FW



- Synthesis of the **current EMTF firmware + NN** in the Virtex-7 FPGA
- They fit in the same FPGA! Nice complementarity in terms of resource usage for EMTF & NN. (NN has taken over the unused DSP's)
- Possible for use as early as Run 3?

Estimates for Phase-2 FPGA (VU9P)

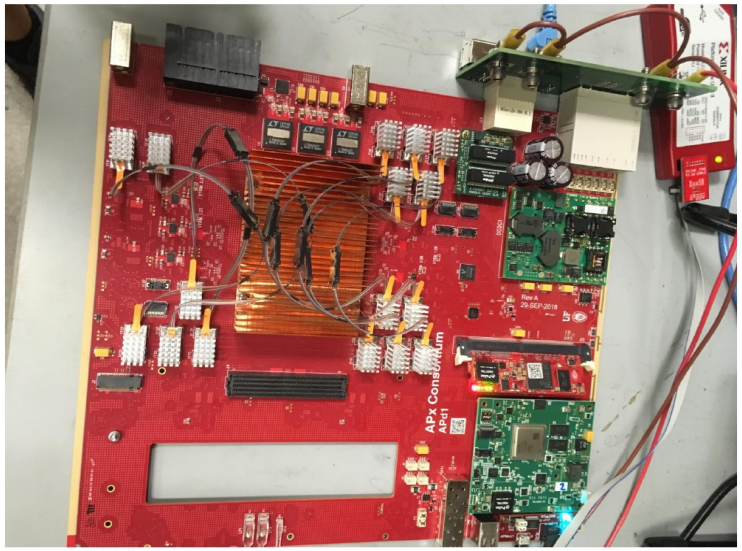
HLS estimates

```
== Utilization Estimates
* Summary:
+-----------------+---------+-------+---------+---------+-----+
|       Name      | BRAM_18K| DSP48E|    FF   |   LUT   | URAM|
+-----------------+---------+-------+---------+---------+-----+
|DSP              |        -|      -|        -|        -|    -|
|Expression       |        -|      -|        0|        6|    -|
|FIFO             |        -|      -|        -|        -|    -|
|Instance         |       39|   2420|    69109|    90580|    -|
|Memory           |        -|      -|        -|        -|    -|
|Multiplexer      |        -|      -|        -|     1404|    -|
|Register         |        0|      -|     4280|       32|    -|
+-----------------+---------+-------+---------+---------+-----+
|Total            |       39|   2420|    73389|    92022|    0|
+-----------------+---------+-------+---------+---------+-----+
|Available        |     4320|   6840| 2364480| 1182240|  960|
+-----------------+---------+-------+---------+---------+-----+
|Utilization (%)  |       ~0|     35|        3|        7|    0|
+-----------------+---------+-------+---------+---------+-----+
```

```
+ Timing (ns):
    * Summary:
    +--------+-------+----------+------------+
    |  Clock | Target| Estimated| Uncertainty|
    +--------+-------+----------+------------+
    |ap_clk  |   3.00|     2.456|        0.38|
    +--------+-------+----------+------------+

+ Latency (clock cycles):
    * Summary:
    +-----+-----+-----+-----+----------+
    |  Latency  |  Interval | Pipeline |
    | min | max | min | max |   Type   |
    +-----+-----+-----+-----+----------+
    |   32|   32|    1|    1| function |
    +-----+-----+-----+-----+----------+
```

APd board being developed



Looking into the Phase-2 APd board [3] with Virtex US+ VU9P FPGA, which has 3X more LUT & FF, and 2X more DSP.

NN should comfortably fit in the VU9P (DSP usage is 35%)

