



Shaping the future CMS event selection with Next-Generation Triggers

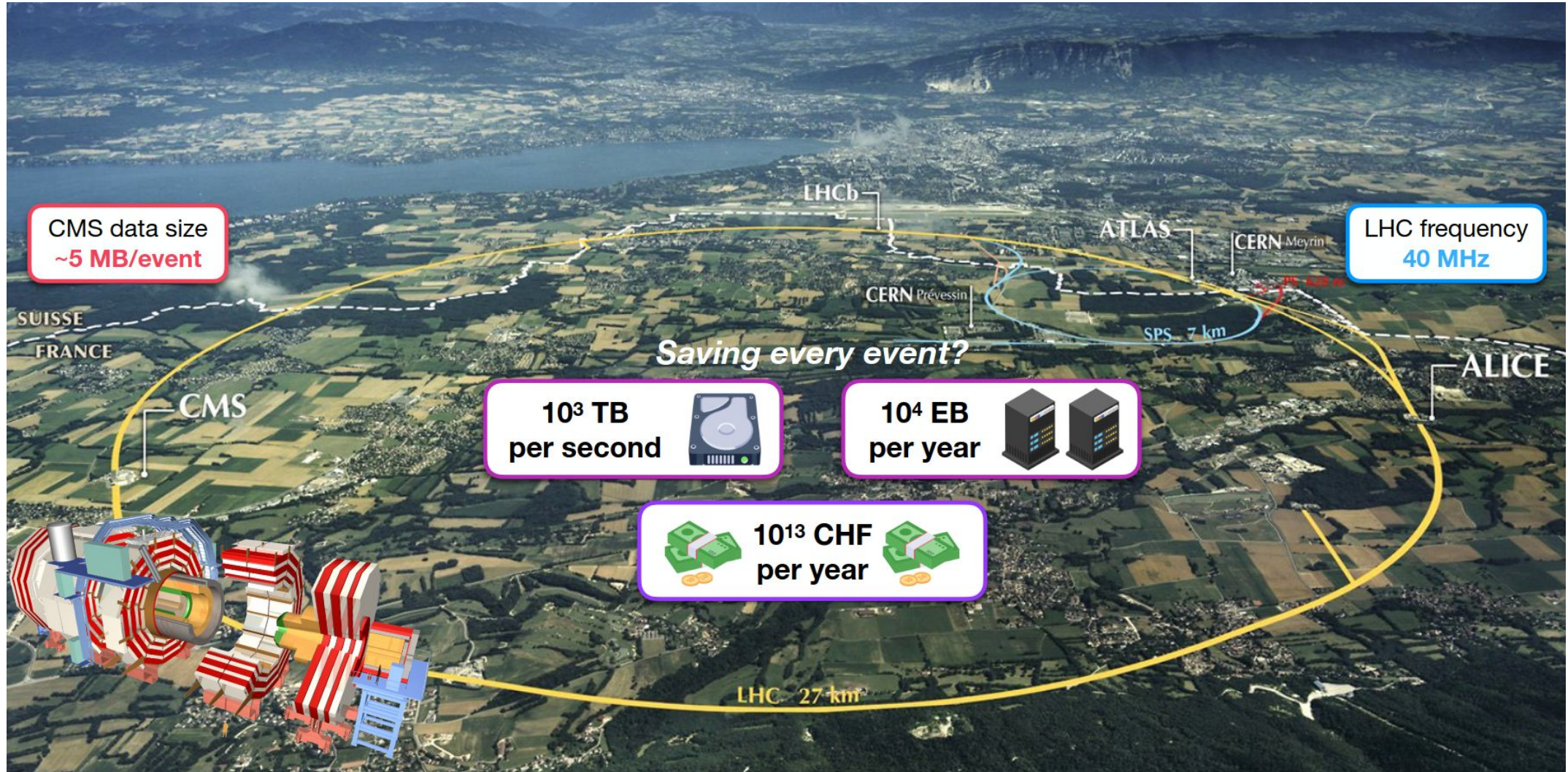
Experimental Particle Physics Seminar
INFN Padova, 9th July 2026

Marco Musich (CERN)



NexTGen
Next Generation Triggers

The CMS Experiment at the LHC



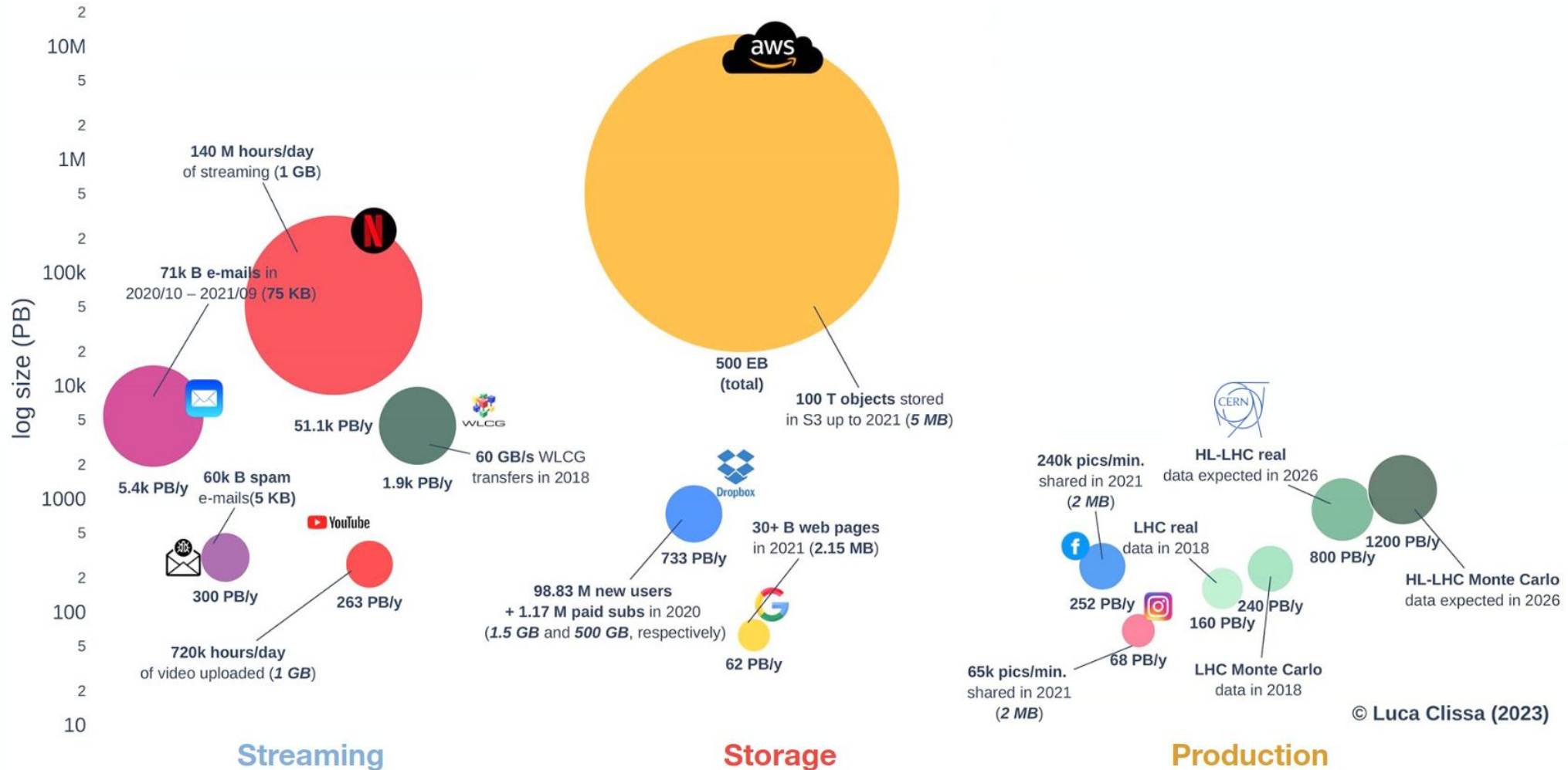
Outline of this presentation

- ❖ The LHC “Big Data”
- ❖ Triggering at the LHC
- ❖ The CMS Triggering system and its Upgrade
- ❖ **Next Generation Triggers**: its vision for the HL-LHC
- ❖ **NGT @ L1**: “A Machine Learning Revolution”
- ❖ **NGT @ HLT**: “Scouting at Full Rate”
- ❖ Conclusions & Outlook

