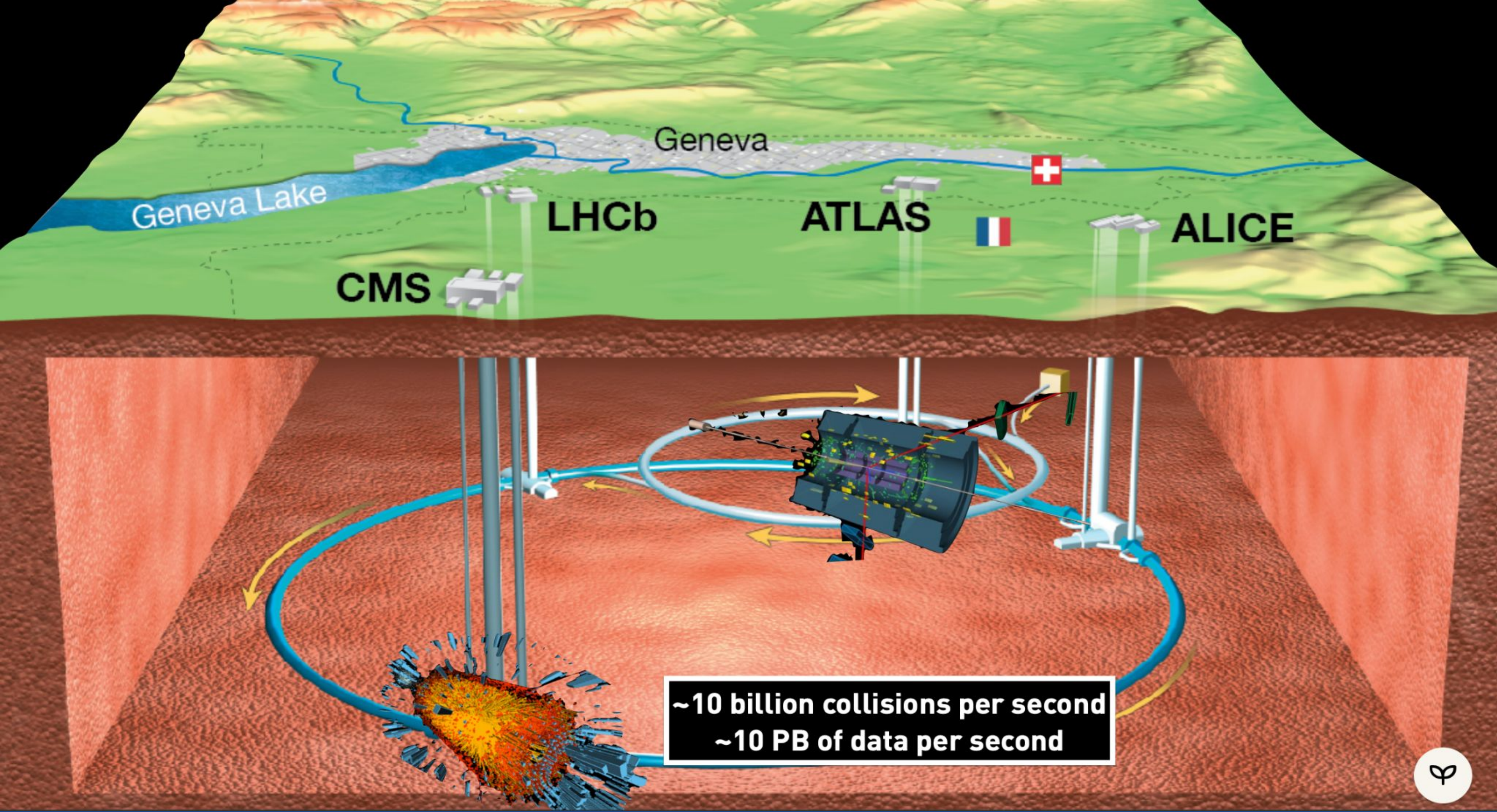
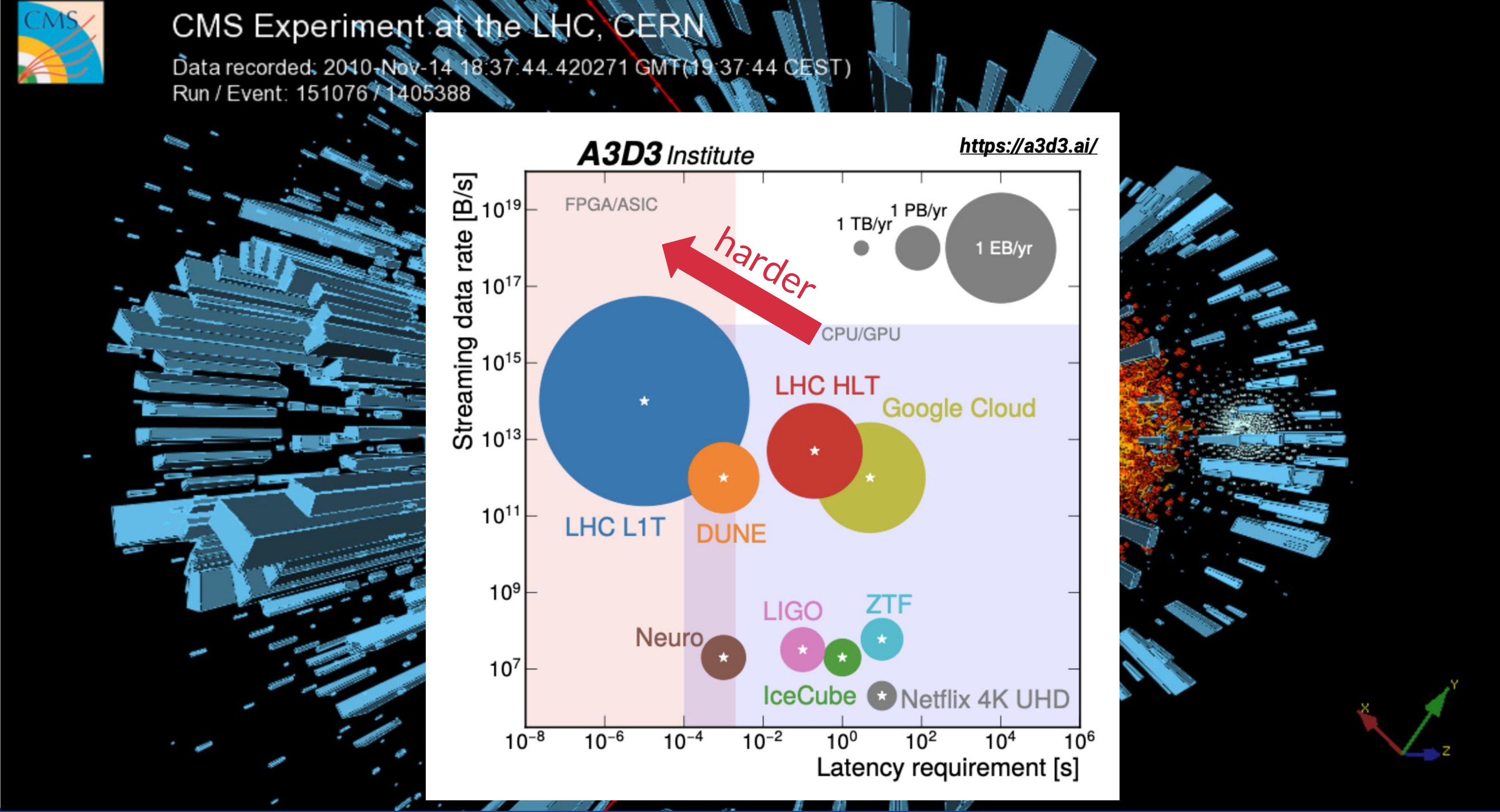
Re-thinking the CMS Level-1 Trigger with Machine Learning