32 clk @ 333 MHz ≈ 100 ns latency

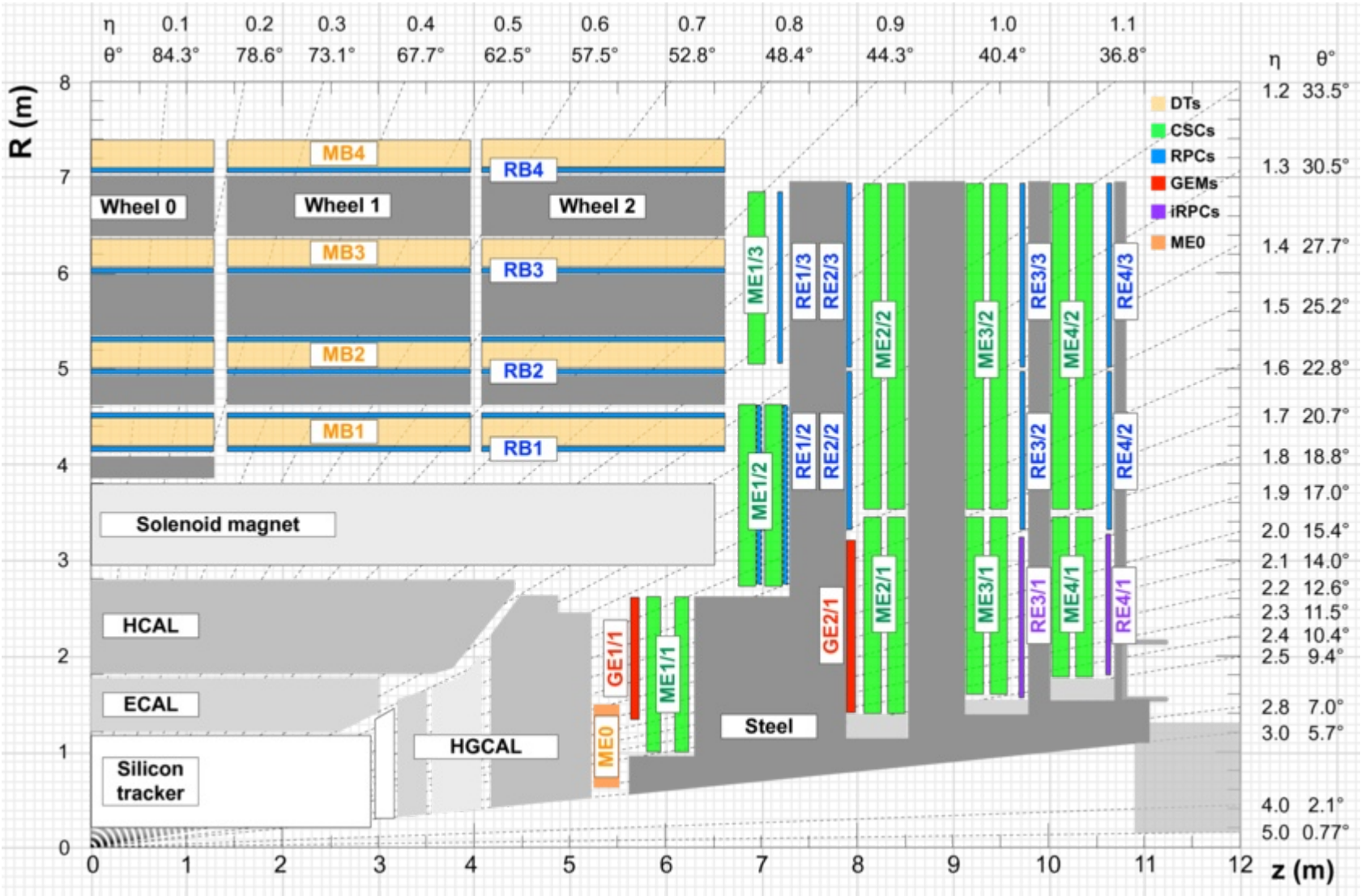
[3] CMS Collaboration, "The Phase-2 Upgrade of the CMS L1 Trigger Interim Technical Design Report", CERN-LHCC-2017-013, CMS-TDR-017