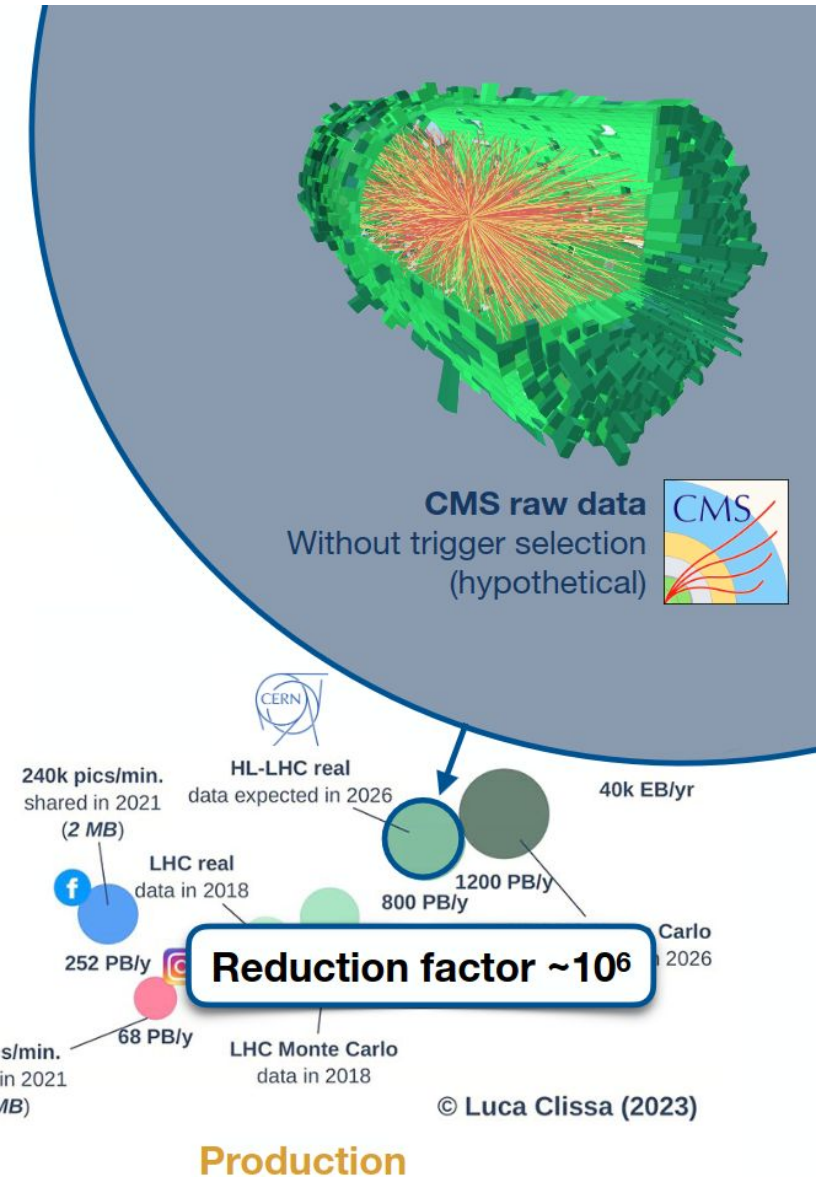
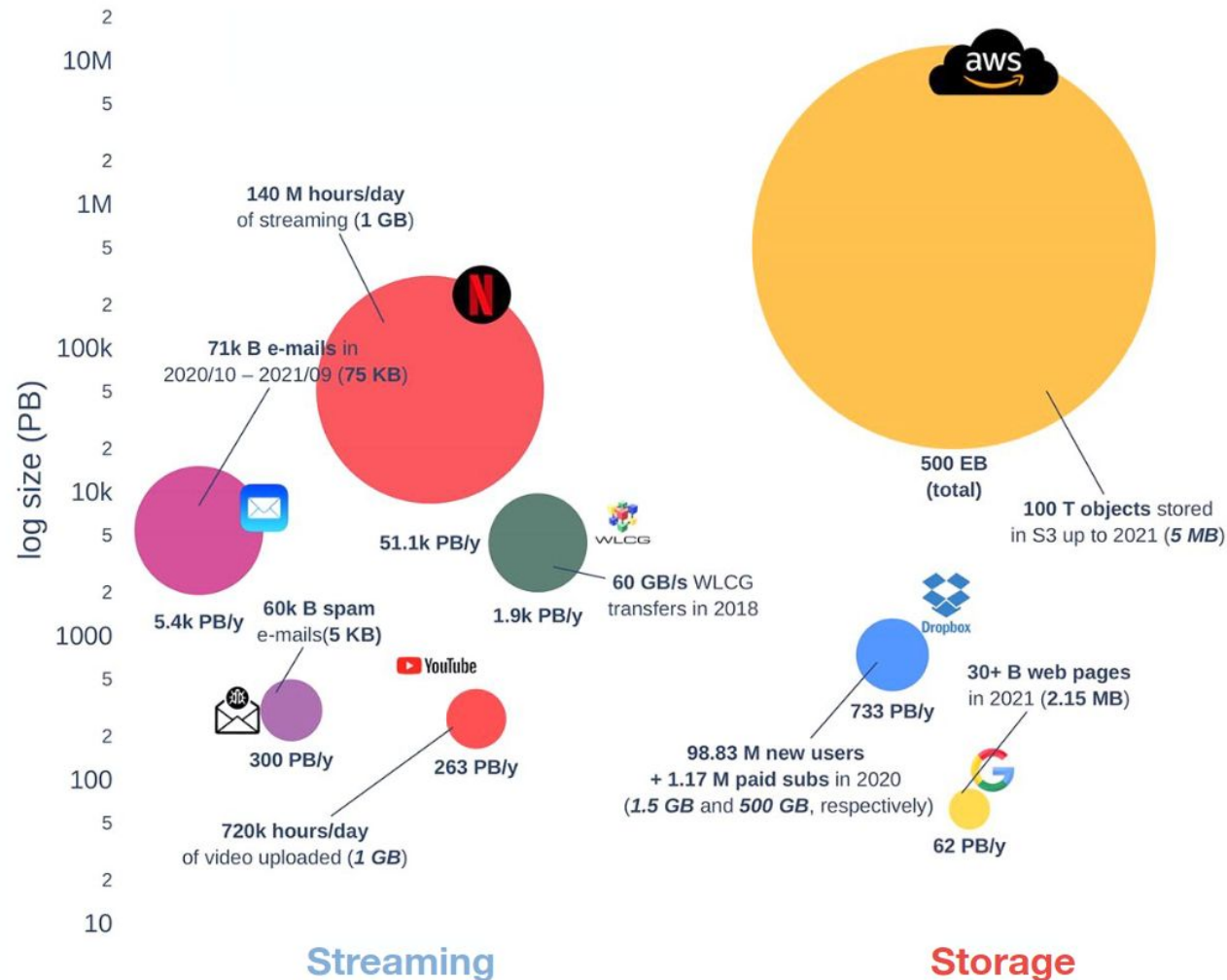
Disclaimer