Maciej Głowacki on behalf of the CMS experiment

April 2024 | Liverpool

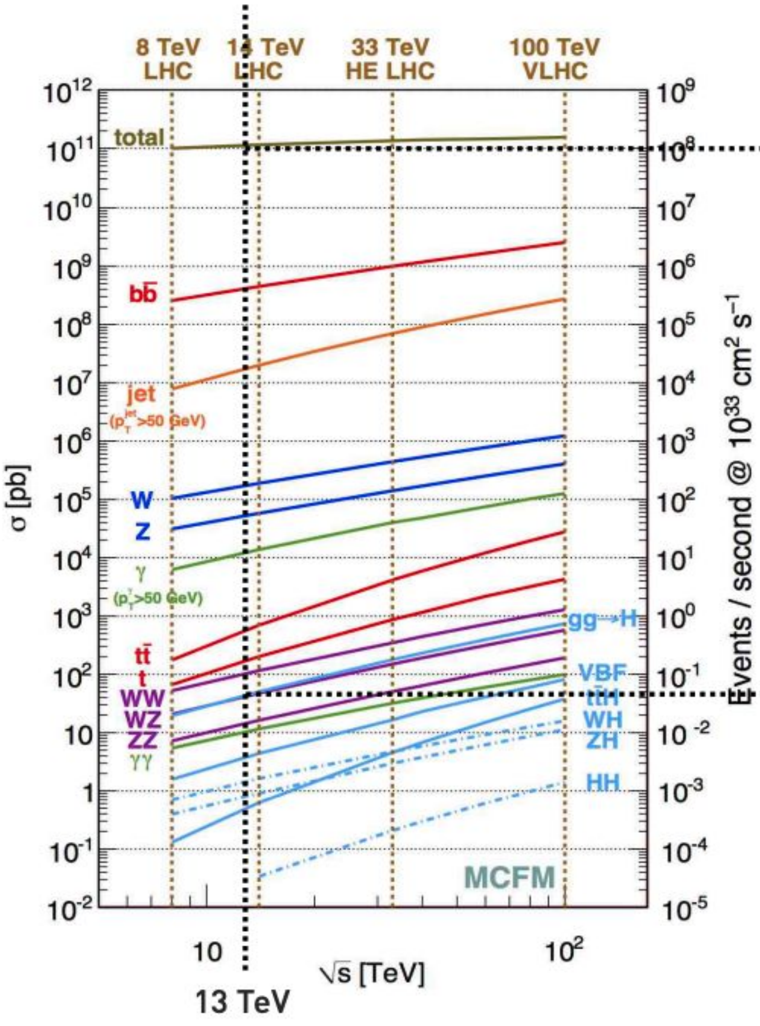
LHC & CMS


Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
~10 billion collisions per second
~10 PB of data per second


CMS Experiment at the LHC, CERN
Data recorded: 2010-Nov-14 18:37:44.420271 GMT(19:37:44 CEST)
Run / Event: 151076 / 1405388
A3D3 Institute
https://a3d3.ai/
Streaming data rate [B/s]
Latency requirement [s]
FPGA/ASIC
CPU/GPU
harder
1 TB/yr
1 PB/yr
1 EB/yr
LHC L1T
DUNE
LHC HLT
Google Cloud
Neuro
LIGO
ZTF
IceCube
Netflix 4K UHD

Towards HL-LHC

- Produce 1 Higgs boson for every billion collisions
- To stress test the SM further, we need more collisions
- Only collected 5% of total LHC luminosity to date
- Saving all this data would not be useful
  - Demand for innovative online selection approaches



"Probability" of producing "anything"

