M. Musich
Università di Pisa & INFN
on behalf of the CMS Collaboration

Triggering Discoveries in High Energy Physics III
Vysoké Tatry (SK), 9-13th December 2024

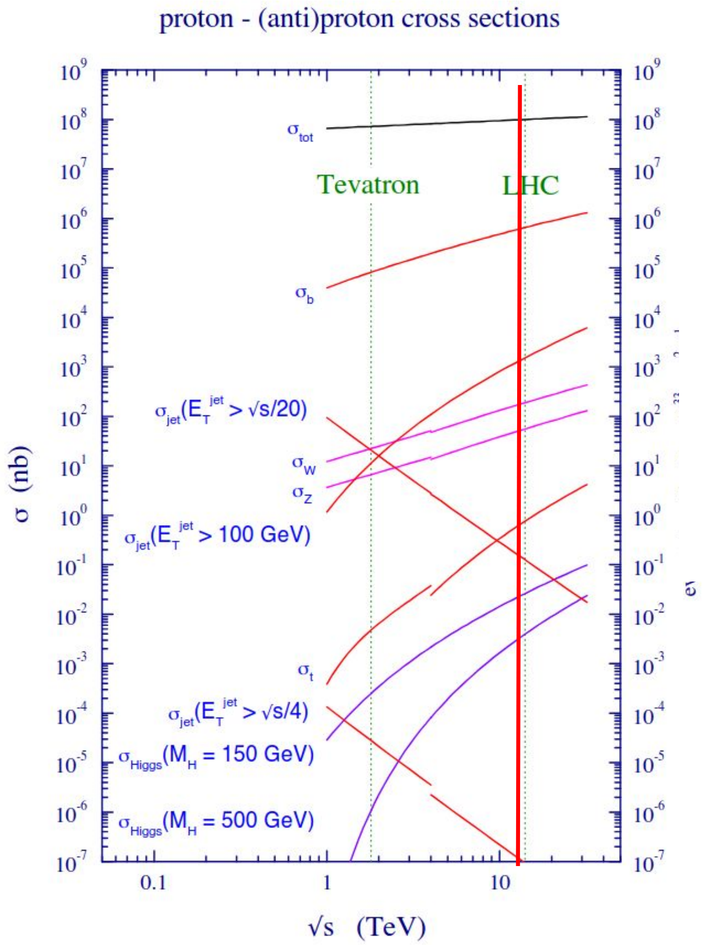
Description & performance of the current CMS trigger

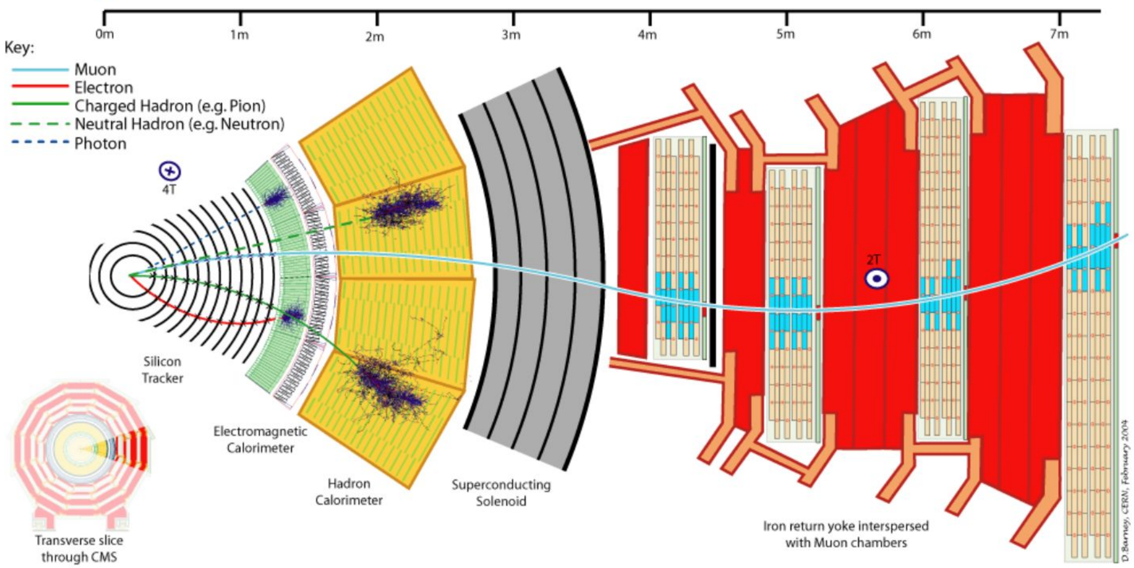
Introduction

- The CMS experiment
  - Description of the CMS trigger system
- LHC Run 3 conditions
  - Run 3 trigger strategies
- HLT Technology
  - Usage of GPUs
  - Impact on HLT timing
- Standard HLT Streams
  - Focus: triggering on LLP
- Parking HLT Streams
  - Focus: B physics and HH
- Scouting HLT Streams
- HLT Objects performance
  - Jets performance
  - MET performance
  - Muon performance
  - E/γ objects performance
  - Tracking performance
  - B-tagging performance
- Triggering on anomalies

The CMS Experiment

- CMS is general purpose detector at the CERN LHC
- Sub-detectors to identify particles & Particle Flow
- Real time decision to store interesting events (Trigger)

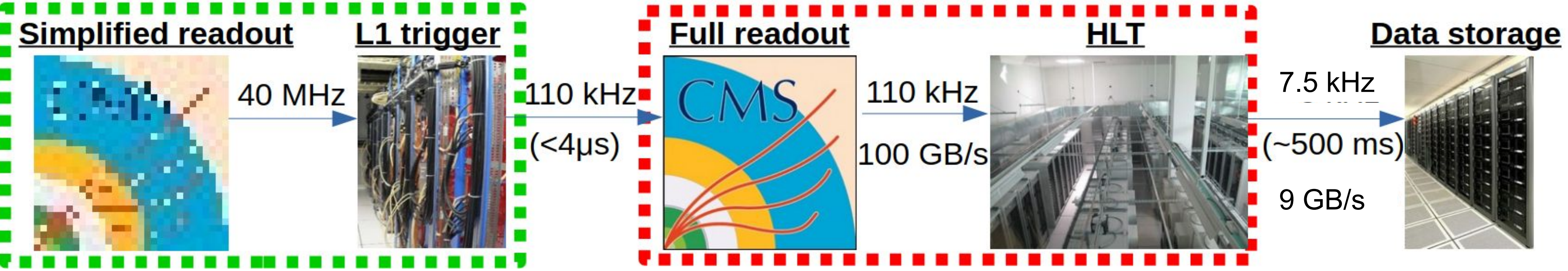




- Lumi: 2-2.4 x 10^{34} cm^2 s^{-1} in LHC Run 3
- Maximum 2556 bunches (2352 in CMS currently), 1.6-1.8x10^{11} p/bunch
- Total collision rate 33 MHz (40MHz Bunch crossings)
    - b-quark production rate 10 MHz
    - W boson production rate 4 kHz
    - Top quark production rate 20 Hz
    - Higgs boson prod. rate only 1 Hz
    - SUSY rate(m@TeV) below 0.1 Hz

Trigger and Data Acquisition

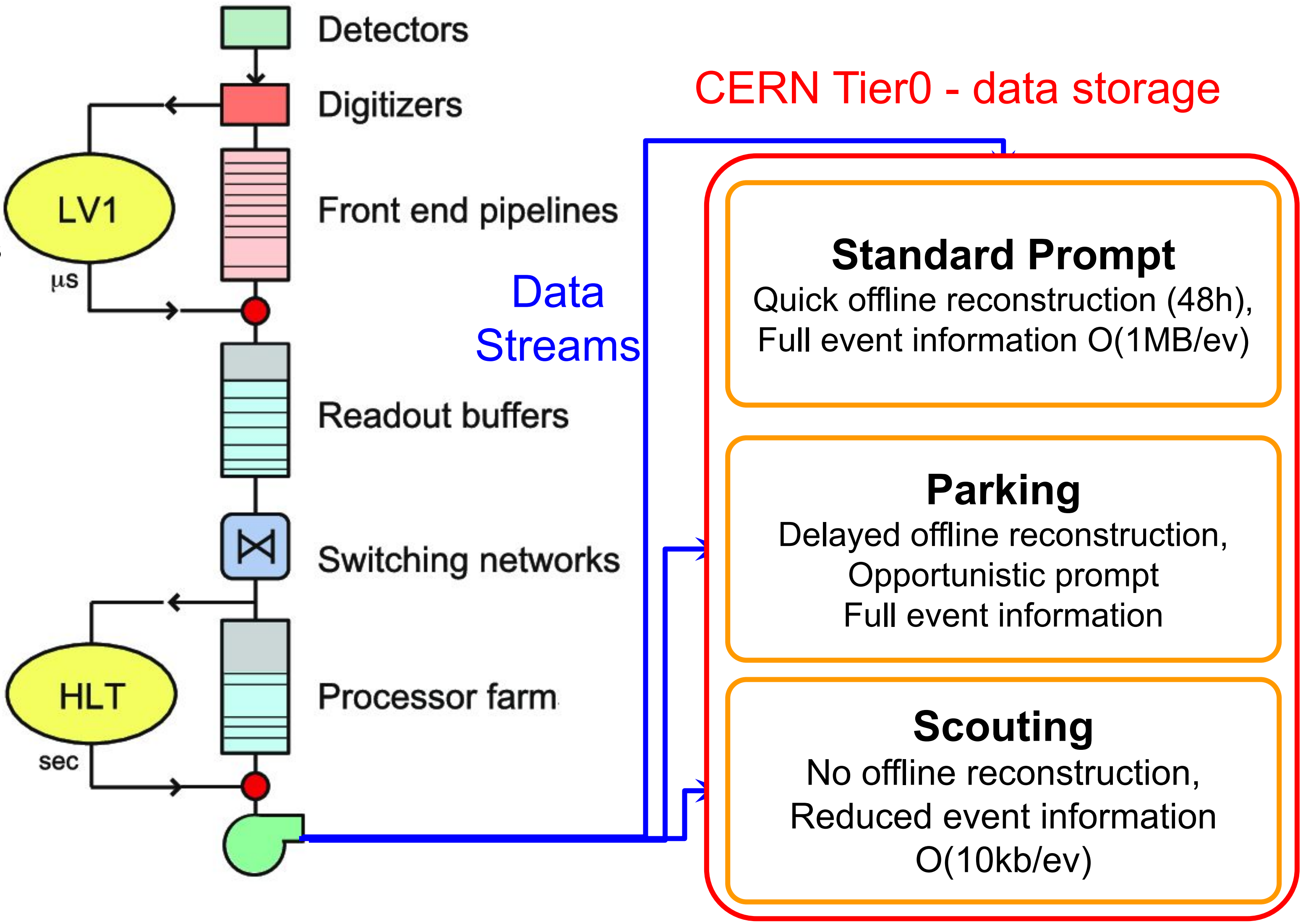
- Hardware trigger (L1): 40 MHz → 110 kHz
  - simplified readout (no tracker), small latency (3.8μs).
- Software trigger (HLT): 110 kHz → ~7.5 kHz.
  - full event readout available (~1.2MB/event @ PU ~ 64);
  - simplified reco: O(50k) CPUs → 420 ms/event on average



CMS: Two level triggering system

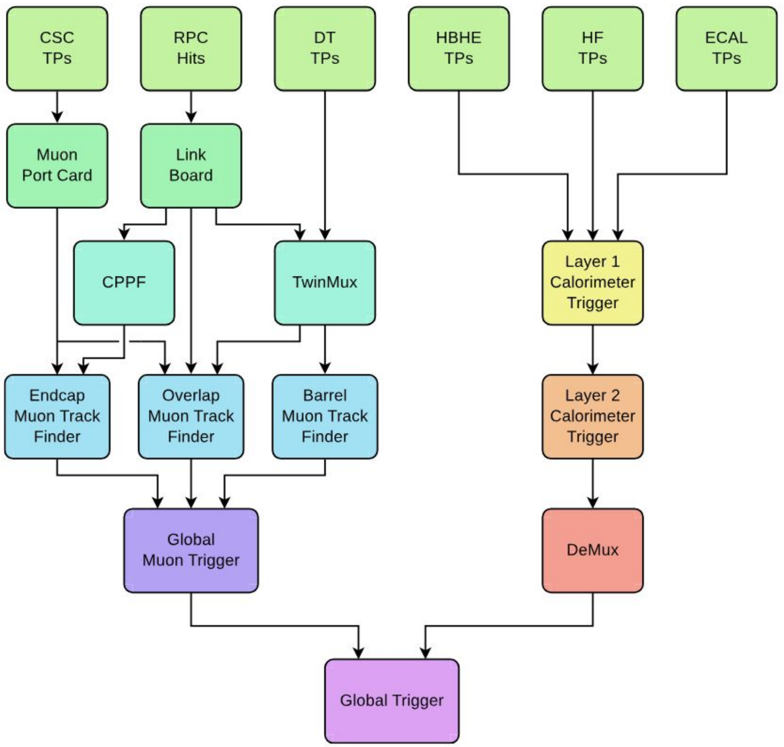
L1 Trigger
hardware-based
coarse granularity,
only muon systems
and calorimeters