This deck of slides is heavily indebted to E. Vernazza’s excellent [seminar on the same topic](#) at UZH, and S. Giorgetti and B. Alves [talk](#) at the 3rd NGT “All-hands” meeting

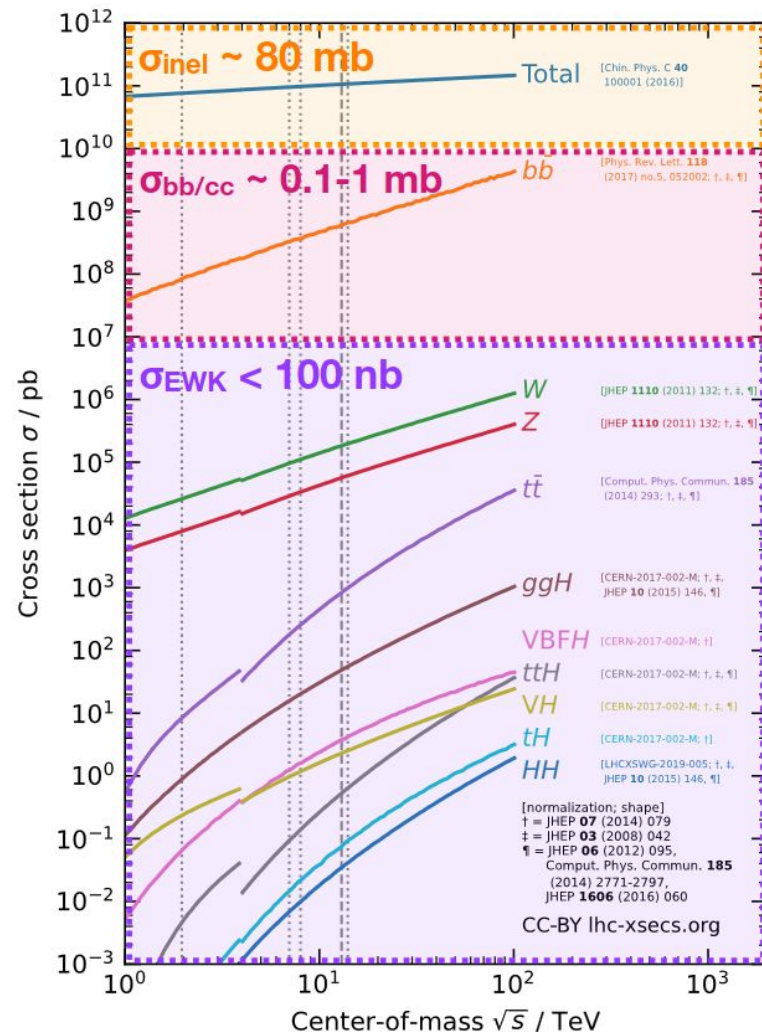
The LHC Big Data



The LHC Big Data



What Physics are we looking for?



- Study of quark-gluon plasma (QGP)
 - Heavy ion collisions
 - Quantum Chromodynamics

LHCb

- Precision flavour physics
 - CP violation
 - heavy hadrons decays



CMS & ATLAS



- Electroweak and Higgs sector
 - SM precision measurements
 - search for new physics

Triggering physics at the LHC



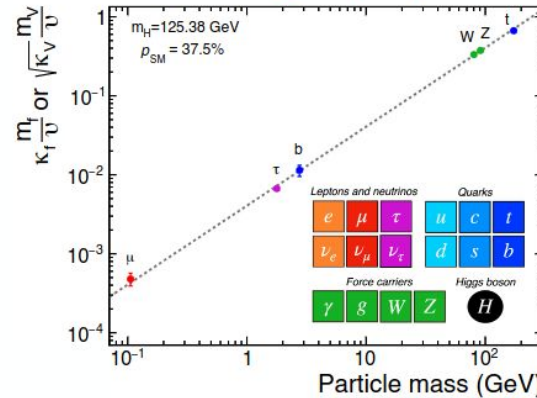
4 July 2012 Higgs boson discovery by CMS and ATLAS Collaborations



Trigger built as a **compromise**

- technological limitations
- theory motivating new physics at O(100-1000) GeV

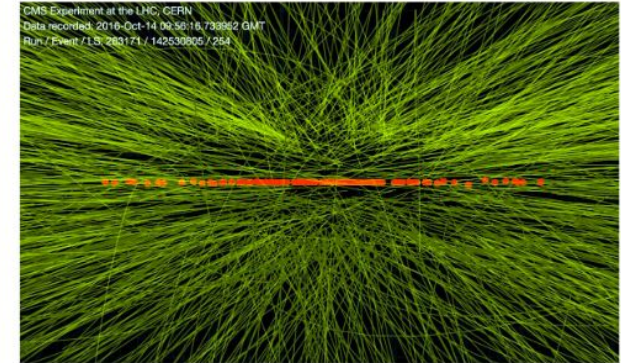
LHC Run 2 — LHC Run 3 Precision measurement of the SM and Higgs properties



Trigger acceptance **enhanced** for new physics (new hardware)

- Data Scouting
- Data Parking

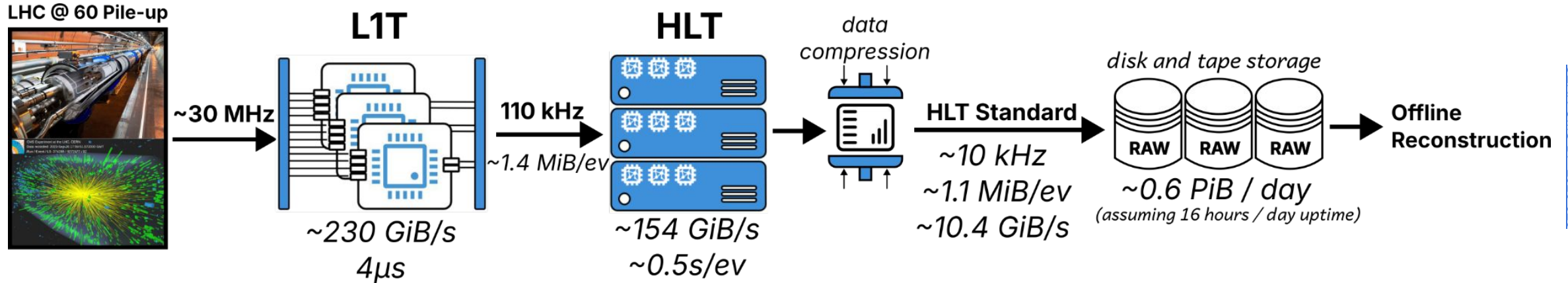
HL-LHC Major detector upgrade for High-Luminosity LHC