14 orders of magnitude

"Probability" of producing a Higgs

LHC

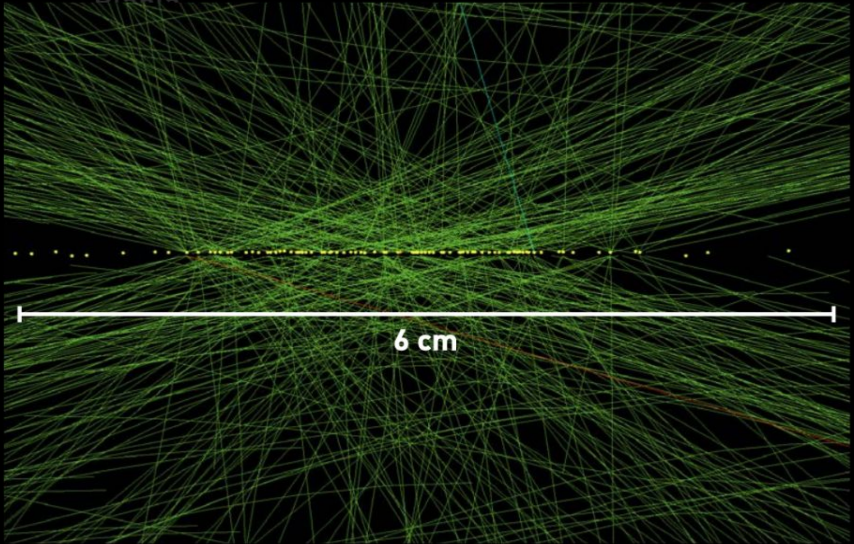
Current

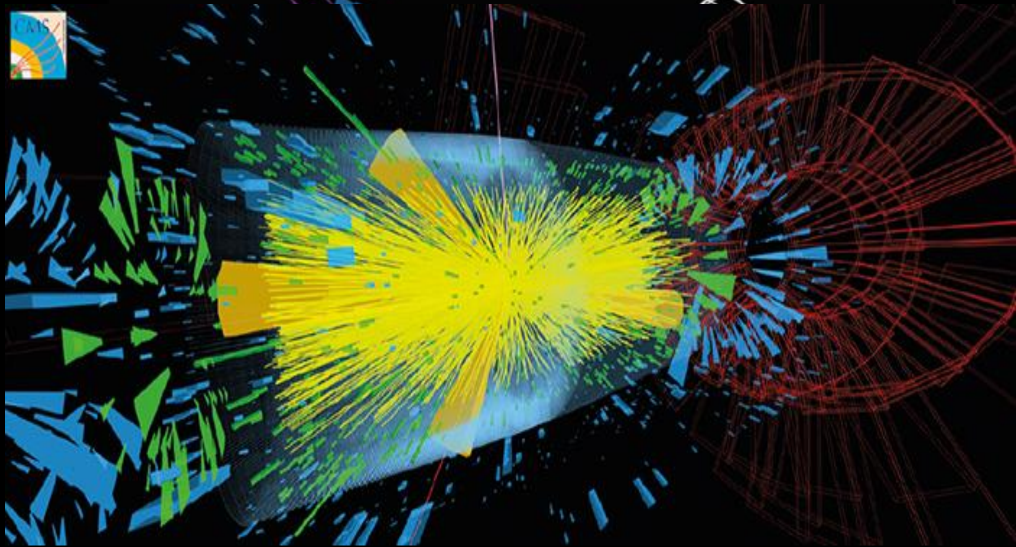
78 vertices
(average 60)

High Luminosity LHC

Phase 2

200 vertices
(average 140)



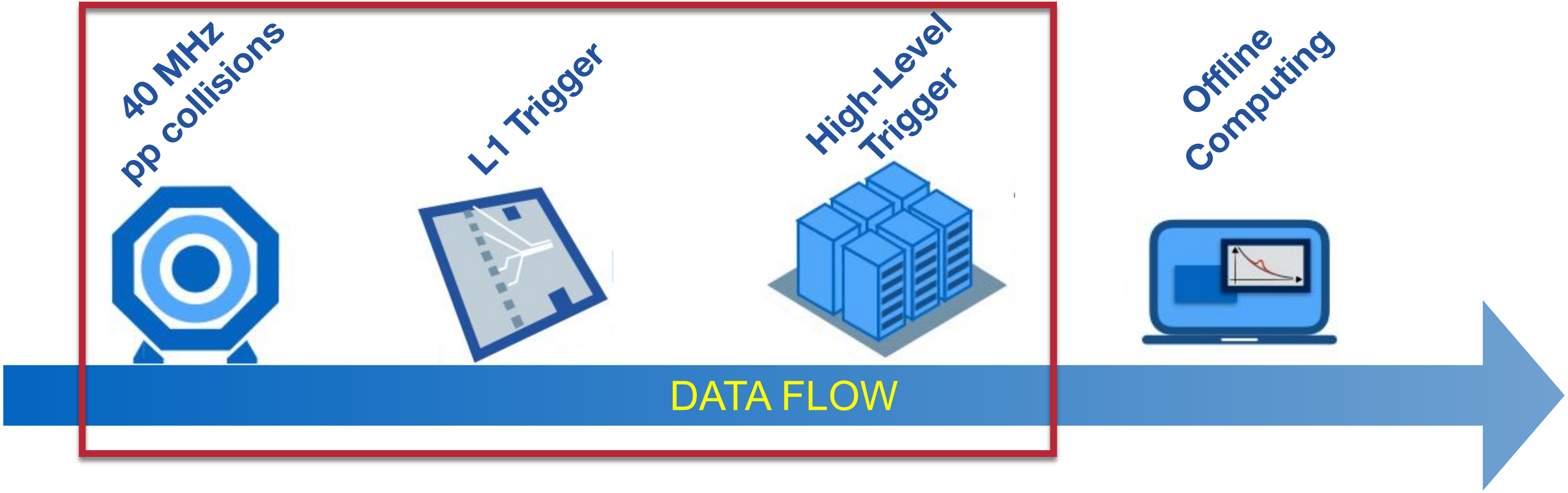


More complex environment with 200 Pile-Up interactions
Current approaches **not** sustainable (acceptance rate would 15x)

Run 3

Run 4+5

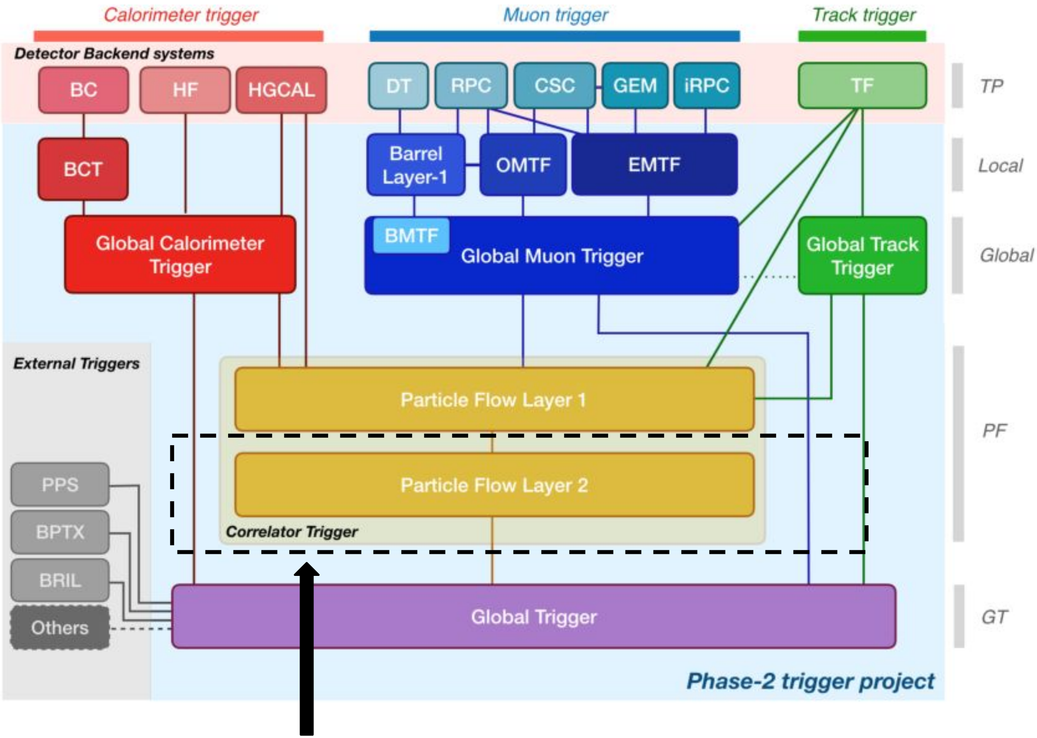
HL-LHC CMS dataflow



Deploy ML algorithms already at Level 1

HL-LHC CMS Level 1 Trigger

- L1 HL-LHC Highlights
  - Larger L1 trigger rate / detector readout rate (100 kHz → 750 kHz).
  - Larger L1 trigger latency (3.8 us → 12.5 us) → **accommodate more sophisticated algorithms.**
  - More info at L1 trigger → L1 **tracking information,** higher granularity calorimetry
  - Full Particle Flow (PF) event reconstruction and PUPPI Pile-Up mitigation.