High Level Trigger
software-based
(CPU/GPU)
full granularity,
entire detecto



The Level-1 Trigger: Architecture & Implementation

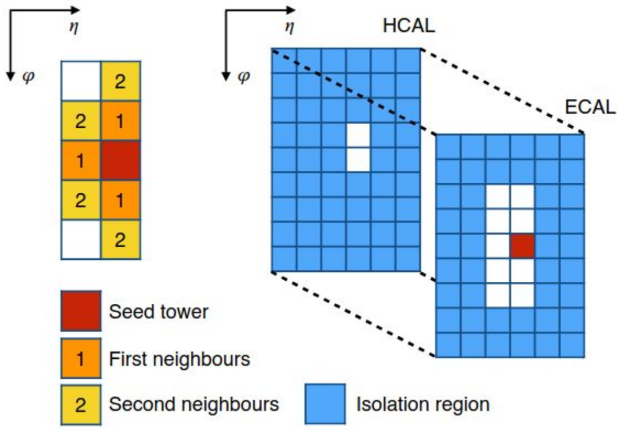
- Each event processed by Muon and Calo trigger
  - **Muon trigger** consists of four muon detection systems combined early in the processing chain of the trigger, in order to improve the efficiency and resolution, but also to reduce trigger rate
  - **Calorimeter** trigger for reconstructing electrons, photons, tau candidates, jets and energy sums
- No inner silicon tracking readout used
  - planned for Phase 2 upgrades
- Global trigger that combines the various objects that are formed by the μGMT and caloL2 triggers
- Set of requirements (including sophisticated operations as invariant mass and ΔR on trigger objects): L1 menu
  - maximums 512 requirements in a logical OR

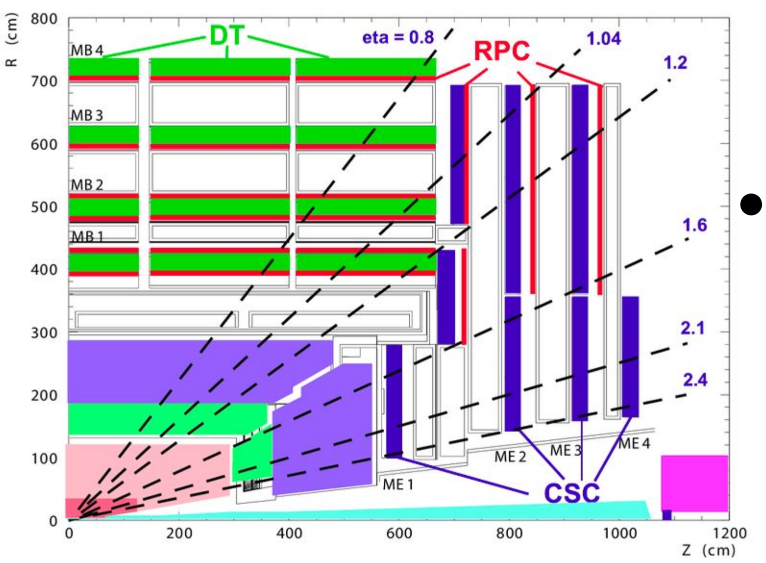


JINST 15 P10017

The Level-1 Trigger: Algorithms

- Electrons and photons are reconstructed using cluster shape and electromagnetic (EM) fraction to discriminate against jets
  - Isolation implemented using LUTs
- Jets reconstructed using sliding window algorithm that looks for trigger tower seeds with an energy over given threshold; 9x9 trigger towers are summed to match offline jets (R = 0.4) after which the jets are also pileup subtracted and calibrated

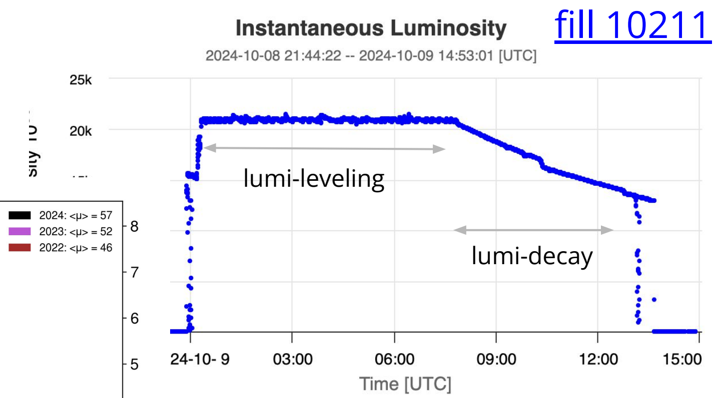


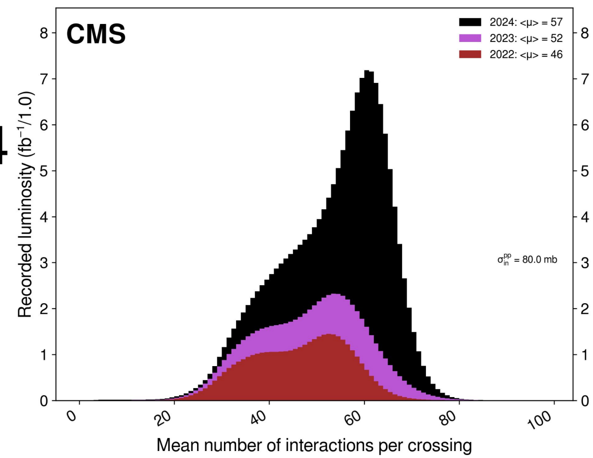


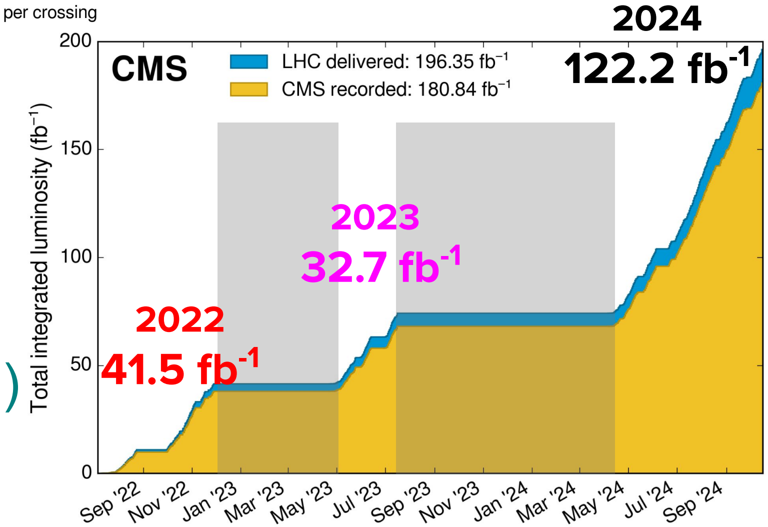
- Energy sums are calculated by summing the jet energies with restrictions to jet energy and to pseudorapidity (for the H_T); for MET: all TTs over ET(η, PU) summed (in full η)
- Muons reconstruction using an extrapolation based track finding in barrel, pattern based in overlap/endcap region
  - muon p_T assignment (both constrained and unconstrained at vertex) based on Δφ in barrel, patterns in overlap region and BDT regression used in the end cap

Run 3 conditions

- LHC:
  - Longer fills, maximise inst luminosity via lumi leveling
  - Larger luminosity:
    - 2.0E34 → 2.15E34 $cm^{-2}s^{-1}$ (2024)
  - Large pileup: 46 →~ 57 (2024)
  - Larger energy 13 TeV → 13.6 TeV
- CMS:
  - Main Phase-I upgrades completed in Run-2 (Pixels, L1 trigger, HCAL endcap)
  - HCAL barrel: new readout
  - PPS: fully integrated in CMS (CT-PPS)
  - New muon detector (GEM)
  - GPU at HLT







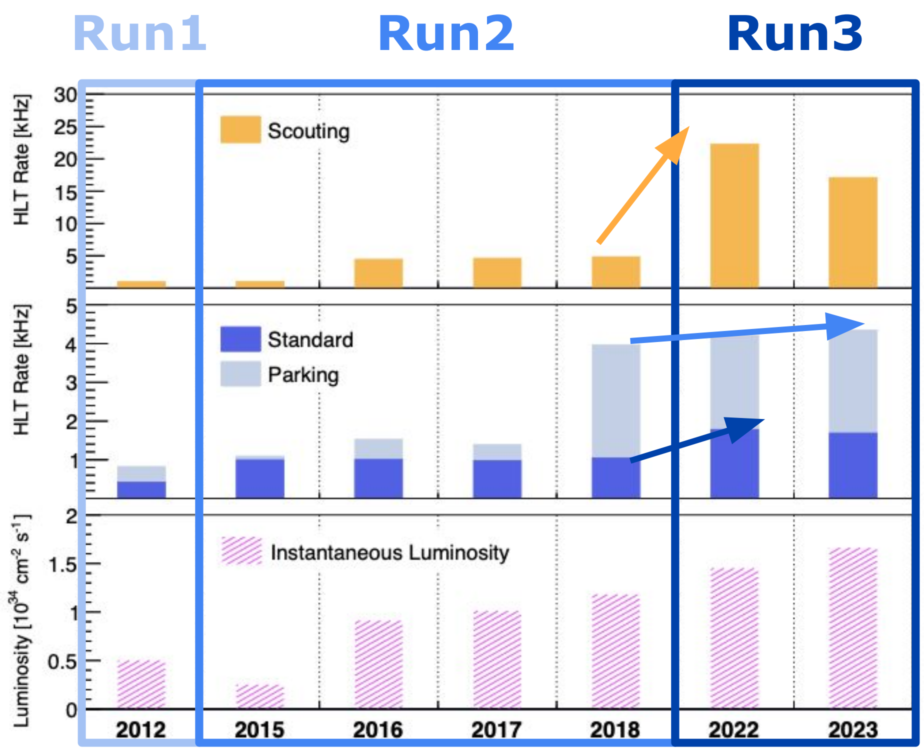
Run 3 Trigger Strategy

- Main principles:
  - Trigger on new objects (more resources to parking b-physics, VBF/S, hh→4b, LLP)
  - Cover more phase space (speed up in HLT reconstruction via GPU ⇒ more bandwidth to scouting di-muons, jets, photons, etc.)

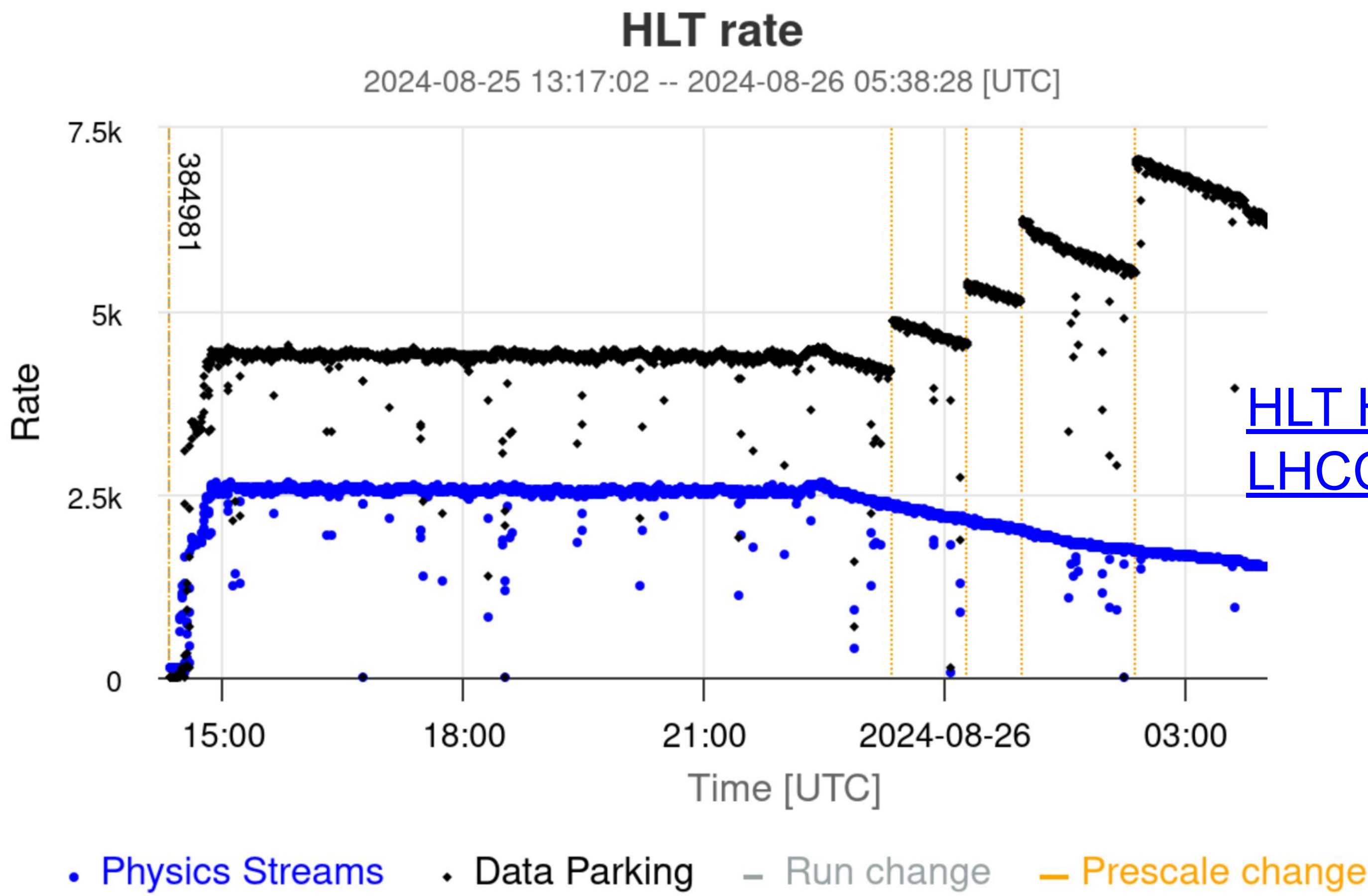
- **Core physics program**
  - promptly reconstructed w/in 48 hrs
- **Data Parking (Delayed Reconstruction)**
  - reconstruction when resources are available→ promptly in 2022 and 2023, 2024 (!)
- **Data Scouting (Trigger Level Analysis)**
  - no offline reconstruction
  - only HLT info [~10 kB vs ~1 MB]