Summary

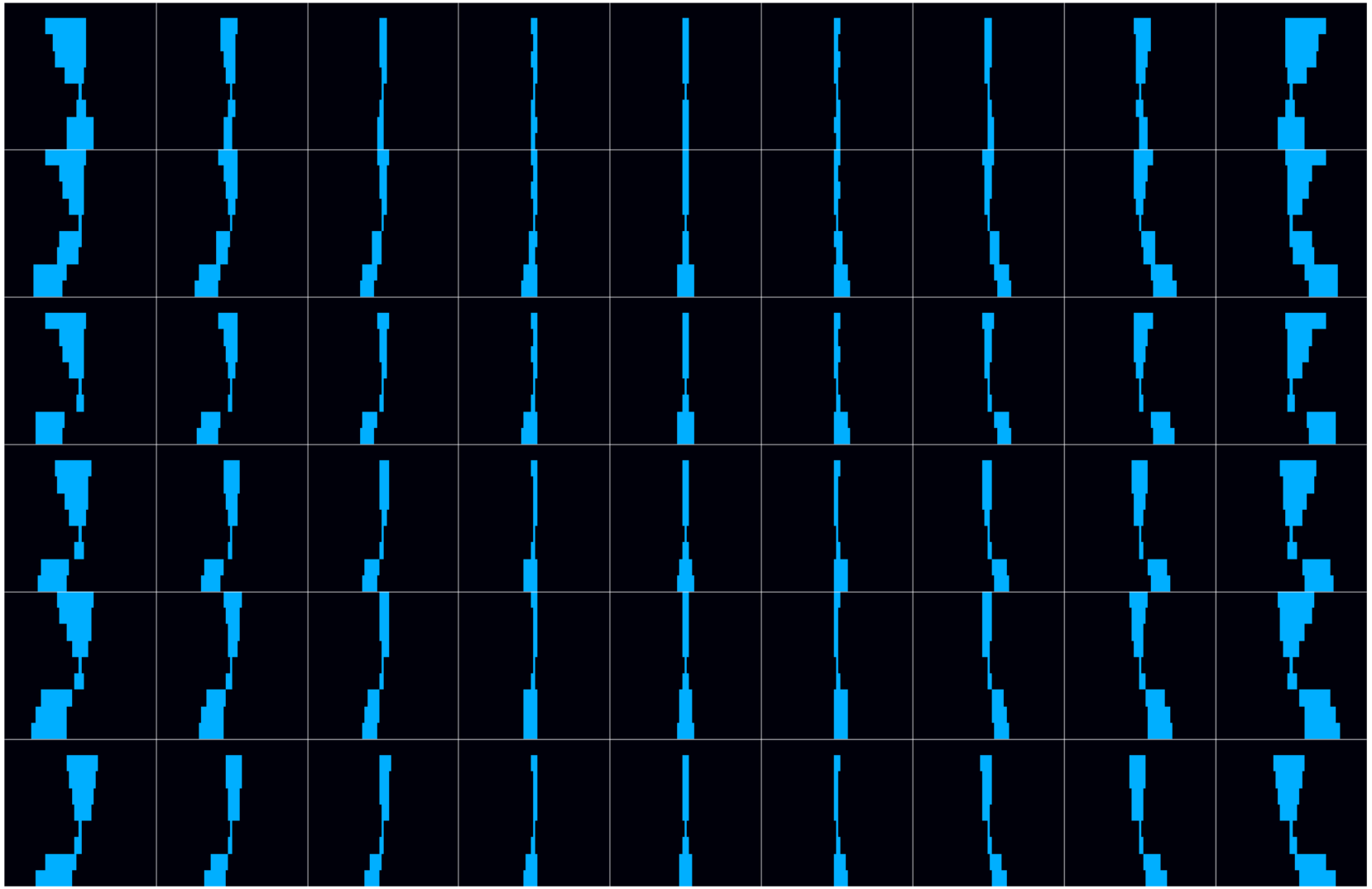
- NN has been used to show promising results for reconstructing prompt and displaced muons in the Endcap
  - which is a difficult region due to non-uniform magnetic field and large background rates.
  - We started experimenting with NN two years ago, and started our own study about displaced muons one year ago. We have made good progress in (i) learning to use the ML technology; (ii) using it to explore new phase space.
- NN implementation into FPGA has been demonstrated. Its highly parallel structure and ability to process large num of inputs make it a good choice for L1 trigger
  - We leveraged `hls4ml` to significantly reduce the firmware development time. We want to continue to study if the resource usage can be further optimized.