House ML in correlator layer 2, downstream from object reconstruction and pile-up removal

Machine Learning at CMS Level 1 Trigger

ML R&D approaches at CTL2

Object Identification

Displaced jets: heavy-flavour (b) & LLP tagging

EM: e, γ, τ

Event Identification

Event classification: topological classification from reconstructed particles (HH → 4b benchmark)

(today's focus)

Event-feature reconstruction

Reconstruction: Missing Transverse Energy (P_T^{miss}) & Hadronic Transverse energy (HT) reconstruction from particle candidates.

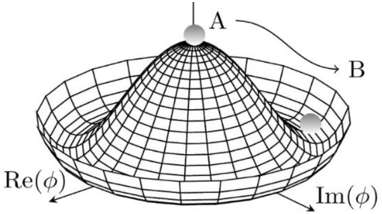
HH → 4b: HL-LHC flagship

$$\mathcal{L}_{scalar} = D_\mu\phi^\dagger D^\mu\phi - V(\phi^\dagger\phi) \text{ with } \phi = (\varphi^+ \ \varphi^0)^T \text{ doublet under } SU(2)$$

$$V(\phi^\dagger\phi) = -\mu^2(\phi^\dagger\phi) + \lambda(\phi^\dagger\phi)^2$$

$$\phi(x) = \frac{1}{\sqrt{2}}\exp\left(i\sigma^i\xi(x)\right)\begin{pmatrix}0\\ v+h(x)\end{pmatrix} \xrightarrow{EWSB} \phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v+h(x)\end{pmatrix}$$

$$\text{Non-zero vacuum expectation value } v = \mu^2/\lambda$$



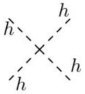
$$\sigma(pp \to HH) \simeq \frac{\sigma(pp \to H)}{1000}$$

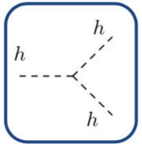
If SM is correct :
→ 4000 HH events during Run-2
… not enough to see HH

$$\mathcal{L}_{scalar} = D_\mu\phi^\dagger D^\mu\phi + \mu^2(\phi^\dagger\phi) - \lambda(\phi^\dagger\phi)^2$$
$$= \frac{v^2}{8}\left(g^2 W^i_\mu W^{i\mu} + g'^2 B_\mu B_\nu - 2g'gB_\mu W^{3\mu}\right)\left(1+\frac{h}{v}\right)^2$$
$$+ \underbrace{\frac{1}{2}(\partial_\mu h\partial^\mu h)}_{\text{kinetic term}} - \underbrace{\lambda v^2h^2}_{\text{mass term}} - \underbrace{\lambda v h^3}_{\text{trilinear coupling}} - \underbrace{\frac{\lambda}{4}h^4}_{\text{quartic coupling}} - \frac{\lambda v^4}{4}$$

$$m_H = \sqrt{2\lambda}v$$

Mass of the weak bosons
+
Mass of the fermions through Yukawa couplings

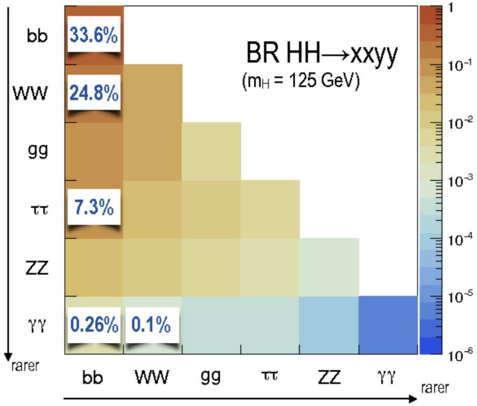




(on our menu today)

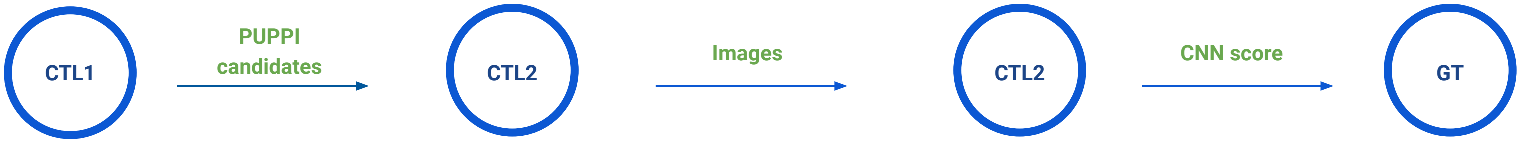
Fully parameterized by **λ**

- Theory value given by v and m_H
- Experimental measurement
  → Test of the SM
  → Probe the shape of the potential
  → Very sensitive to BSM



Dataflow

Topological trigger at Level 1



Particle Flow reconstruction and pileup mitigation

1. Calorimeter and track information used in PF reconstruction and PUPPI pile-up mitigation to deliver *PUPPI candidates* to CTL2

Histogramming & preprocessing

2. PUPPI candidate P_T binned in the η - ϕ space of the detector to produce 2-D *images* used concurrently by topological classifier and Jet Finding Algorithm. Preprocessing is applied to refine images to serve as input classifier.