Major **upgrade** of trigger systems to exploit higher luminosity

- from 60 to 200 PU interactions
- enable same physics program

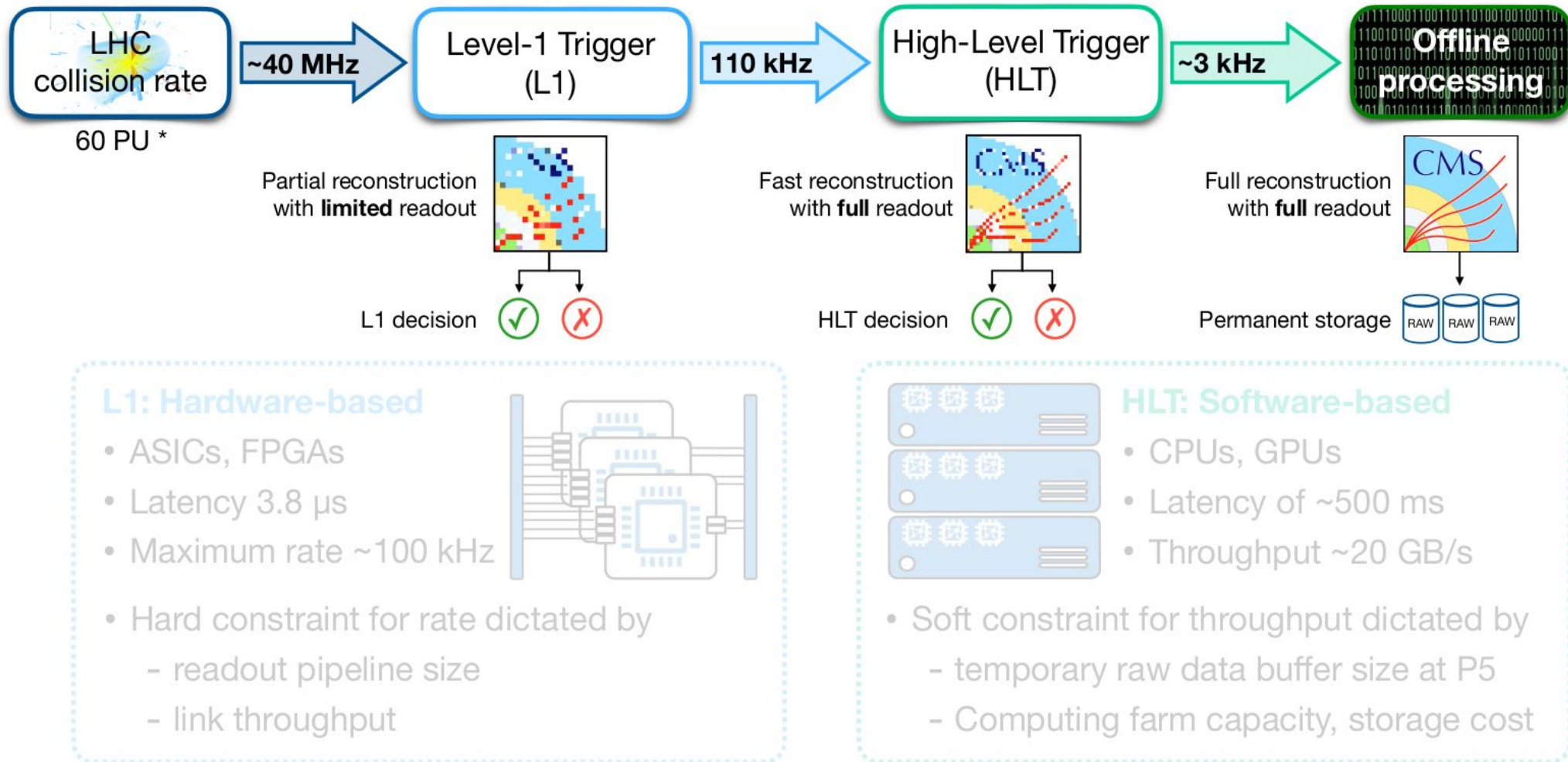
The Original CMS Trigger system



CERN-CMS-DP-2026-024
CERN-CMS-DP-2026-023

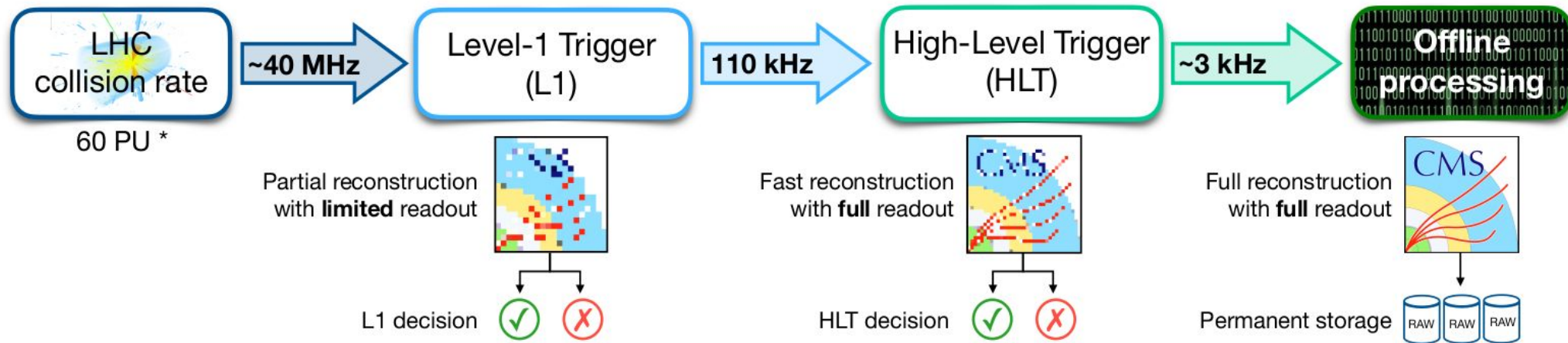
- **CMS two-tiered trigger design:** optimized tradeoff between technological limitations and physics exploration
- Enables **efficient selection of rare physics events** while reducing the overwhelming event rate: **98%** events rejected at **Level-1 Trigger (L1T)**, **99%** rejected at **High-Level Trigger (HLT)**.

The CMS Trigger original design



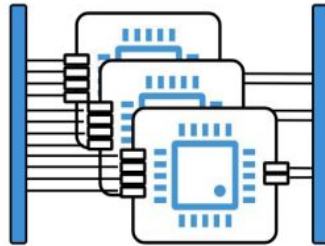
* LHC Run-3 conditions

The CMS Trigger original design (L1)



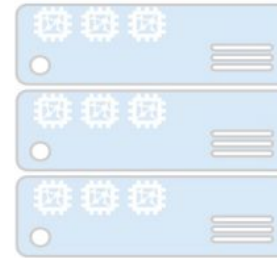
L1: Hardware-based

- ASICs, FPGAs
- Latency 3.8 μ s
- Maximum rate ~100 kHz
- Hard constraint for rate dictated by
 - readout pipeline size
 - link throughput



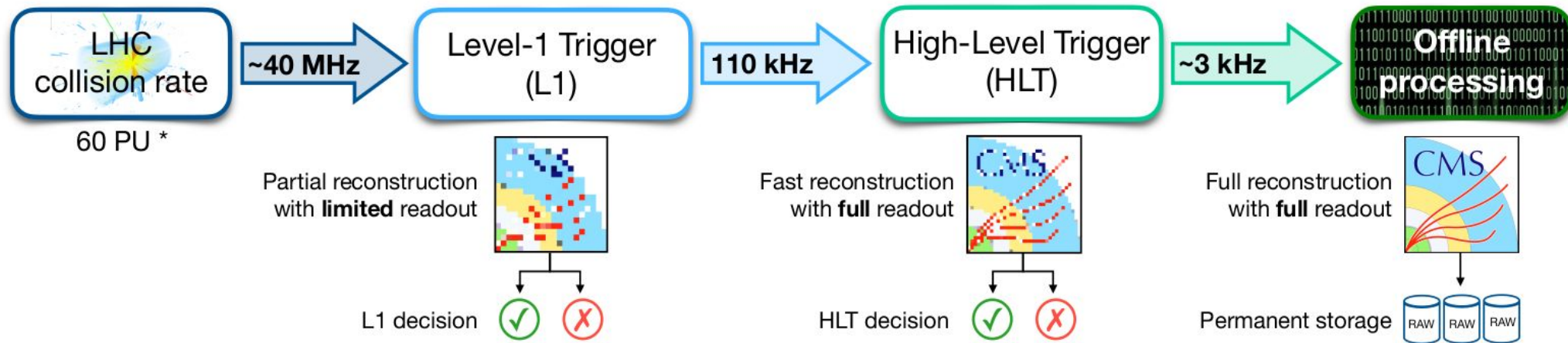
HLT: Software-based

- CPUs, GPUs
- Latency of ~500 ms
- Throughput ~20 GB/s
- Soft constraint for throughput dictated by
 - temporary raw data buffer size at P5
 - Computing farm capacity, storage cost