→ analysis done w/ HLT objects & calibrations

From EXO-23-007



Anatomy of a 2024 Fill

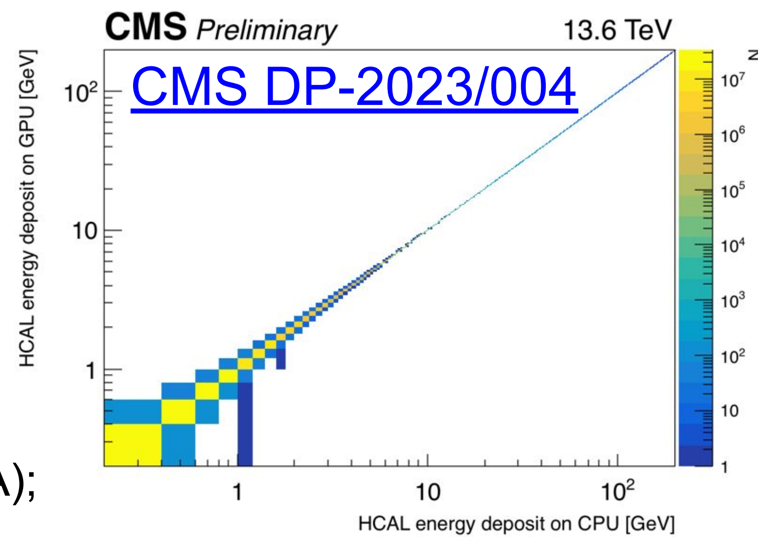


[HLT Highlights for LHCC 159]

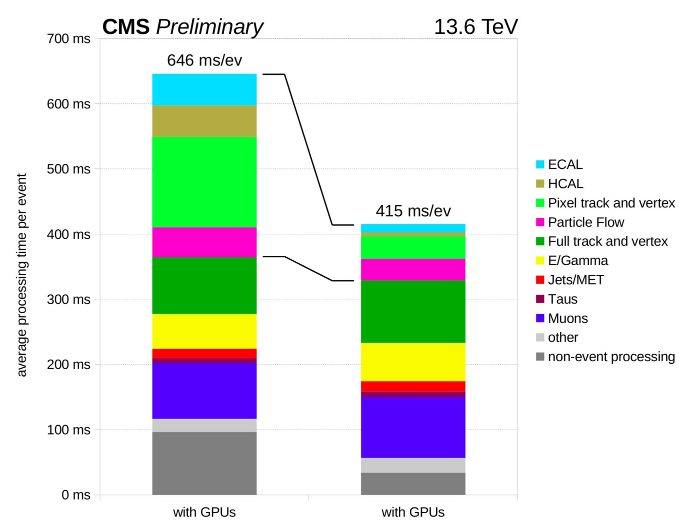
The standard streams follows the luminosity profile, while the parking one shows the strategy of optimizing the output bandwidth.

HLT Technology CPU & GPU

- GPUs very powerful in parallel computing,
  - exponential increase!
- Increasing usage in High Energy Physics
  - especially for Machine Learning.
- CMS HLT uses heterogeneous resources CPUs + GPUs to increase computing power
  - CMS is using GPUs in the trigger software starting from Run 3.
    - Big effort at the start of the run in porting Pixel, HCAL, ECAL code on GPU (CUDA);
  - Particle Flow reconstruction ported to run on GPUs in 2024
    - GPUs require re-writing of HLT code and AI (Alpaka)
  - first step towards an heterogeneous era!
- More computing power allowed CMS to:
  - develop more accurate object reconstruction in HLT
    - Better resolution → lower rates and higher efficiency
  - Lower rates → extend the physics program
    - Allows to running HLT scouting at much higher rate than Run 2

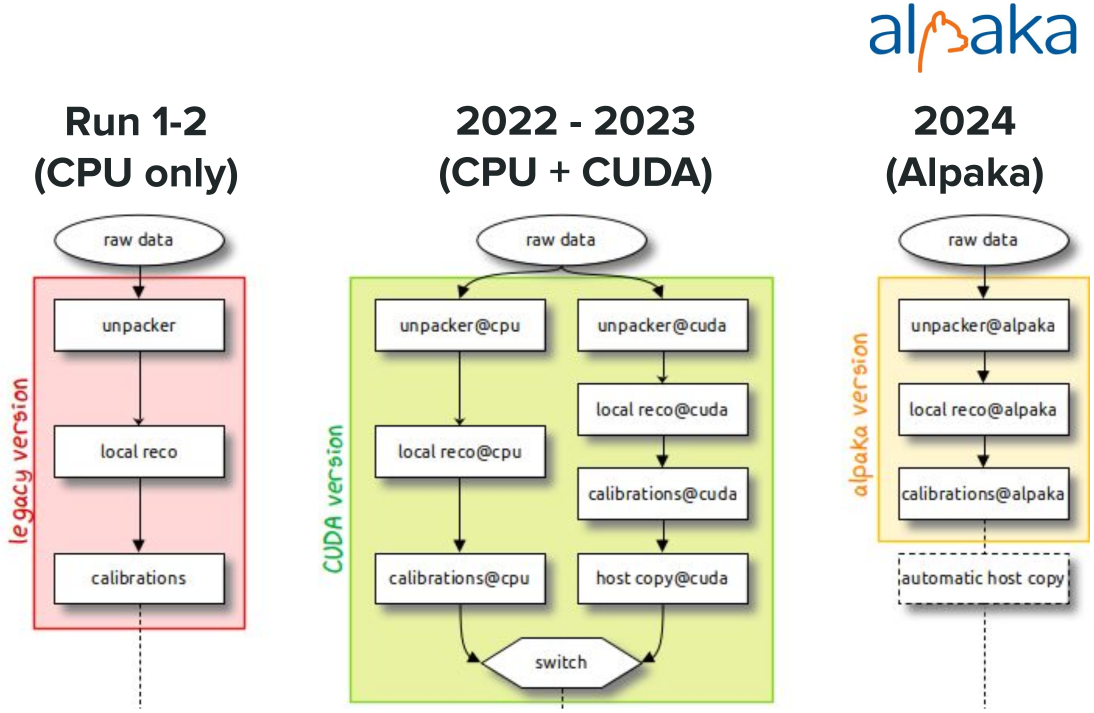


CMS DP-2023/004



CMS DP-2024/082

Heterogeneous Reconstruction at HLT



[Alpaka](#) *performance portability* library allows a single source to be built for and run on:

- x86 and ARM CPUs
- NVIDIA and AMD GPUs
- experimental support for Intel GPUs (and FPGAs) → not yet enabled in CMSSW

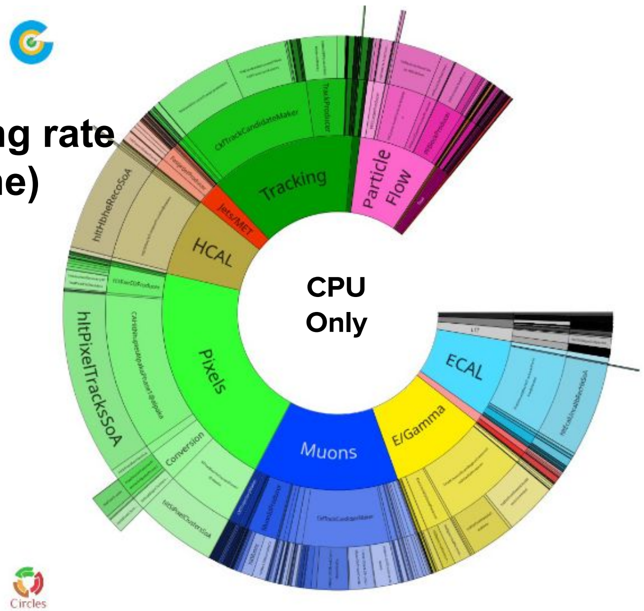
- 2017-2021: **CUDA** code integrated in CMSSW to perform pixel, ECAL, HCAL local reco and pixel tracking on **NVidia GPU**
- 2022: first collisions collected using an **heterogeneous farm** (CPU+GPU)
- 2022-23: work on the migration of CUDA code to **Alpaka (portability library)**
  - Pixel
  - ECAL
  - PFRecHit/Cluster of HBHE **(new)**
- 2024: **deployment of Alpaka at HLT**
  - ECAL, PF, and Pixel reco (March)
  - HCAL reco (July)

HLT Timing in 2024

- Currently offloading to GPU ~35% of HLT reco: CMS DP-2024/082
  - Heavy usage of tracking due to scouting updates;
    - **~50% speedup** compared to CPU only

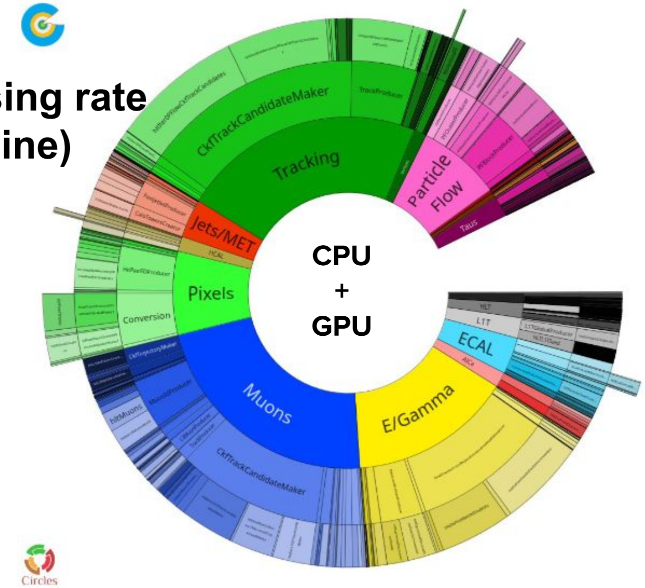
timing of current HLT menu (PU~63)

processing rate (1 machine) 400 Hz



638 ms

processing rate (1 machine) 603 Hz



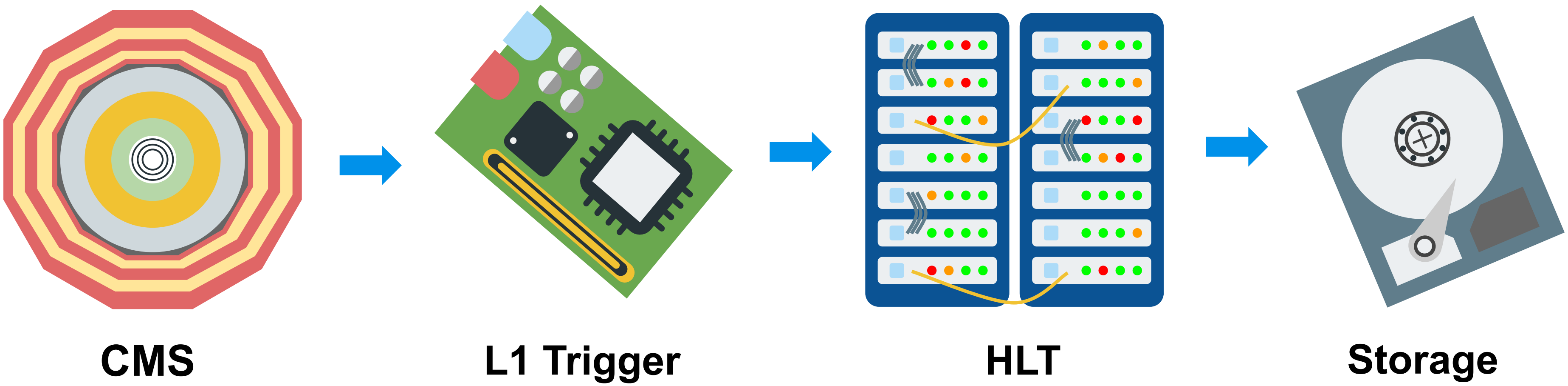
421 ms

→ GPUs as a way to reduce power usage

		each GPU		CPU-only
throughput:	N_{GPUs} ×	0.51 kev/s	+	1.20 kev/s
power draw:	N_{GPUs} ×	0.10 kW	+	1.04 kW