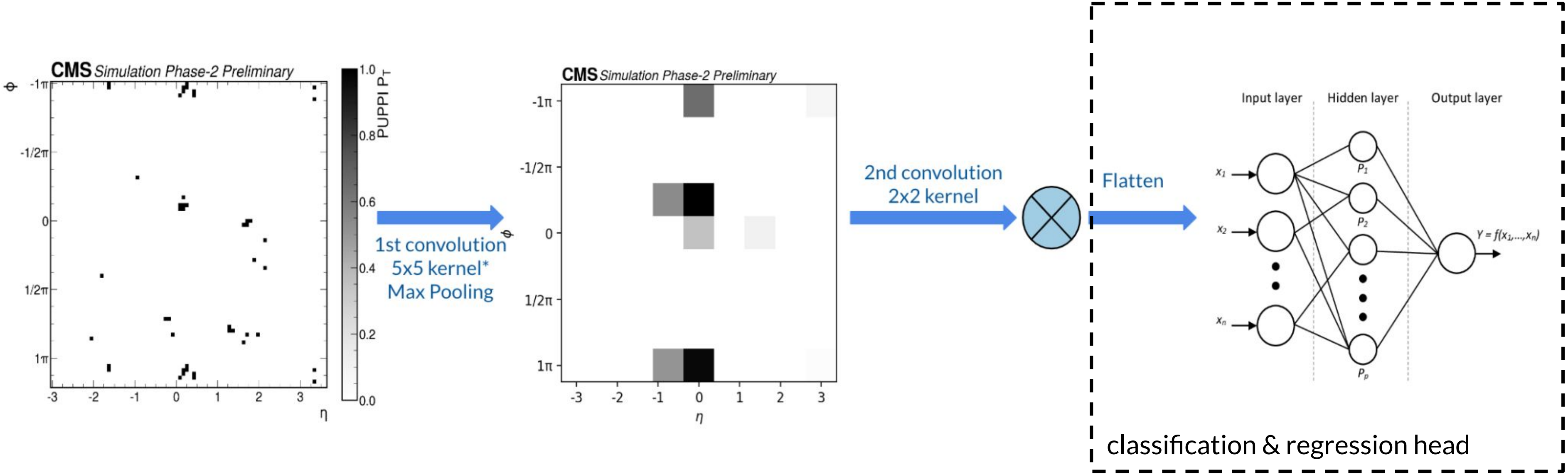
Classification

3. Convolutional Neural Network (CNN) executes its inference procedure from input images.

Global Trigger input

4. CNN probability score delivered to GT to be used alongside existing menu bits.

Architecture

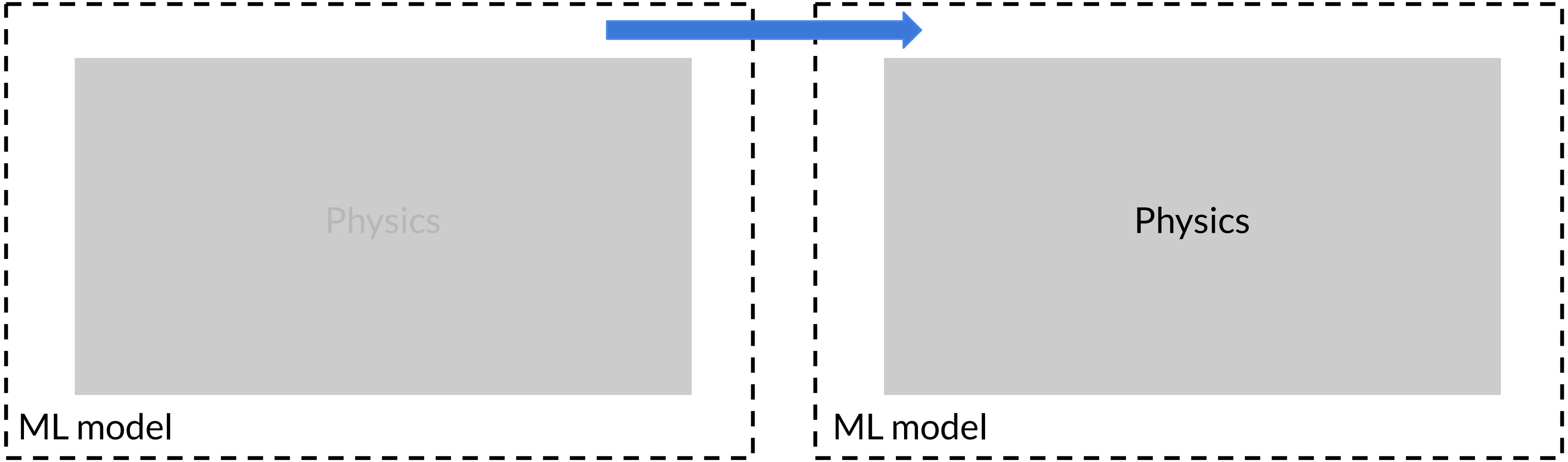


*Kernel regularisation $\propto \sum(\text{weights}_{\text{kernel}})^2$

Explainability

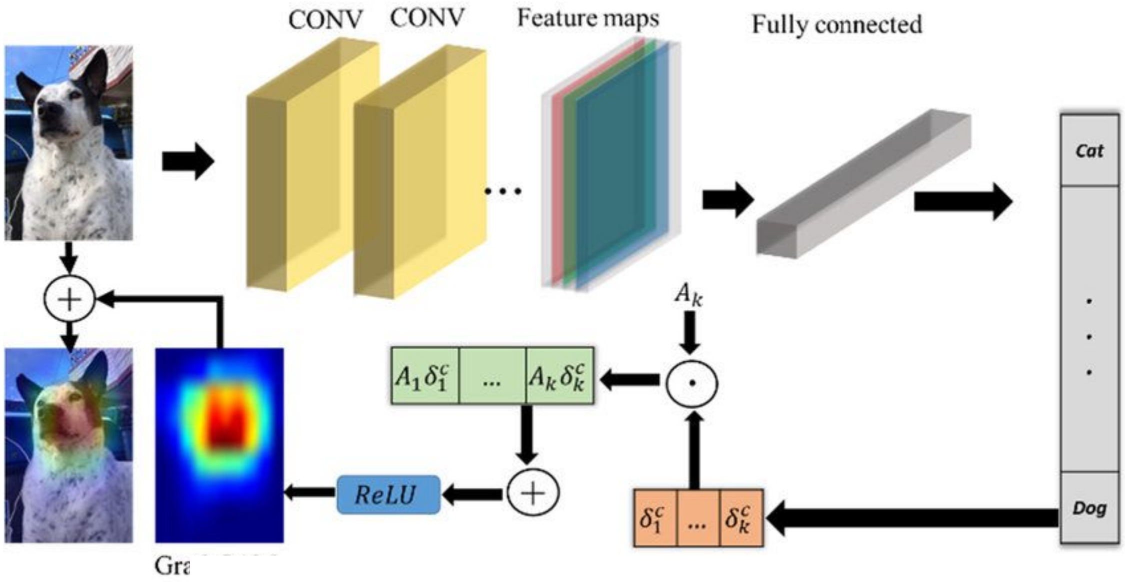
Making physics transparent is vital at the triggering stage