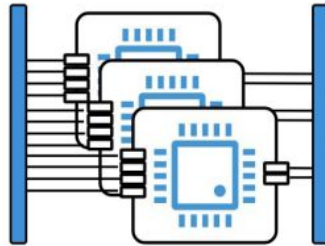
* LHC Run-3 conditions

The CMS Trigger original design (HLT)

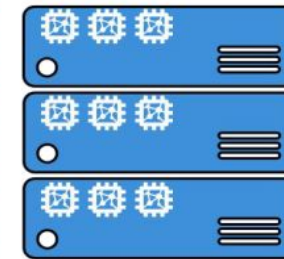


L1: Hardware-based

- ASICs, FPGAs
- Latency 3.8 μ s
- Maximum rate ~100 kHz
- Hard constraint for rate dictated by
 - readout pipeline size
 - link throughput



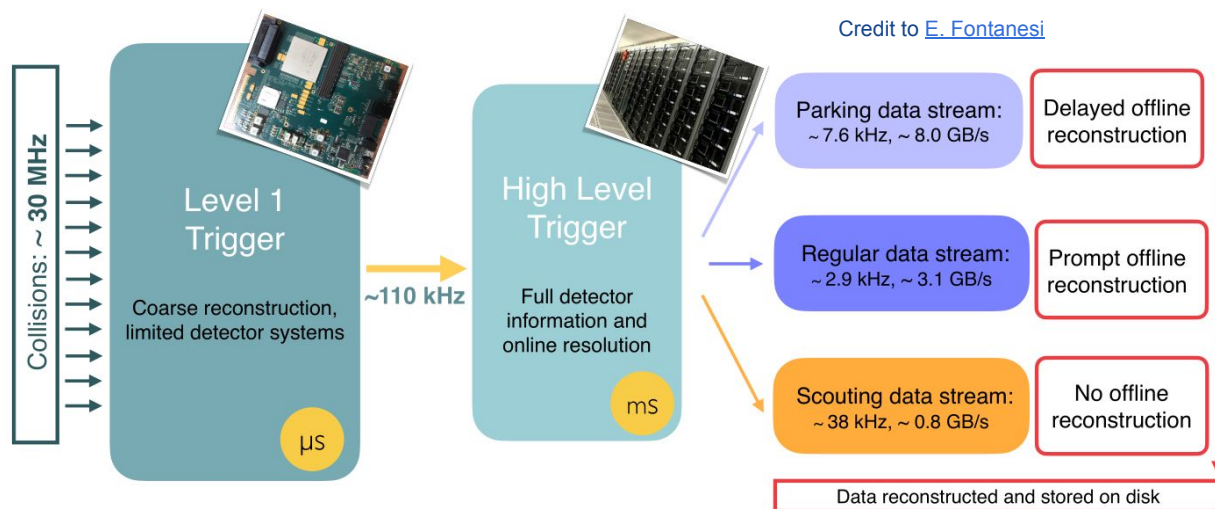
HLT: Software-based



- CPUs, GPUs
- Latency of ~500 ms
- Throughput ~20 GB/s
- Soft constraint for throughput dictated by
 - temporary raw data buffer size at P5
 - Computing farm capacity, storage cost

* LHC Run-3 conditions

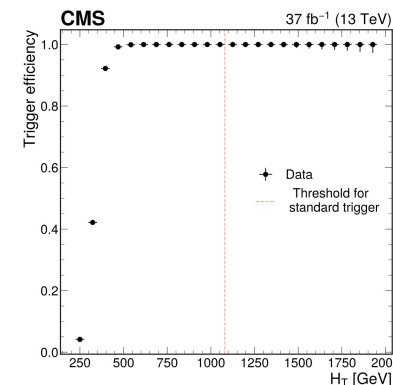
Extending the CMS Trigger system



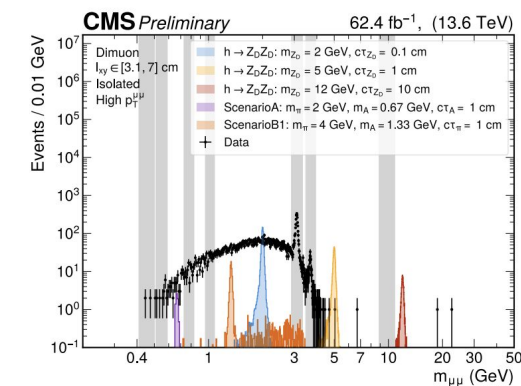
- **These strategies leverage on heterogeneous computing @ HLT**
 - Flexible systems with CPUs and GPUs for timing speed-ups
 - Bring lower cost and lower power consumption
- The new streams enlarged the acceptance for **low mass resonances** (di- μ , di-jets), **B-physics, VBF, Long-Lived particles and HH production**

Extend the acceptance with novel data streams

- **Parking:** increase the processed data size and delay its reconstruction for when resources are available.
- **Scouting:** select data lowering trigger thresholds and run an online analysis with reduced event information.



Lower trigger thresholds: from 1050 to 410 GeV
in H_T enhancing hadronic final state searches.



Enhanced searches for displaced dimuon resonances at low mass.