Standard HLT streams

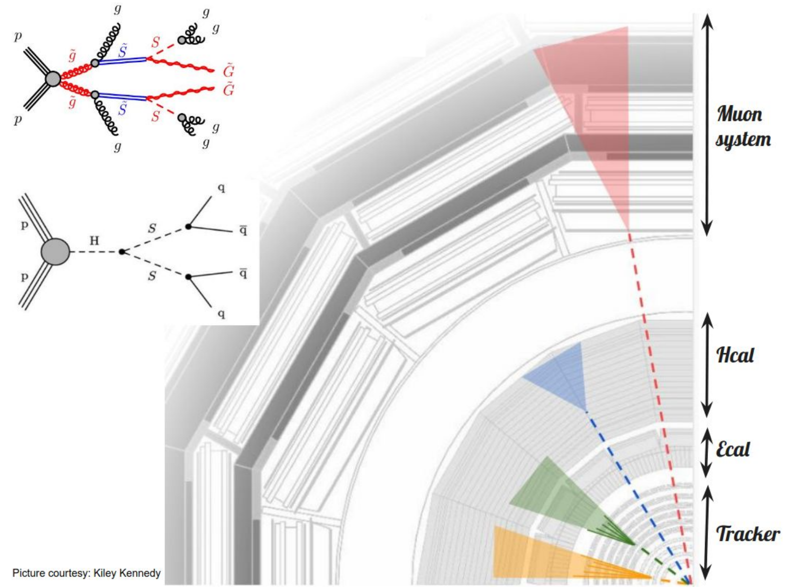
- Quick offline reconstruction (within few days), full event information
- Most of HLT paths (hundreds)
- Collect data for a wide range of CMS needs (Physics program + Alignment and Calibration)
- Physics program:
  - Generic HLT paths covering multiple physics analysis needs (broadly used, well studied, high efficiency)
  - Dedicated HLT paths for particular physics analysis that require special requirements for sufficient stats
  - Dedicated HLT paths to catch anomalies to the known physics signatures

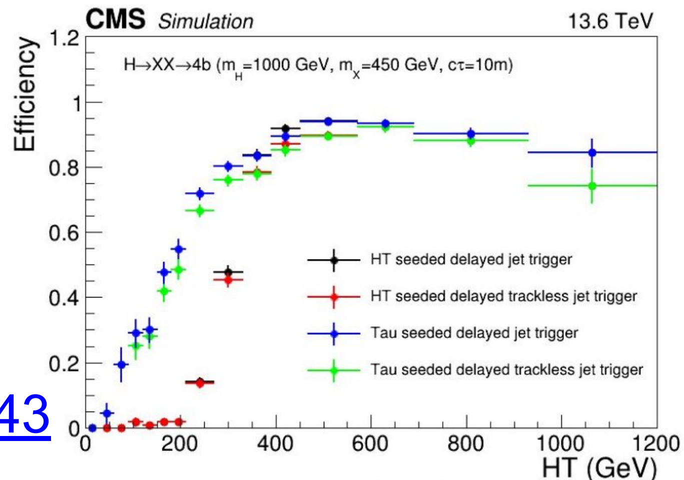


Triggering on long lived particles

- Run 3 – look for new physics. Eg. LLP.
- Several displaced-jet HLT triggers to capture various detector signatures, depending of LLP's lifetime (decay length).
  - **Tracker-based** – Reconstruct objects with non-prompt tracker-tracks seed L1 HT > 450 GeV (or Use L1 HT > 240 GeV + μ)
    - HLT jets reconstructed with displaced tracks (prompt veto) Run 3 result limits public [EXO-23-013]()
  - **ECAL-based** - Exploit timing of ECAL that measures arrival within ~200 ps seed L1 HT>430 GeV or (L1 Tau pT>120 GeV and HT>360 GeV)
    - HLT jets (nominal track match to ECAL, or ECAL only) w/ timing > 2 ns
  - **HCAL-based**
  - **Muon system-based**

[CMS DP-2023/043]()





Jets with p_T>40 GeV, number of ECAL cells>5, |η|<1.48 and jet timing>2 ns.

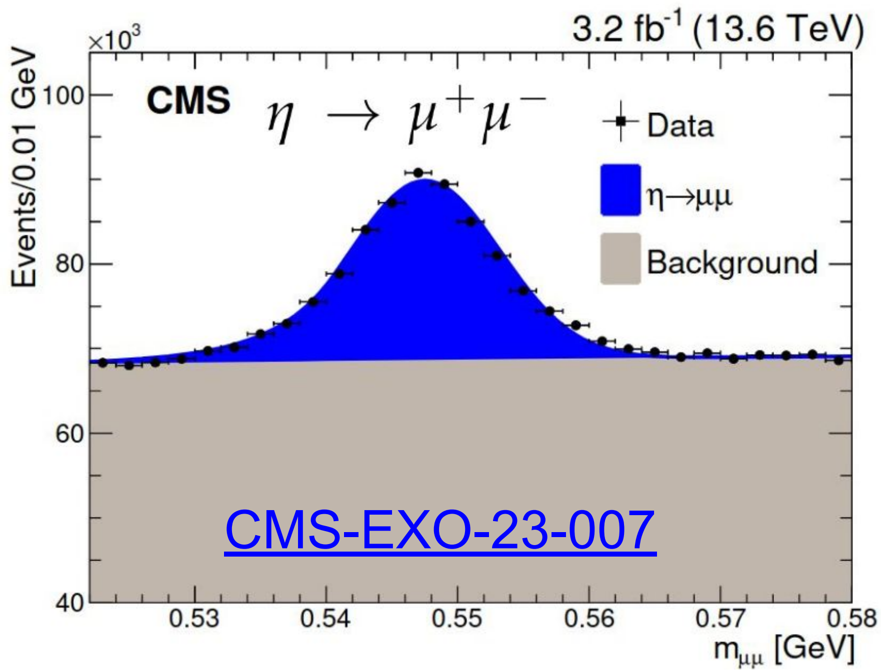
Parking HLT streams

- Delayed offline reconstruction (Opportunistically prompt)
  - Full event information…. no double copy of files
  - for physics analysis that need special triggers with rates that don't fit in Standard stream bandwidth
- Stream content is flexible and adjusted to actual physics needs
  - Current CMS priorities are signatures of LLP, di-Higgs, VBF+X and VBS+X process, Flavour-Physics
- Novel triggers for Run 3 or standard triggers (Run 2) but with lower thresholds

- Exceptionally rich B-Physics program with low pT muon and electron triggers
  - Various searches for LFU violation are being considered: measuring R(D*), searching for LFV in tag-side

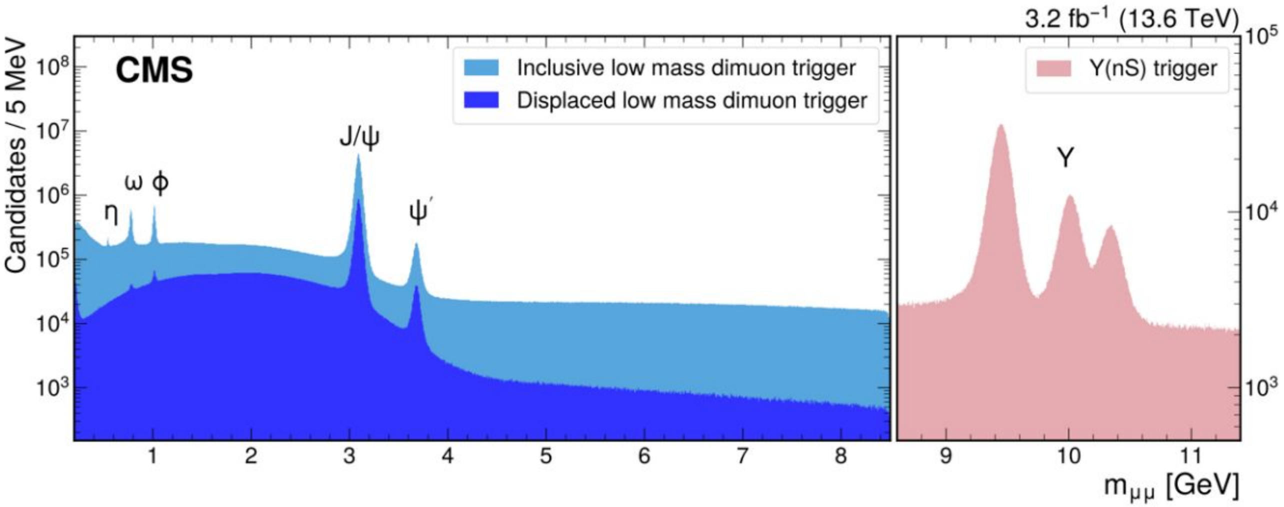
$D^0 \to \mu^+\mu^-$ $B^+ \to K^+e^+e^-$

$B^0_s \to \mu^+\mu^-$ $B^0 \to J/\psi K^0_S$



B-Physics Trigger

- In Run-3 we increased the rate of B-physics triggers using delayed reconstruction from ~300 Hz to 1.6 kHz (2022).
  - New inclusive dimuon trigger in mass range.
- In 2024, single muon parking as well
- New soft di-electron trigger
  - Improved soft electron reconstruction
- Single displaced muon active in 2018 and 2022.

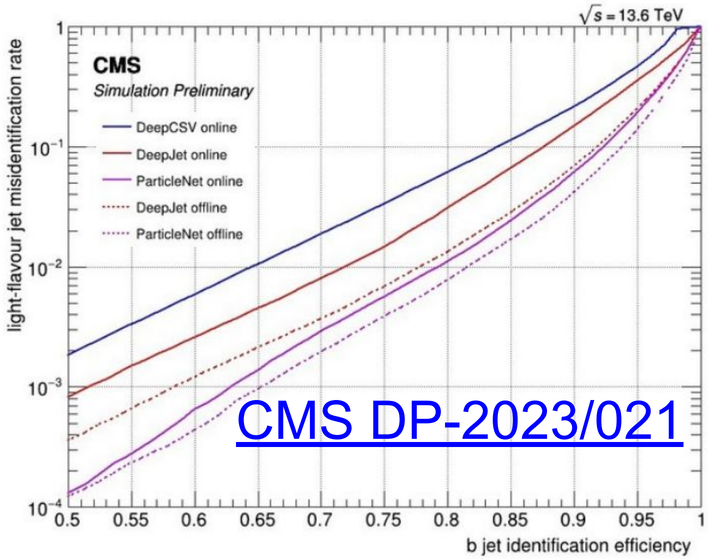


Mass distribution for pairs of μ's oppositely charged, originating from a common vertex (inclusive & displaced). Improved L1 (Kalman) and HLT (GPU)