- Understand model decision making
- Condition on/regress classically derived quantities



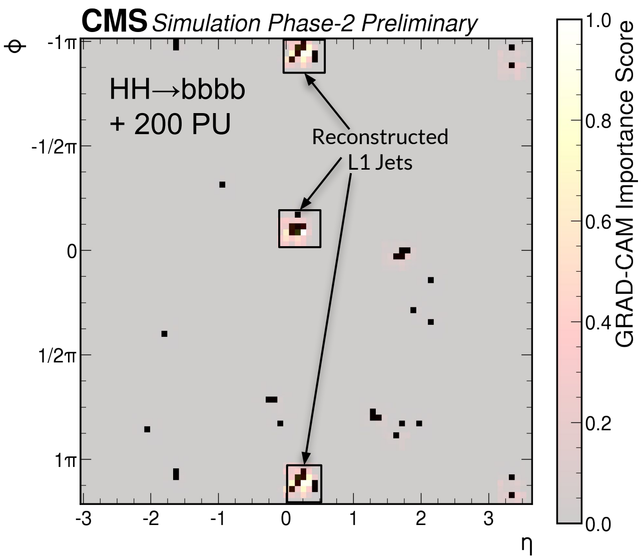
GradCam

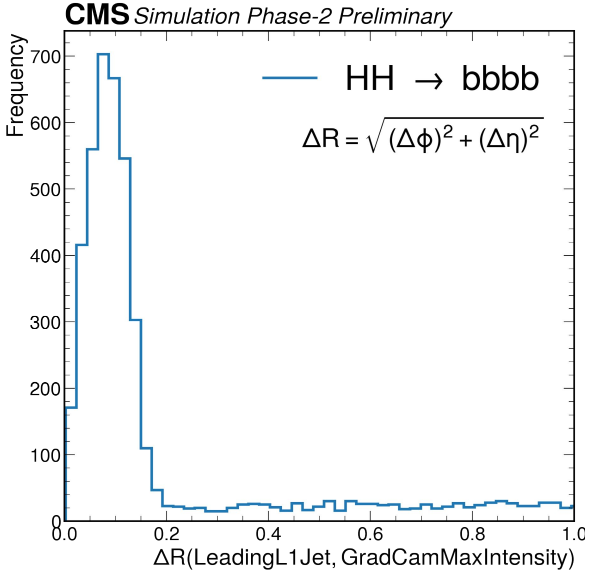
- Pixel-wise Importance scores computed from backpropagated gradients of class scores w.r.t. final feature map activations.
- Gradients are then averaged, producing a weight for each feature map point that represents its importance in the classification decision.
- Visualises areas of high importance in the input on the final classification.

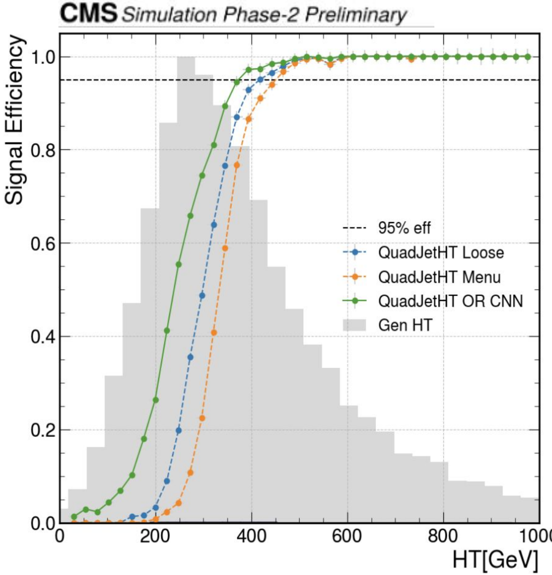


$$\text{Grad-CAM}_c(x,i,j) = \text{ReLU}\left(\sum_k \frac{\partial Y^c}{\partial A^k_{i,j}} A^k_{i,j}\right)$$ [2]

Physics Performance







- Grad-CAM scores indicate model is learning to cluster PUPPI candidates into jets.
  - Good agreement seen between highest Grad-CAM intensity and leading Level 1 reconstructed jet.
- ~10-15% gain in integrated and offline efficiencies compared to “QuadJetHT” trigger path at constant rate.