CMS-EXO-23-007, CMS-PAS-EXO-24-016

The CMS Upgrade for the HL-LHC

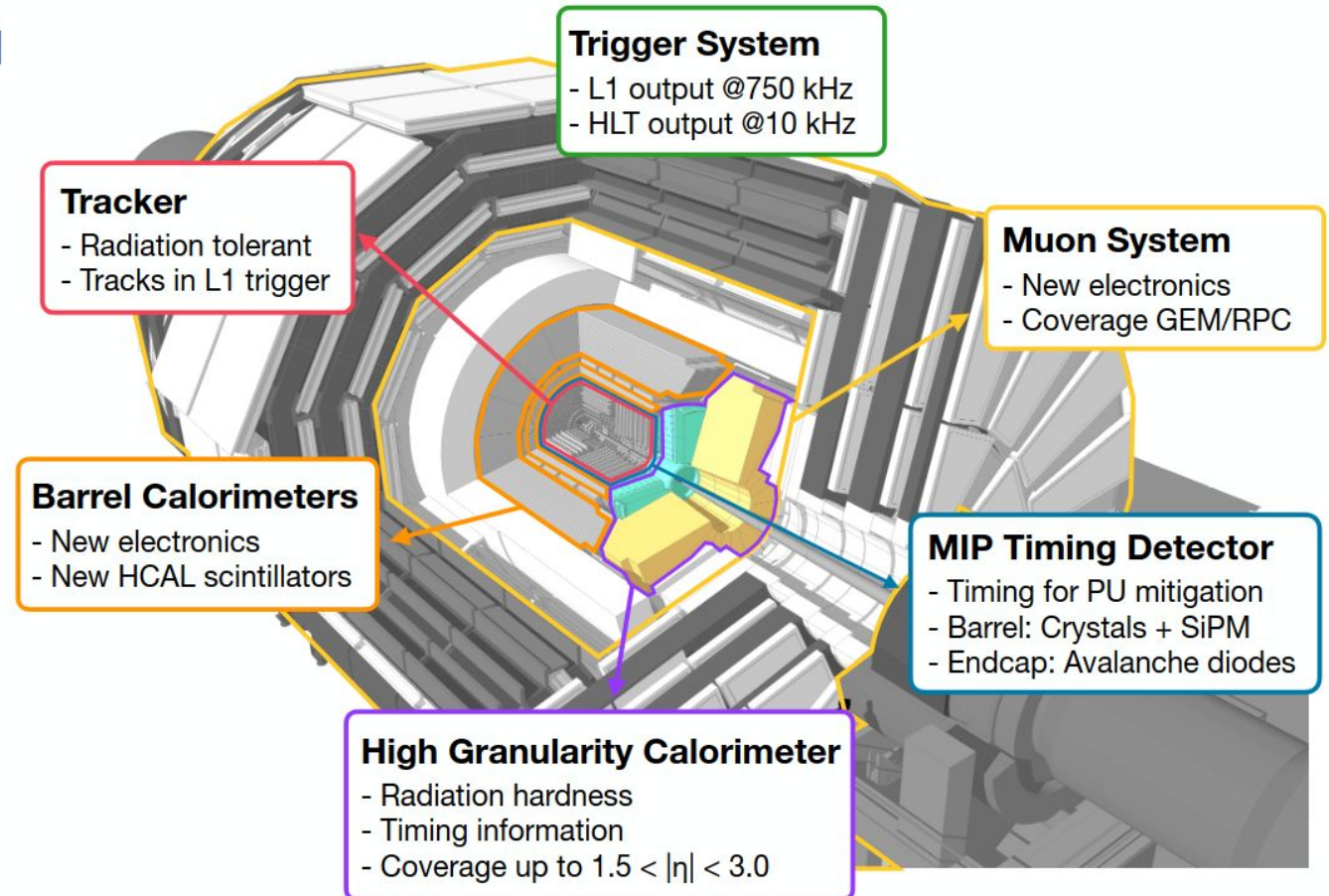
- The High-Luminosity LHC will pose several **challenges** for the CMS detector
 - More **simultaneous collisions**: ~200 PU
 - High **radiation levels**: $O(100)$ Mrad

Sub-detector upgrades

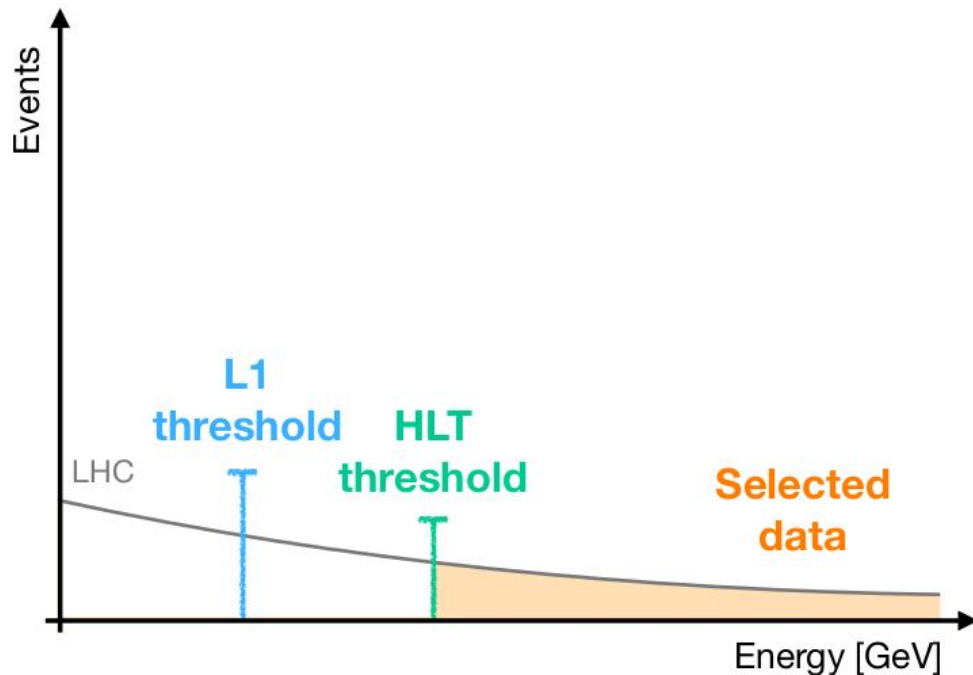
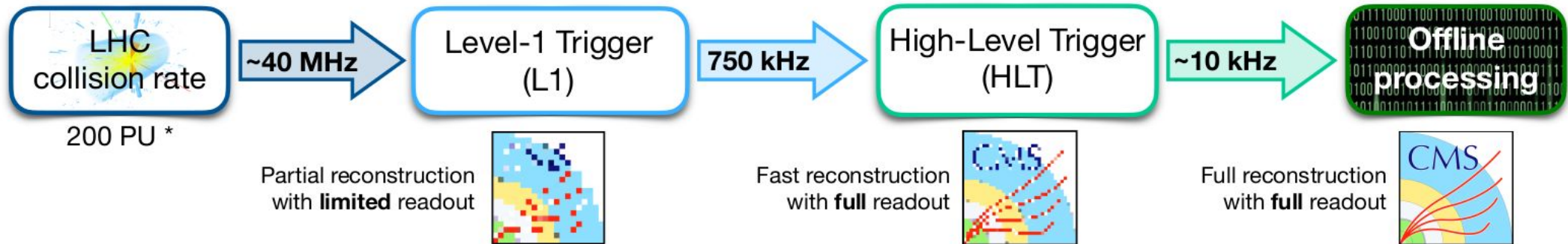
- **Granularity and Timing**
to separate different interaction vertices
- **Radiation hardness**
to sustain unprecedented dose

Trigger upgrades

- **Faster Processing**
to fit within available latency
- **Higher Computational Power**
to reconstruct all particles



Towards an upgraded trigger



Main Goal : Maintain **same physics acceptance** moving from 60 to 200 PU



More **powerful** trigger ...