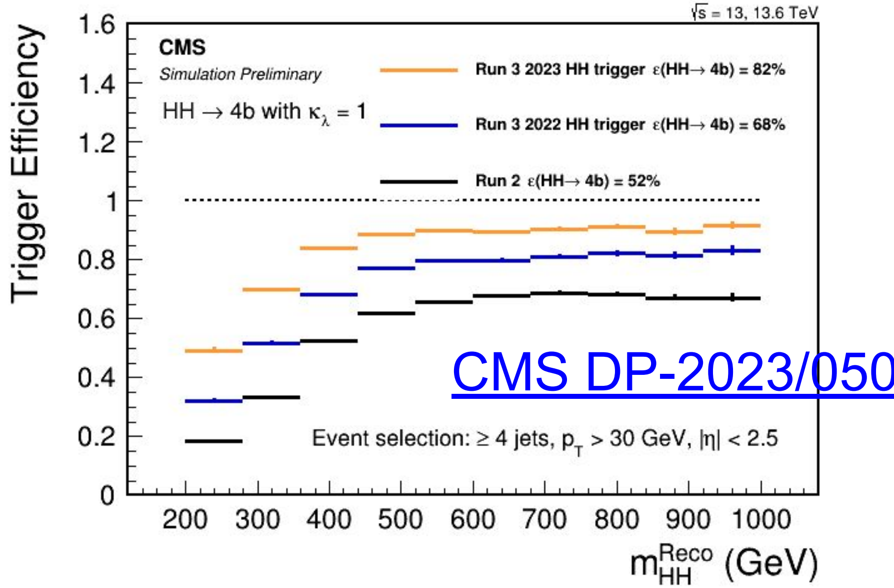
From B parking in 2022: [CMS-EXO-23-007]

HH Triggers

- New b-tagging algorithm based on graph net (ParticleNet) integrated at HLT
  - Excellent performance both offline and at HLT



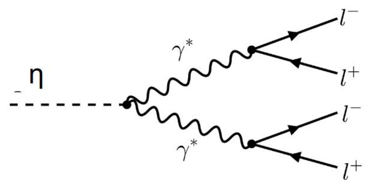
- Large efficiency increase for HH → 4b
- Further increase in 2023 using delayed reconstruction and new L1 seed

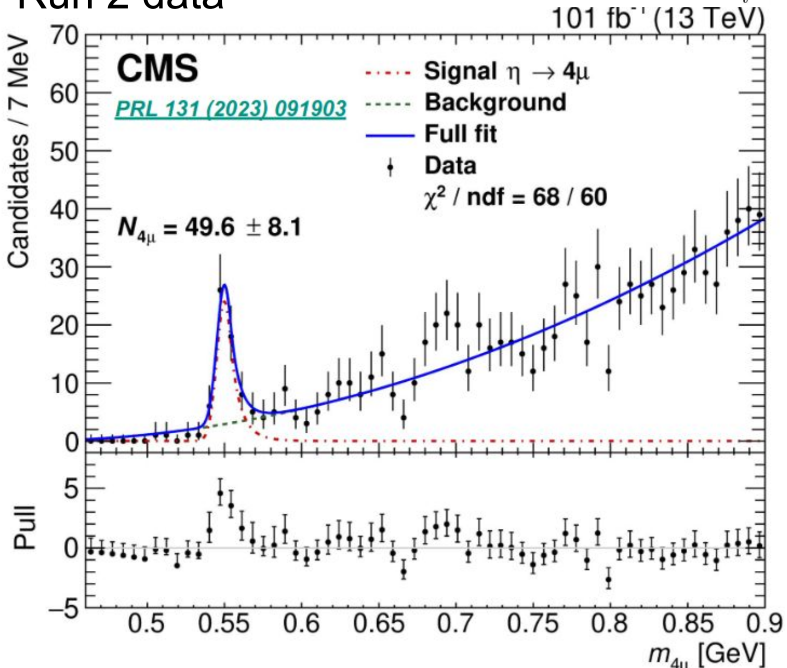


Scouting HLT streams

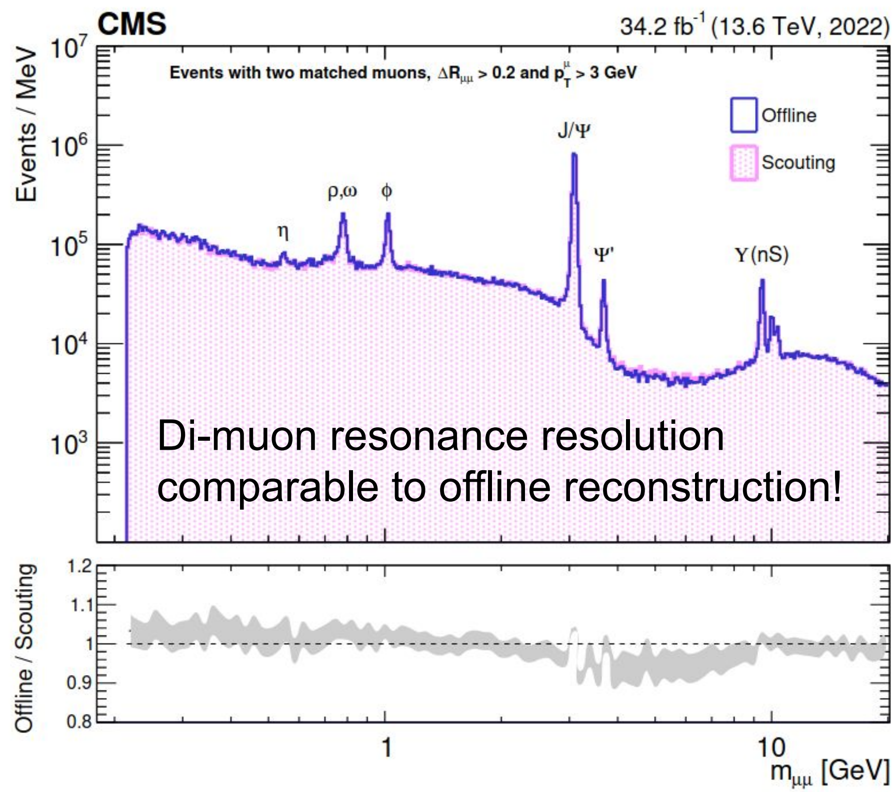
- Improvements in HLT reconstruction (use of GPUs) allowed for improved scouting strategy in Run 3
  - PF algorithm w/ tracker tracks seed by algorithm offloaded to GPUs
  - Run 3 scouting rate ~ 30 kHz

First observation of η meson decaying into four muons in Run 2 data



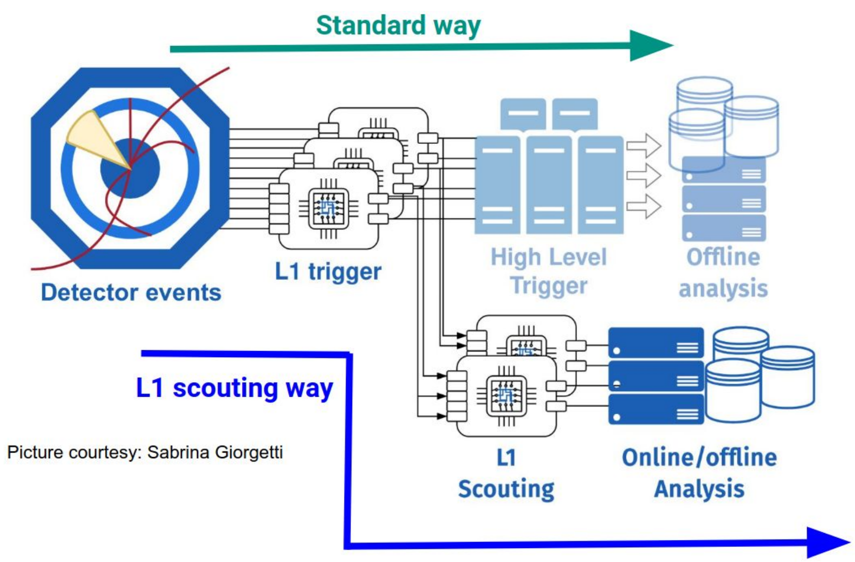


[CMS-EXO-23-007](CMS-EXO-23-007)



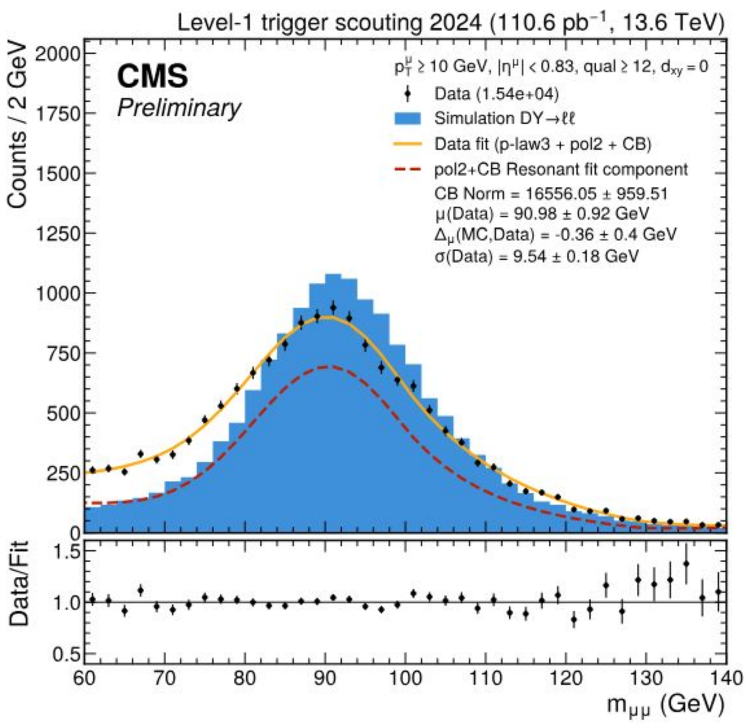
[CMS-EXO-23-007](CMS-EXO-23-007)

L1T Scouting (40 MHz Scouting)



Picture courtesy: Sabrina Giorgetti

- Idea: Store trigger-less data with limited resolution before L1 decision.
- L1 trigger data Scouting is being developed for high-lumi LHC.
- A demonstrator has been operational since the start of Run 3.

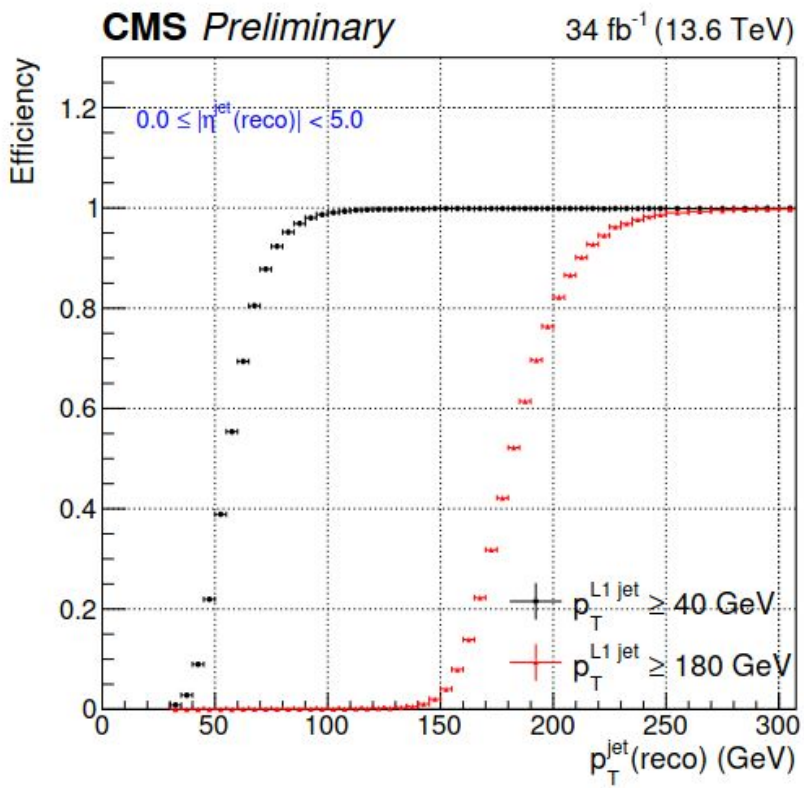


[CMS DP -2024/056]

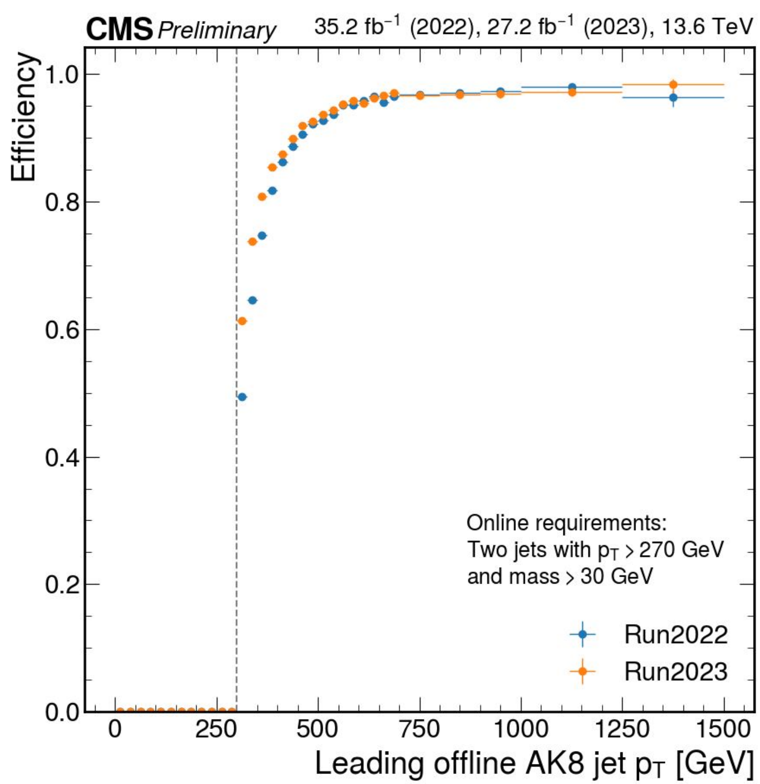
- Standard L1 rejects 99.75% events.
  - L1 scouting will allow CMS to have a look at those events.
- Tremendous capability. Enables studies of otherwise inaccessible region of phase space.
  - Next step: Properly identify all potential signatures unreachable through standard trigger and let L1 scout those events.