Machine Learning on the edge

Machine Learning on FPGAs

TensorFlow

Keras

ONNX

PyTorch

API

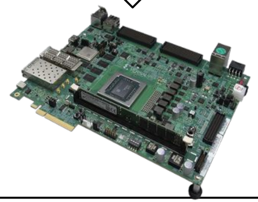
hls4ml

https://arxiv.org/abs/2103.05579

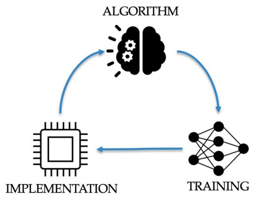
C++

Vivado™ HLS

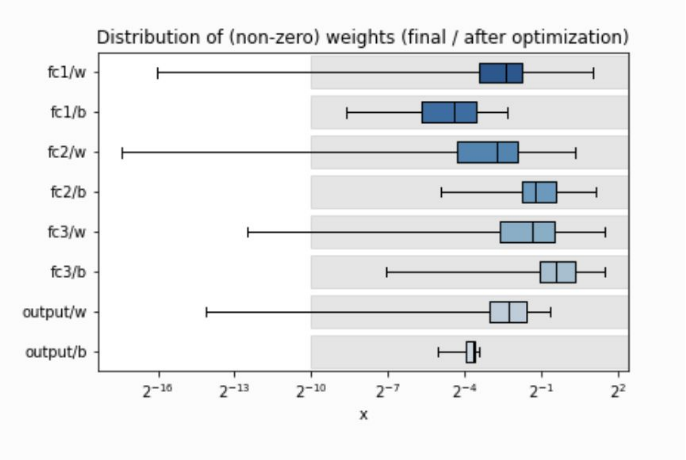
VHDL



Co-design

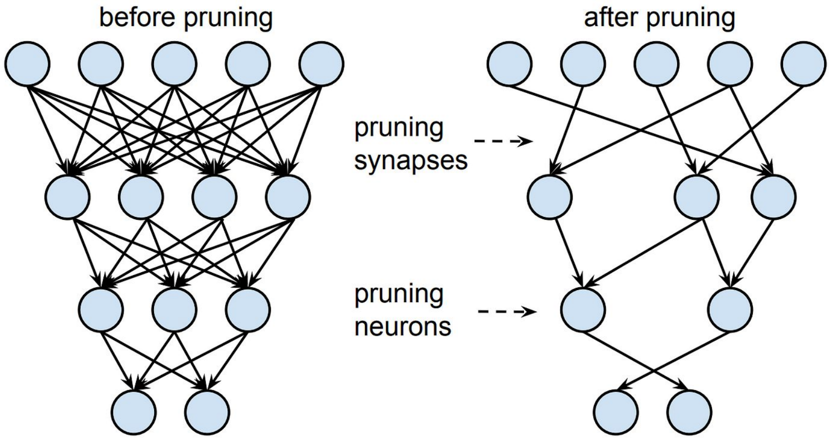


Quantisation



QKeras

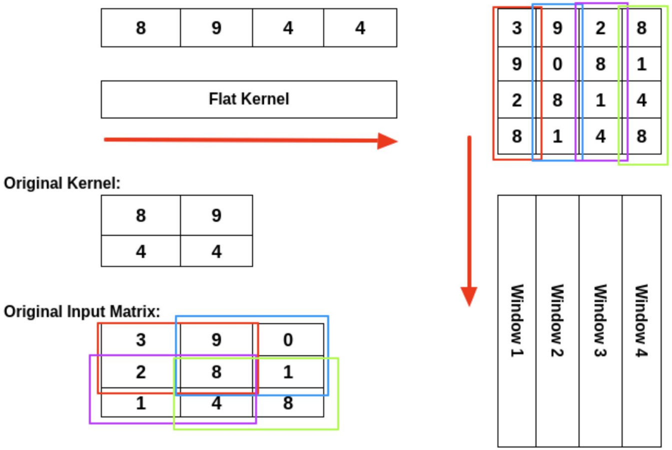
Pruning



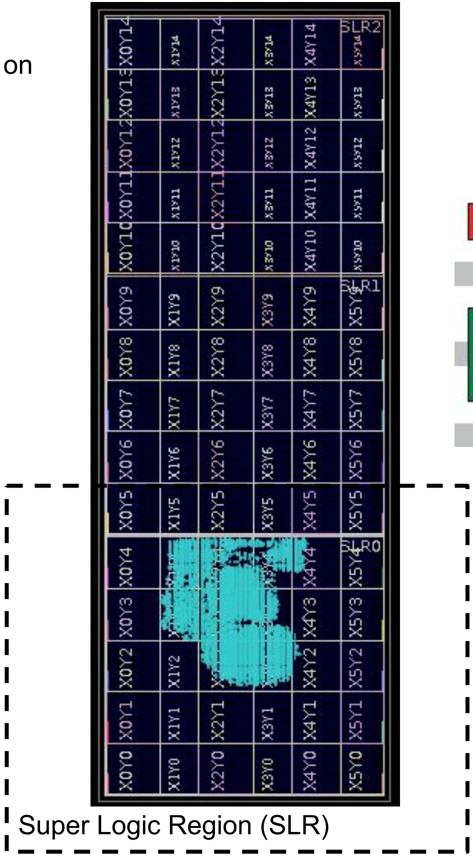
Implementation on FPGA

FPGA: Xiling xcvu9p-flga2577-2-e

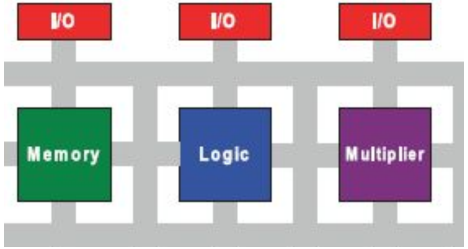
- Im2col algorithm used to increase parallelism on each clock cycle
- Benefits from highly optimsied vector-wise operations



CNN implementation on FPGA chip



Super Logic Region (SLR)



Summary

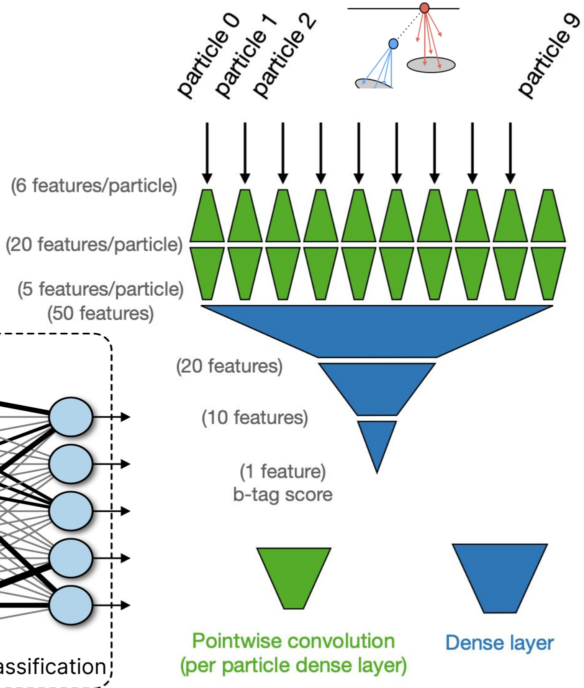
- High-Luminosity LHC:
  - Ability to study increasingly rare phenomena (BSM will have less places to hide)!
  - Highly complex data taking environment (<PU> 200)
- Upgrade to the CMS level 1 trigger:
  - Tracking and high granularity calorimeter information
  - Particle Reconstruction & Pile-Up mitigation
  - Increased read-out rate and relaxed latency constraints
- Machine Learning approaches viable:
  - We have the tooling to translate ML models to firmware running on FPGA
  - Software-hardware co-design enables ML use at Level 1 with optimised algorithms meeting latency and resource constraints
  - ML outperforms classical methods
- Considerations:
  - Changing environments: i.e. tracker degradation (neccesiates continual training & deployment)

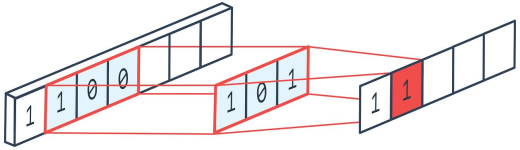
Back up

b-tagging

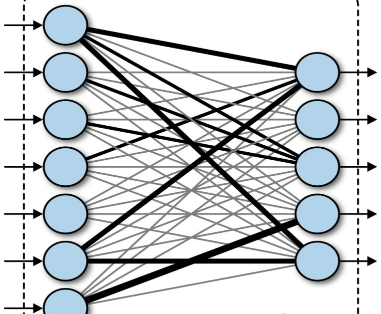
Particle ID	pT, η, φ	Impact parameter z-position

Input Features / PUPPI candidate (extended tracking)