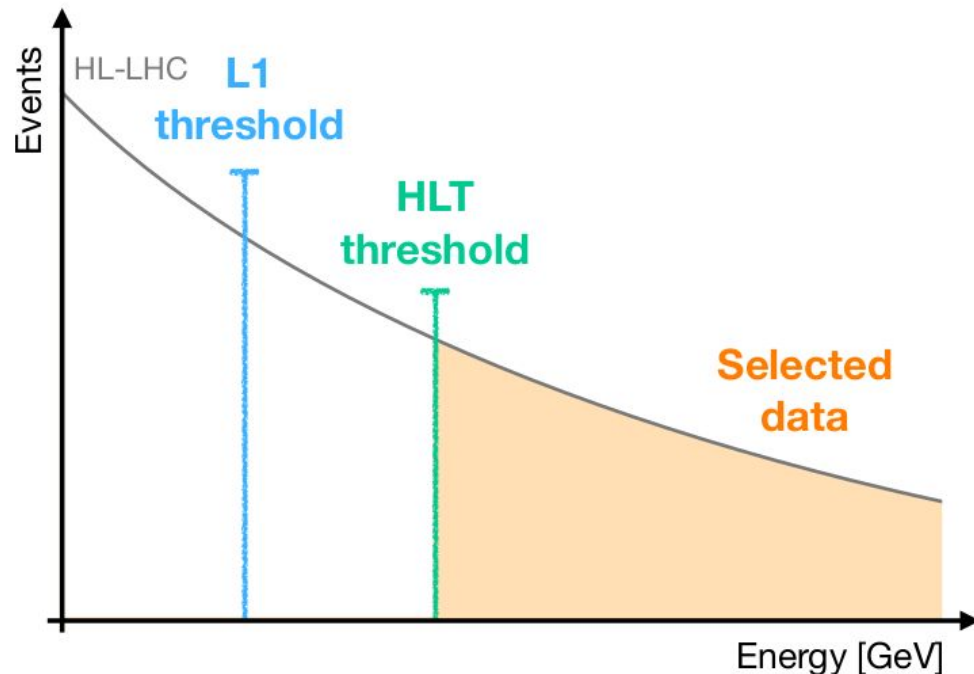
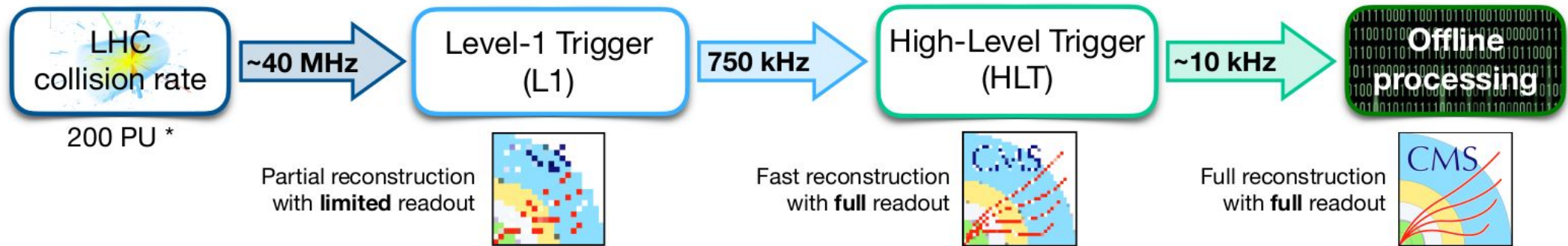
... for **smarter** decisions!



L1

HLT

Towards an upgraded trigger



Main Goal : Maintain **same physics acceptance** moving from 60 to 200 PU



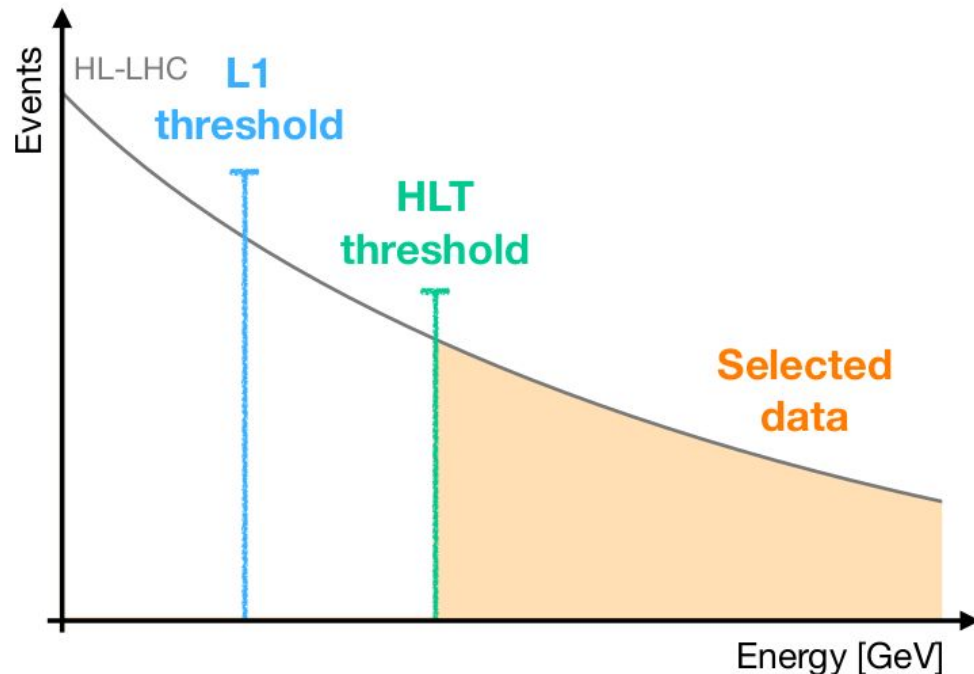
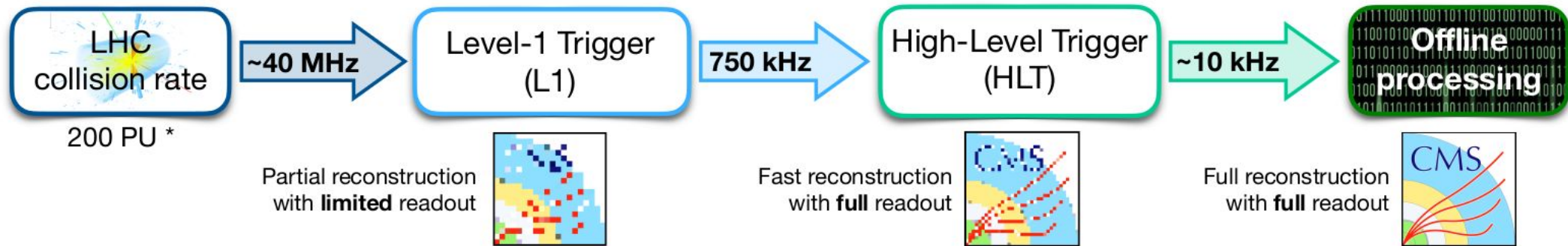
More **powerful** trigger ...
... for **smarter** decisions!



L1

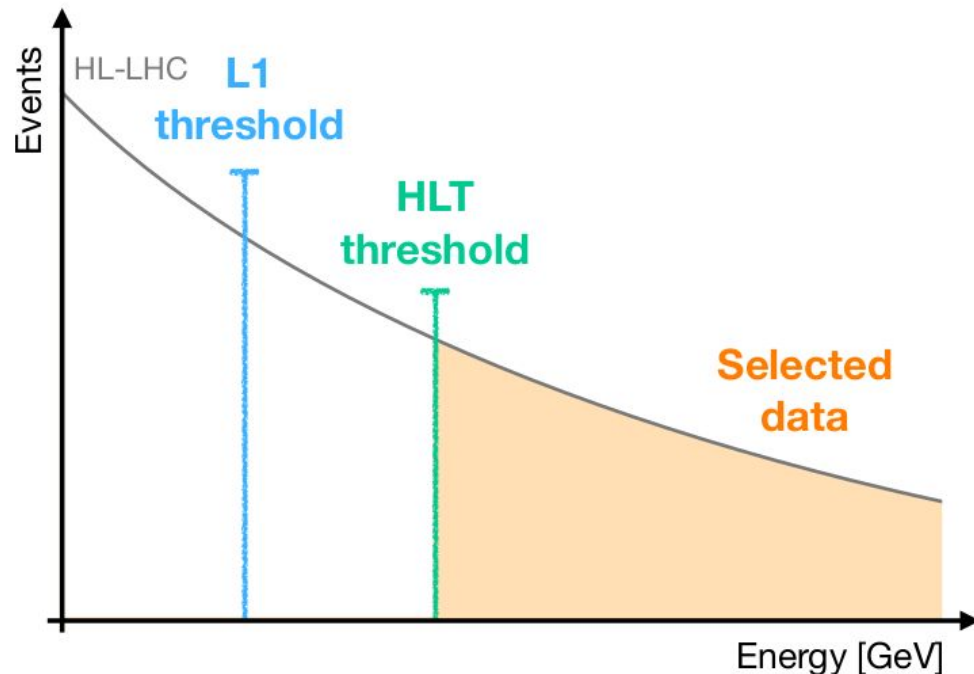
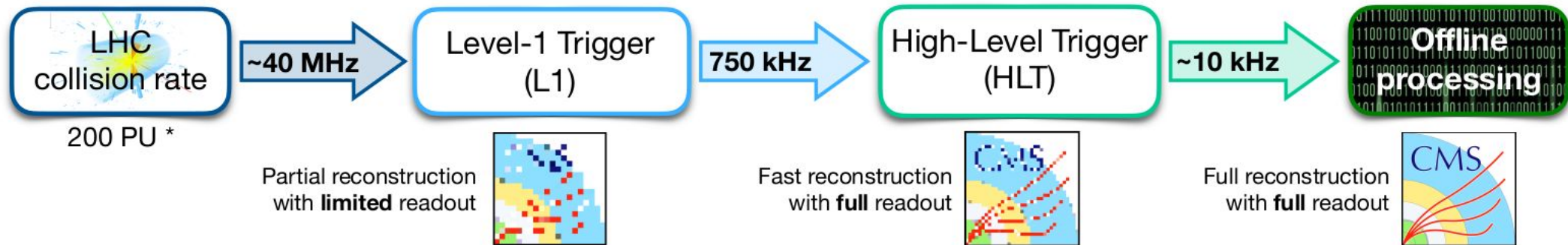
HLT