Jets at HLT performance

- Good jet performance in Run-3
- New boosted algorithm at HLT:
  - “trimmed mass“ → “soft-drop mass“



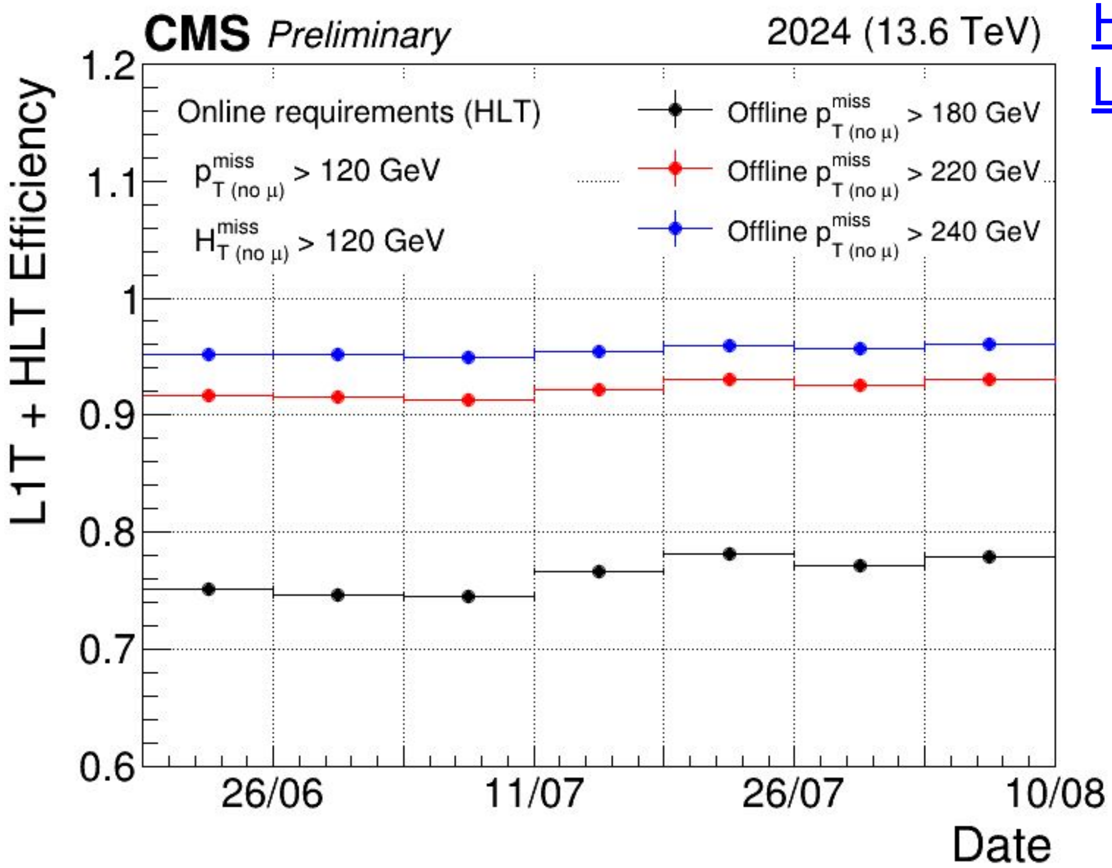
CMS DP-2023/054



CMS DP-2023/094

MET at HLT performance

- Improved pileup subtraction at L1 trigger
- Smaller rate at fixed trigger efficiency
- Excellent stability as a function of time

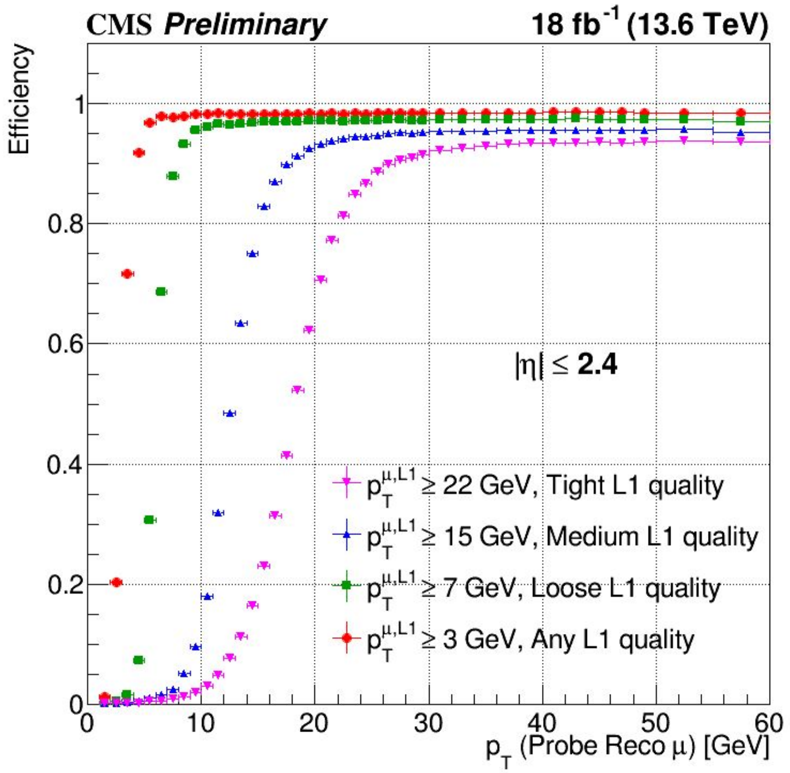


[HLT Highlights for LHCC 159]()

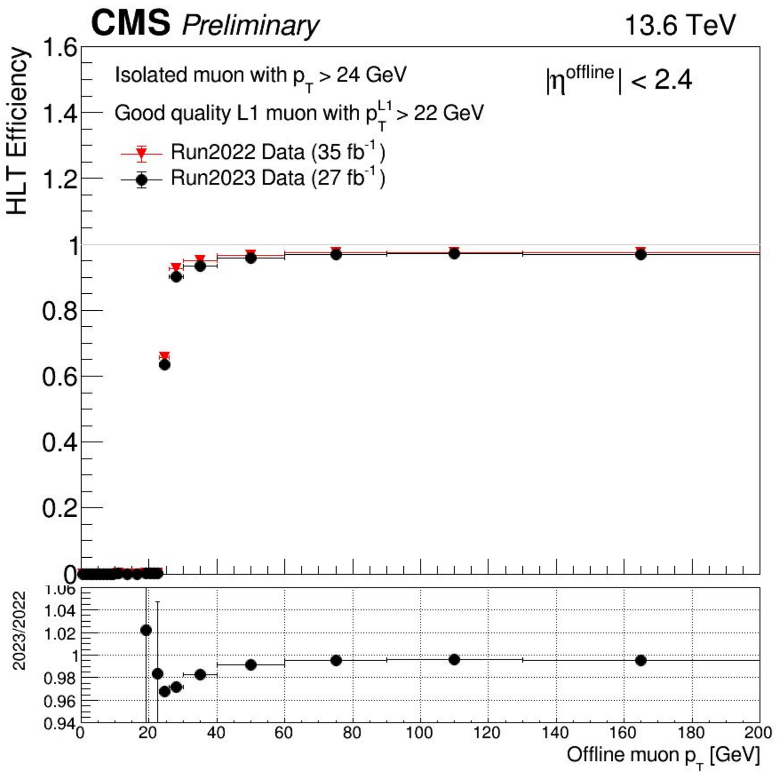
- Time evolution of L1+HLT Efficiency of the lowest unprescaled MET/MHT(NoMu) trigger during.
- The efficiency is measured for different working points of offline PUPPI Type1 MET(NoMu) selection for capturing information from both turn-on and plateau.
- Stable performance is shown in all cases. The efficiency is measured on events with real MET by the orthogonal method using the Muon dataset.

Muons performance at HLT

- Excellent muon efficiency both at L1 and HLT.
- ML technique used to select seeds in HLT muon reconstruction
  - speed up of +16% in the full HLT reconstruction.



[CMS DP-2023/057]

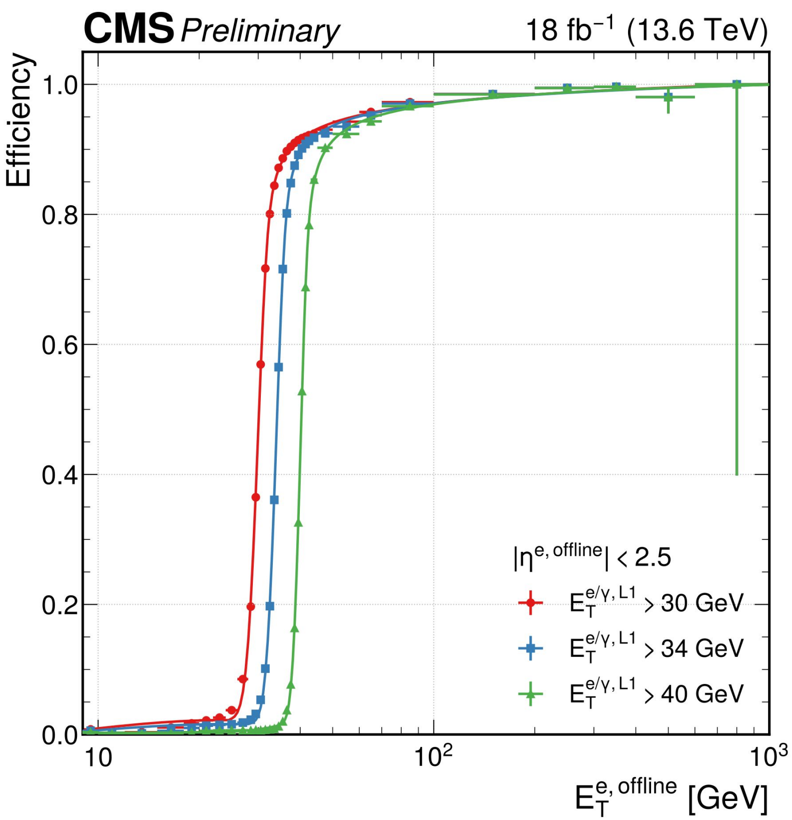


[CMS DP-2024/05]