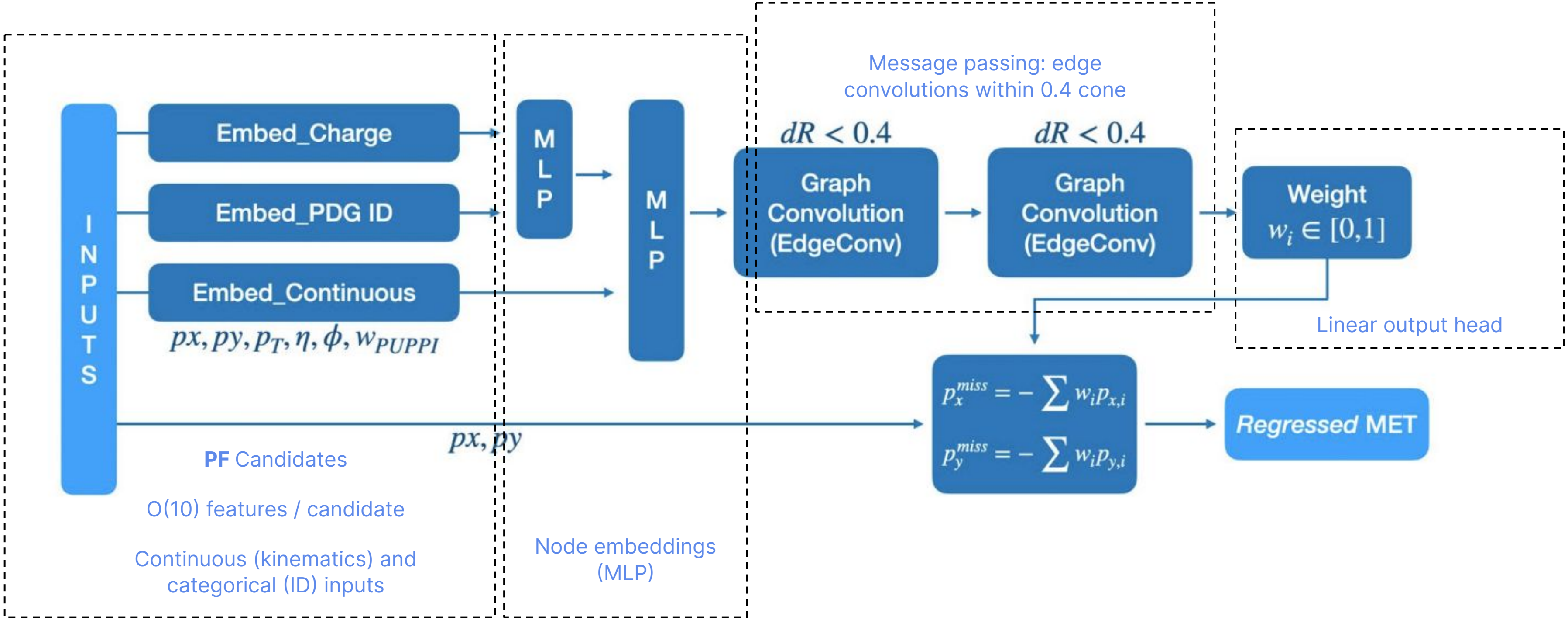


1D convolutional feature extractors



Dense layer classification

Missing-Transverse Energy



Phase II trigger menus

- Quad jet and HT requirements (reconstructed jets and summed)
- ~50-60% efficiency at ~10 kHz rate

L1 Trigger seeds	Offline Threshold(s) at 90% or 95% (50%) [GeV]	Online Threshold(s) (Barrel) [kHz]	Rate* $\langle PU \rangle = 200$ [kHz]	Additional Requirement(s) [cm, GeV]	Objects plateau efficiency [%]
Single/Double/Triple Lepton (electron, muon) seeds					
Single TkMuon	22	20	12	$\|\eta\| < 2.4$	95
Double TkMuon	15,7	13,6	1	$\|\eta\| < 2.4$, $\Delta z < 1$	95
Triple TkMuon	5,3,3	4,2,2	16	$\|\eta\| < 2.4$, $\Delta z < 1$	95
Single TkElectron	36	32	24	$\|\eta\| < 2.4$	93
Single TkIsoElectron	28	25	28	$\|\eta\| < 2.4$	93
TkIsoElectron-StaEG	22, 12	19,8	36	$\|\eta\| < 2.4$	93, 99
Double TkElectron	25, 12	22,10	4	$\|\eta\| < 2.4$, $\Delta z < 1$	93
Single StaEG	51	46	25	$\|\eta\| < 2.4$	99
Double StaEG	37,24	32,20	5	$\|\eta\| < 2.4$	99
Photon seeds					
Single TkIsoPhoton	36	33	43	$\|\eta\| < 2.4$	97
Double TkIsoPhoton	22, 12	19,9	50	$\|\eta\| < 2.4$	97
Taus seeds					
Single CaloTau	150(119)	109	21	$\|\eta\| < 2.1$	99
Double CaloTau	90,90(69,69)	65,65	25	$\|\eta\| < 2.1$, $\Delta R > 0.5$	99
Double PuppiTau	52,52(36,36)	36,36	7	$\|\eta\| < 2.1$, $\Delta R > 0.5$	90
Hadronic seeds (jets,H_T)					
Single PuppiJet	180	121	70	$\|\eta\| < 2.4$	100
Double PuppiJet	112,112	72,72	71	$\|\eta\| < 2.4$, $\Delta\eta < 1.6$	100
PuppiH_T	450(377)	363	11	jets: $\|\eta\| < 2.4$, $p_T > 30$	100
QuadPuppiJets-PuppiH_T	70,55,40,40,400(328)	41,30,19,19,316	9	jets: $\|\eta\| < 2.4$, $p_T > 30$ safety online cut $p_T > 25$ for jets	100,100

Path	Inclusive acceptance	Loosely presel. evts. acceptance	YR presel. evts. acceptance
QuadJet_70_55_40_40	59%	85%	99%
QuadJet_70_55_40_40_HT320	50%	76%	91%
QuadJet_40_40_40_40_MuJet40	23%	36%	44%
QuadJet_40_40_40_40_MuJet40_HT250	22%	35%	43%
QuadJet_70_55_40_40_HT320 OR QuadJet_40_40_40_40_MuJet40_HT250	52%	79%	94%

Datasets (DAS)

MinBias:
/MinBias_TuneCP5_14TeV-pythia8/Phase2HLTTDRWinter20DIGI-PU200_110X_mcRun4_realistic_v3-v3/GEN-SIM-DIGI-RAW

HH→ bbbb:

/GluGluToHHTo4B_node_SM_TuneCP5_14TeV-madgraph_pythia8/Phase2HLTTDRWinter20DIGI-PU200_110X_mcRun4_realistic_v3-v5/GEN-SIM-DIGI-RAW