Backup

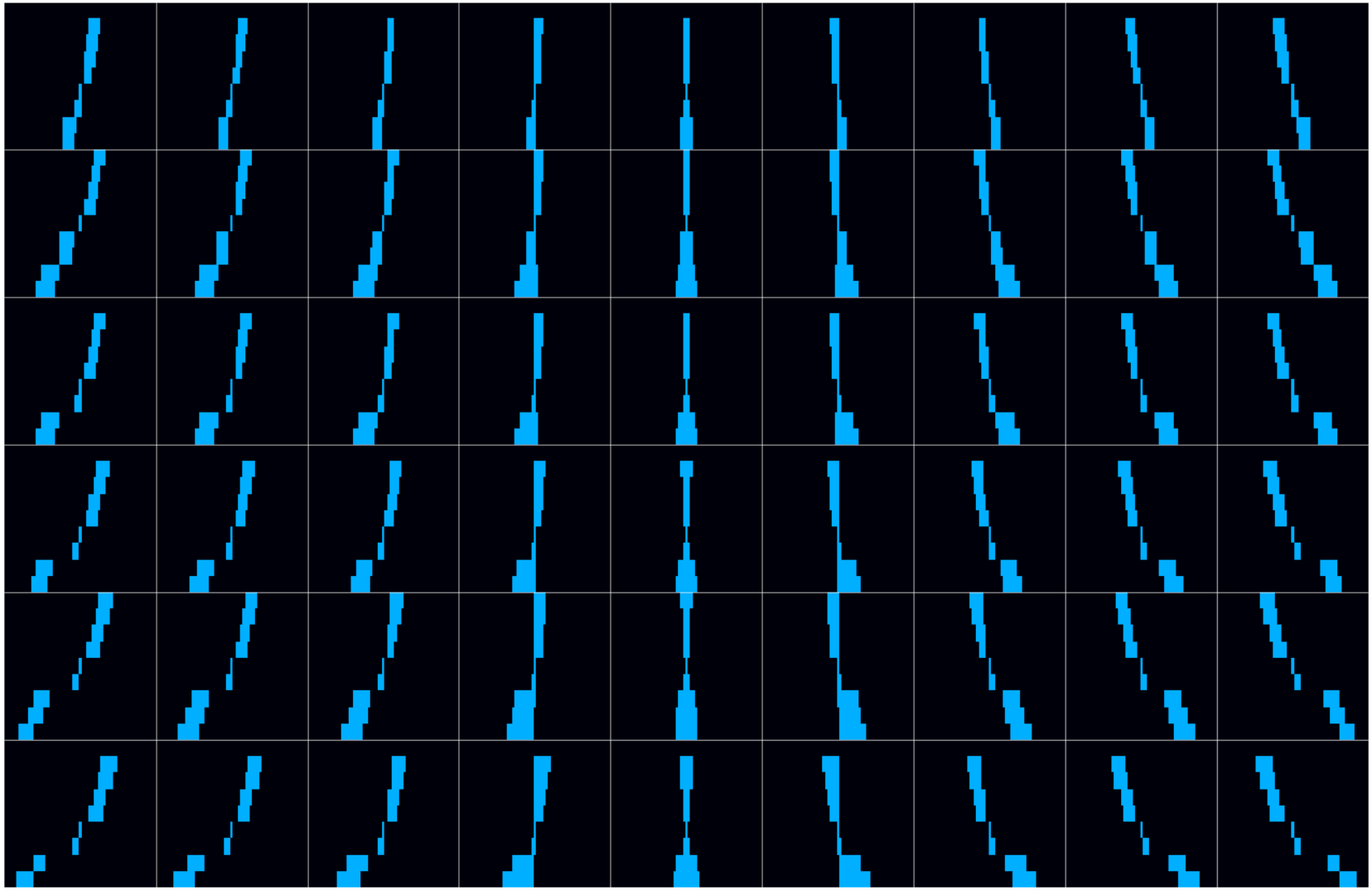
Phase-2 CMS detector quadrant



EMTF++ patterns (prompt)



EMTF++ patterns (displaced)



Input variables for BDT

Mode		Δϕ						Δϕ	Δθ						Bend + RPC				F/R				θ	Md	Bits
		1-2	1-3	1-4	2-3	2-4	3-4	sign	1-2	1-3	1-4	2-3	2-4	3-4	1	2	3	4	1	2	3	4			
15	1-2-3-4	7			5		4	2			2				2	1	1	1	1				3	1	30
14	1-2-3	7			5			1		3					2	1	1		1	1			5	3	30
13	1-2-4	7				5		1			3				2	1		1	1	1			5	3	30
11	1-3-4		7				5	1			3				2		1	1	1		1		5	3	30
7	2-3-4				7		5	1					3			2	1	1		1			5	4	30
12	1-2	7							3						3	3			1	1			5	7	30
10	1-3		7							3					3		3		1		1		5	7	30
9	1-4			7							3				3			3	1			1	5	7	30
6	2-3				7							3				3	3			1	1		5	7	30
5	2-4					7							3			3		3		1		1	5	7	30
3	3-4						7							3			3	3			1	1	5	7	30

Mode		Mode identifier in LUT address	Bits	Addresses
15	1-2-3-4	1XXX XXXX ...	1	2^29
14	1-2-3	011X XXXX ...	3	2^27
13	1-2-4	010X XXXX ...	3	2^27
11	1-3-4	001X XXXX ...	3	2^27
7	2-3-4	0001 XXXX ...	4	2^26
12	1-2	0000 111X ...	7	2^23
10	1-3	0000 110X ...	7	2^23
9	1-4	0000 101X ...	7	2^23
6	2-3	0000 100X ...	7	2^23
5	2-4	0000 011X ...	7	2^23
3	3-4	0000 010X ...	7	2^23

The inputs to the BDT must be compressed into the 30-bit address space. The compression scheme depends on the track “mode”.