Electromagnetic objects performance

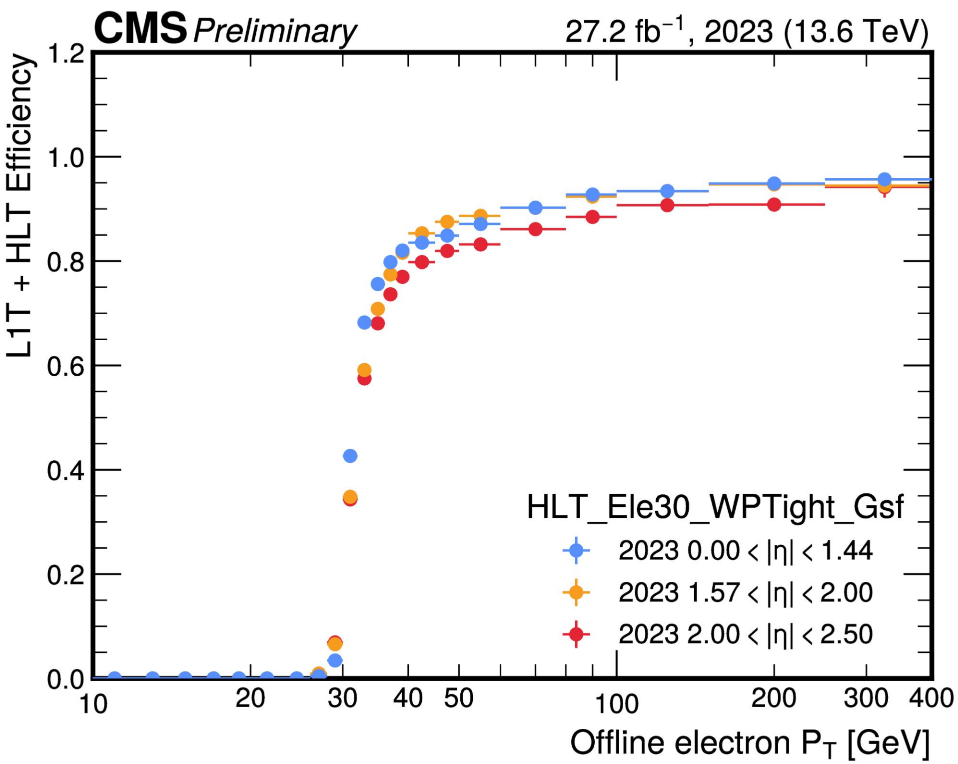
- Very good performance of e/γ objects

L1 Electron/photon efficiency



CMS DP-2023/055

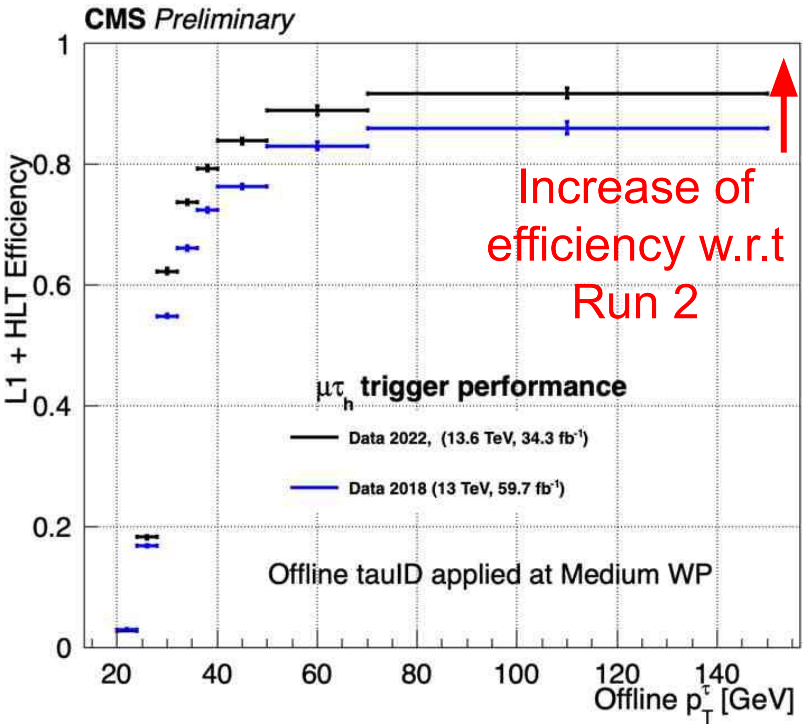
L1+HLT Electron efficiency



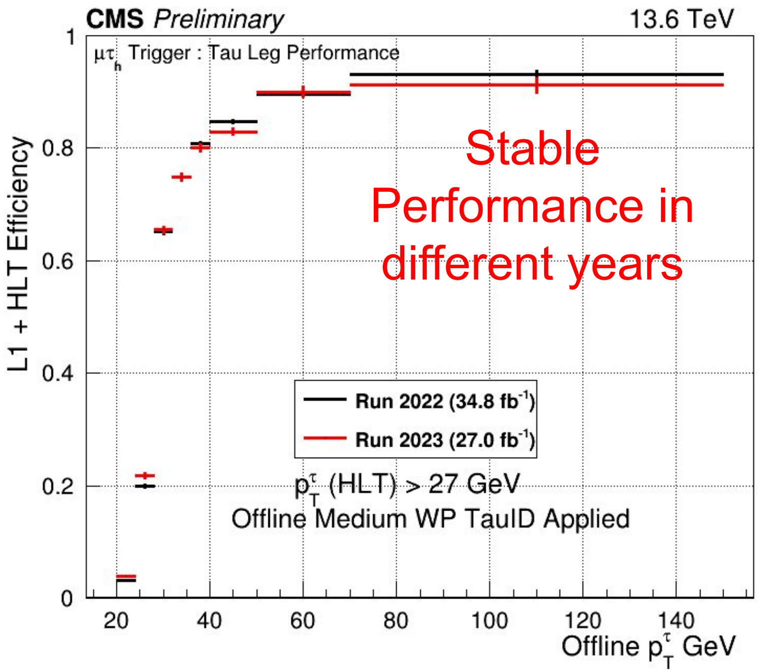
CMS DP-2024/041

Tau peformance at HLT

- New Tau reconstruction at HLT based on Convolutional Neural Network (DeepTau)
  - Faster reconstruction and better performance.
- In 2024 transitioned to novel PNet Tau model



[CMS DP-2023/024]

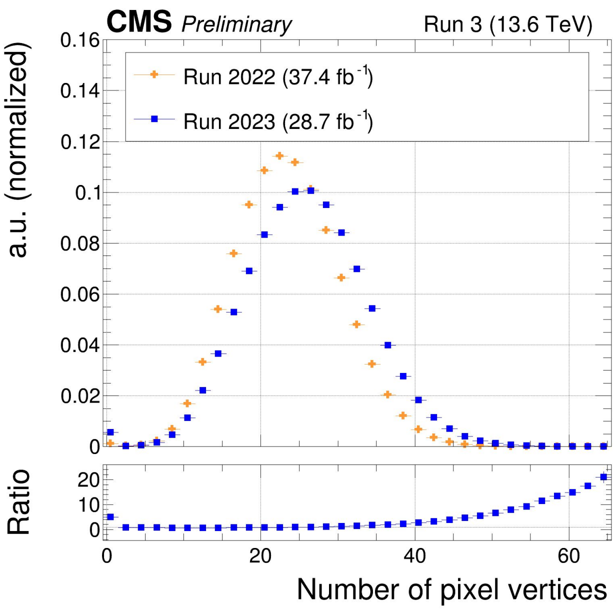


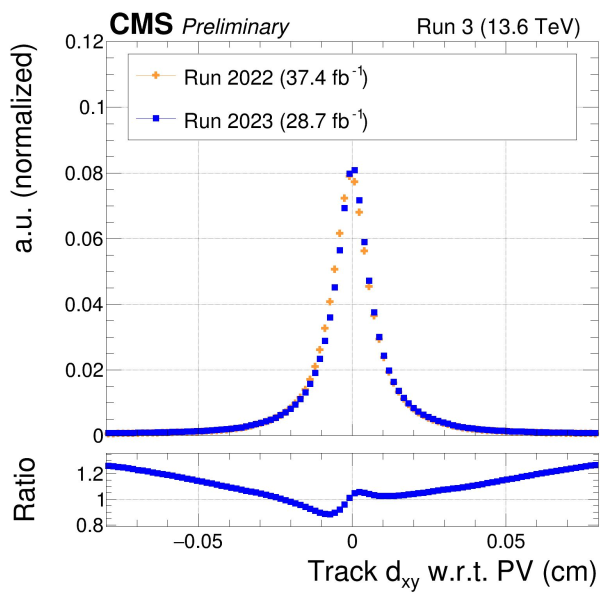
[CMS DP-2024/042]

Tracking and Vertexing performance at HLT

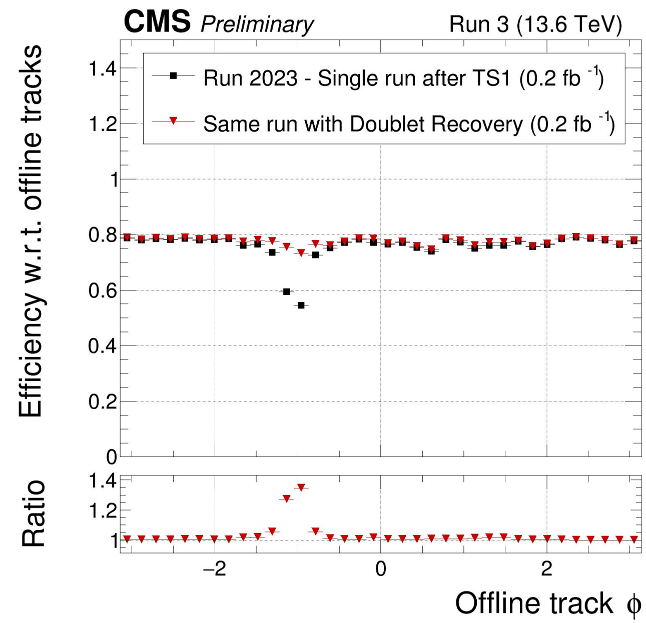
- Tracking based on single iteration in 2022 and 2023.
  - Pixel tracking and vertexing running on GPUs
- In 2024 additional tracking iteration (“doublet recovery”) included to cope with pixel detector failures localized in η,φ

[CMS DP-2024/013]





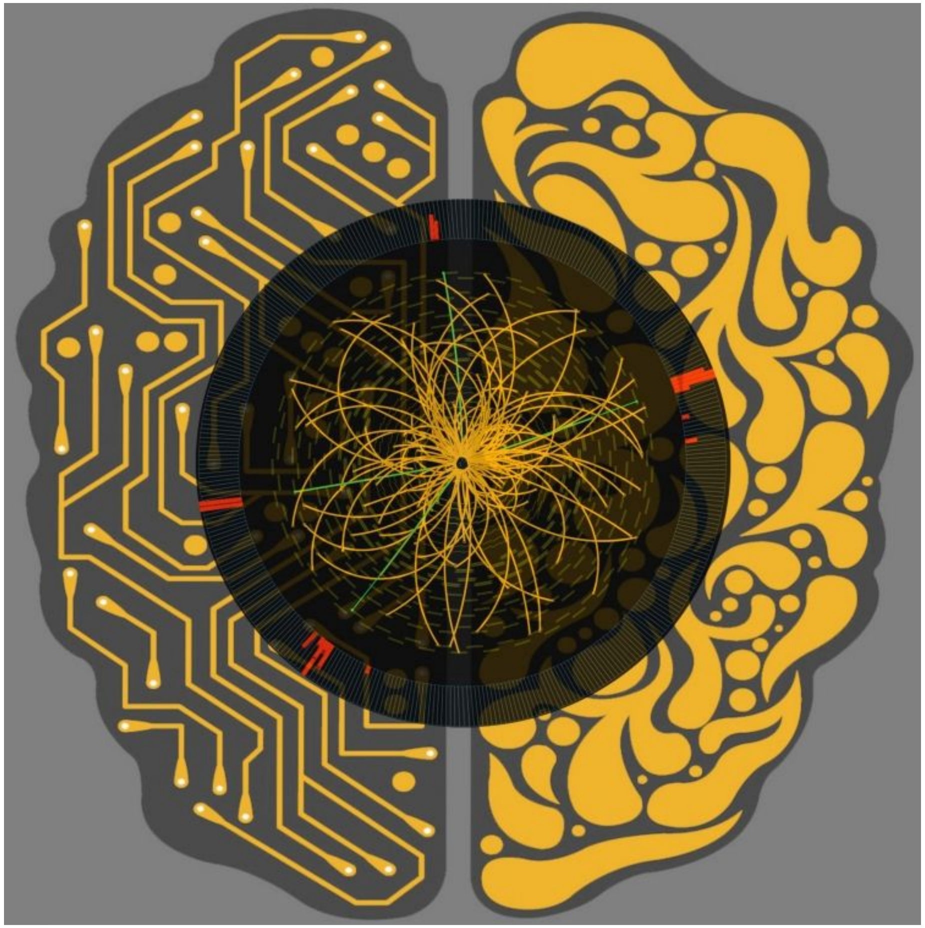
Comparison of performance of 2023 data-taking w.r.t 2022



Tracking efficiency with respect to offline tracks as a function of the offline track azimuthal angle φ for a run reconstructed at HLT without (black) and with (red) the doublet recovery iteration.

Machine Learning at the Trigger level

- ML is an essential and versatile tool that we use
  - to improve existing approaches
  - to enable new approaches

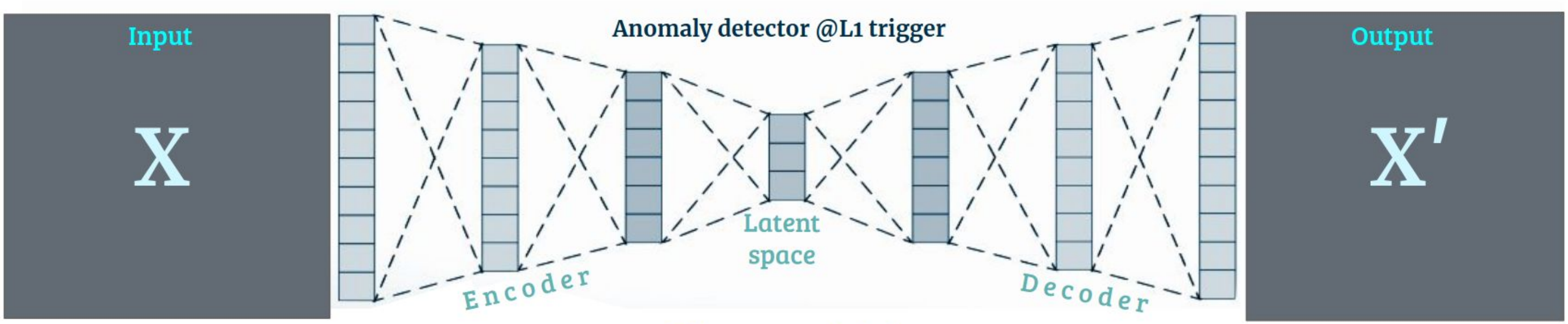


https://cms.cern/news/cms-releases-open-data-machine-learning

The unknown-unknown territory: how to approach it?

- If we knew the exact signature we are looking for, we'd build a trigger for it!
- In absence of that, what else can we do?

- Use of ML to learn the features of typical standard model events
- Then, pick events that are not typical, using autoencoder (AE)
- Train AE on typical events ("ZeroBias" data) and use reconstruction error (loss) as a metric for anomalous-ness



$$\mathcal{L} = || X - X' ||$$

Anomaly detector @L1 trigger in CMS

AXOL1TL

Anomaly eXtraction Online Level-1 Trigger aLgorithm

Inputs:

- pT, η, ϕ of Jets(x10) , e/γ (x4),
- μ (x4), and MET (from Calo layer-2 and Global Muon Trigger

Ref:
https://cds.cern.ch/record/2876546

CICADA

Calorimeter Image Convolutional Anomaly Detection Algorithm

Inputs:

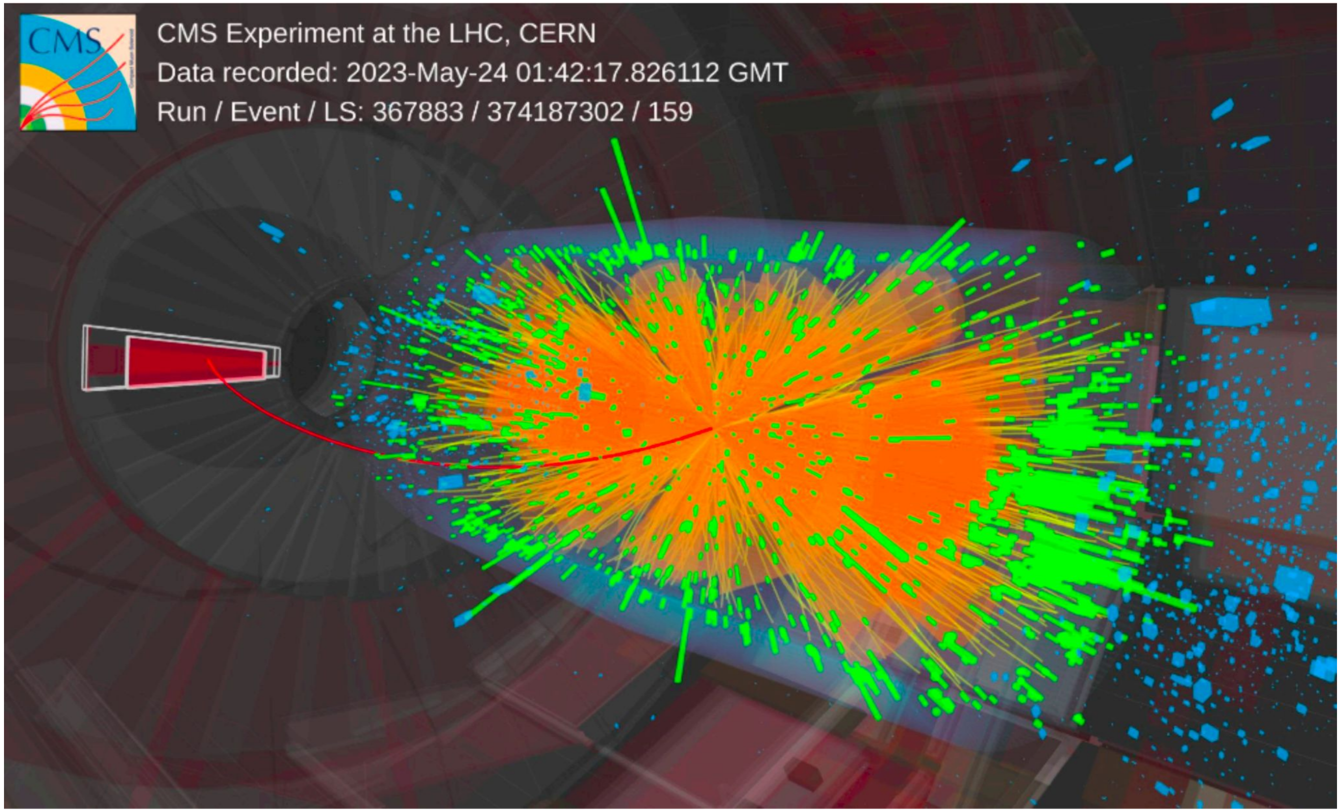
- Low-level information (from Calo layer-1) in image format.

Ref:
https://cds.cern.ch/record/2879816

ML@L1 trigger becoming important. Tools for ML@FPGA developed.

- Neural Nets → HLS4ML (documentation)
- Boosted Decision Trees → Conifer (github, paper)

An event selected by AXOL1TL



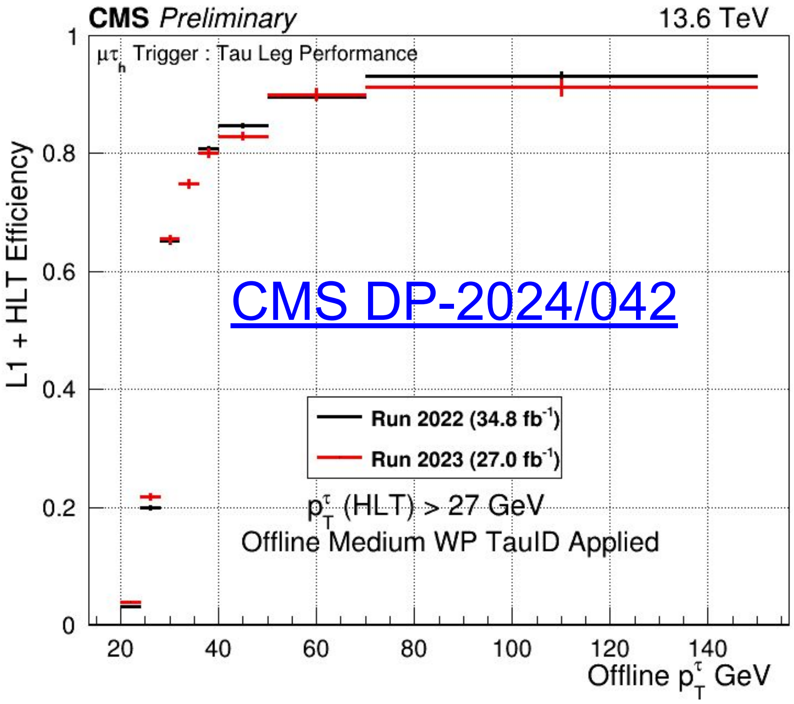
SUEP?
Emerging jet?
Or just normal QCD

Example of CMS event selected by AXOL1TL **not selected by any other CMS trigger**. Features at L1T

- 12 jets, 11 w/ pT >20 GeV
- 1 μ, pT > 3 GeV
- Large number of primary vertices (75)

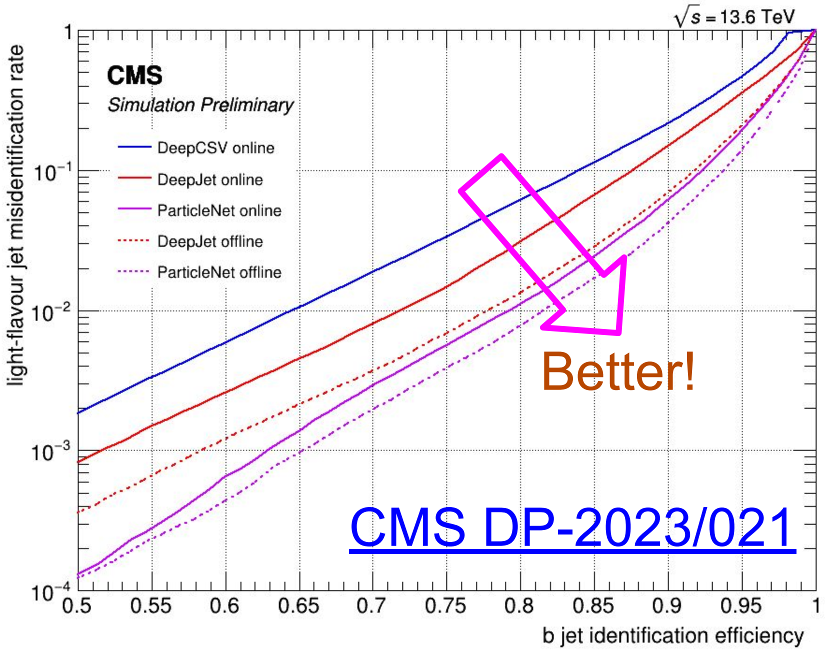
Machine Learning at the HLT

- Tau @ HLT
  - Reconstruction: Hadron plus strip
  - Identification: CNN+DNN based tagger (DeepTau)



CMS DP-2024/042

- ParticleNet b-jet tagger@HLT. GNN-based.
  - Jets treated as a permutation-invariant point cloud.
- Performance gain, specially for HH(4b), HH(2b2τ) and HHH(6b) processes, compared to Run 2



CMS DP-2023/021

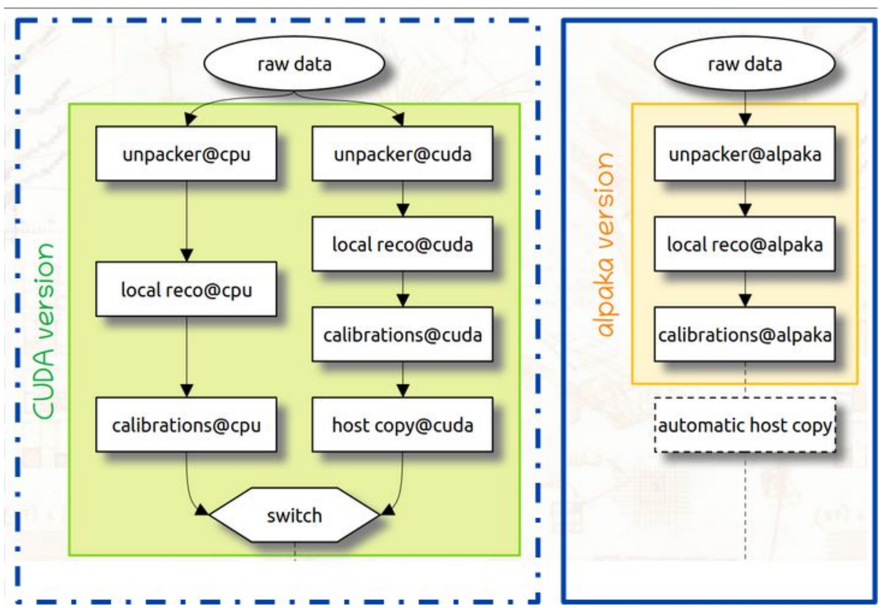
Conclusions

- Many improvements have been implemented in the CMS trigger after Run-2 both at L1 and HLT trigger
  - Trigger for long-lived particle (eg. ECAL and HCAL timing)
  - GPU at HLT → more powerful scouting
  - Large rate for flavour physics and HH (delayed reconstruction)
- The Run-3 data confirm the good performance of the L1 and HLT
- Thanks to the new triggers Run-3 is not just “a copy of Run-2” but it is an opportunity to look for New Physics in new final states.
  - Many more news will come with Phase-2 upgrades and HL-LHC (>2030)
- Onset of a New Era for CMS Trigger:
  - Leveraging mature technologies with advanced Machine Learning for anomaly detection and increasingly powerful classification tools, driving innovation in data processing.

BACKUP

Migration to Alpaka

- alpaka is a portability library. Same code able to run on
  - multiple hardware vendors (eg. AMD GPU, Intel GPU)
  - multiple kinds of accelerators (eg. GPU, FPGA)
- Pixel and ECAL and HCAL code migrated from CUDA to Alpaka in 2024.
  - Part of the Particle Flow recently ported directly to Alpaka from CPU-only.



Early Run 3: 2022 & 2023 **Run 3: ≥ 2024**