Towards an upgraded trigger



- ▶ **More powerful FPGAs**
7.5× more resources
- ▶ **Increased input stream**
from 2 Tb/s to 63 Tb/s input
 - higher detector granularity
 - tracks for the first time at L1
- ▶ **More time to make the decision**
extended latency from 3.8 to 12.5 μ s
- ▶ **Increased output rate**
from 110 kHz to 750 kHz output

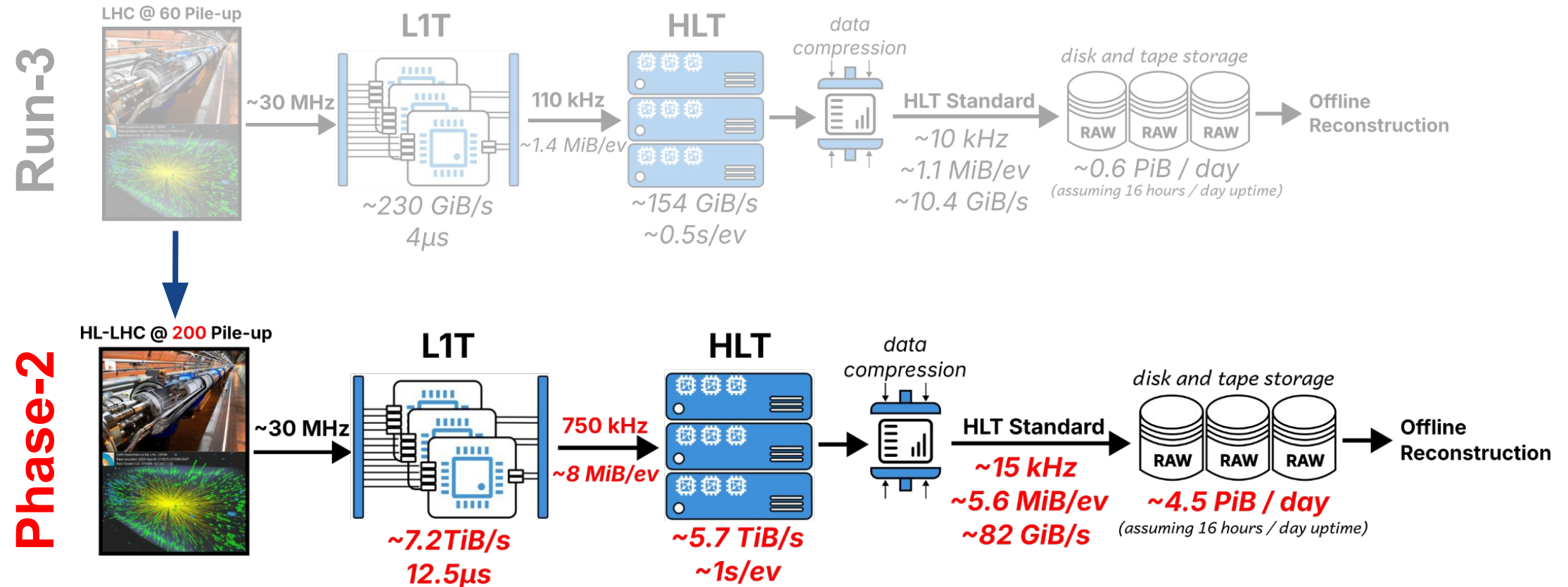
Towards an upgraded trigger



- ▶ **More processing power**
40× increase in processing demand
proportional farm size scaling not possible
- ▶ **More heterogeneous computing**
from 30 up to 50 (80) % of code execution
offloaded to GPUs in Run 4 (Run 5)
- ▶ **Fast/smart reconstruction algorithms**
 - Patatrack Pixel tracking
 - Line Segment Tracking (LST)
 - The Iterative CLustering (TICL)
 - Trimmed tracking

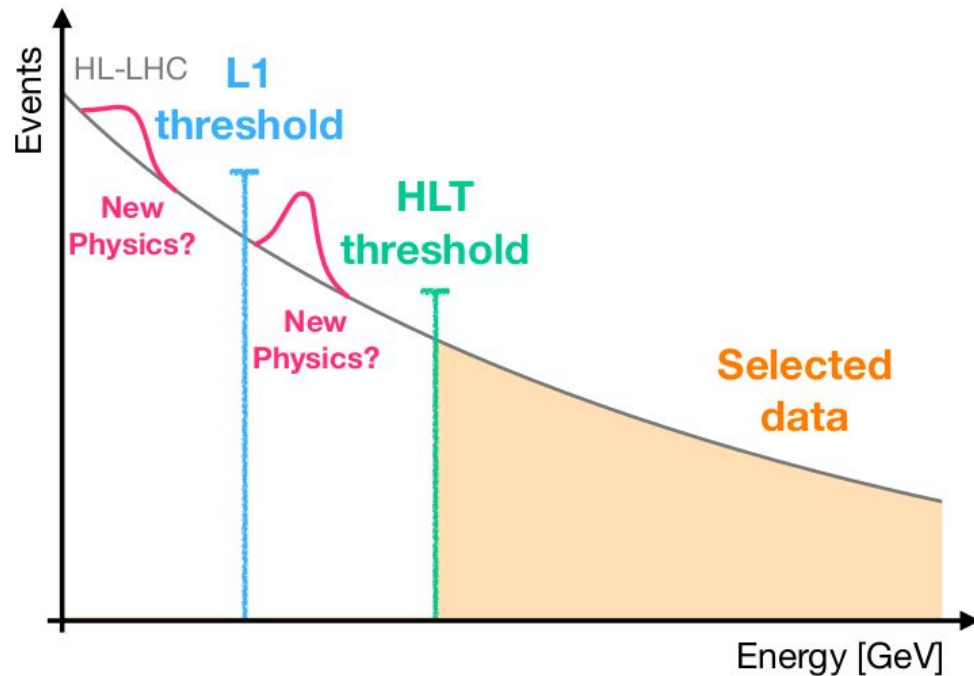
The CMS Trigger system for the HL-LHC

CMS is upgrading the L1T and HLT to enable the same physics program we are doing now:
from ~62 to 200 pile-up interactions at the HL-LHC!



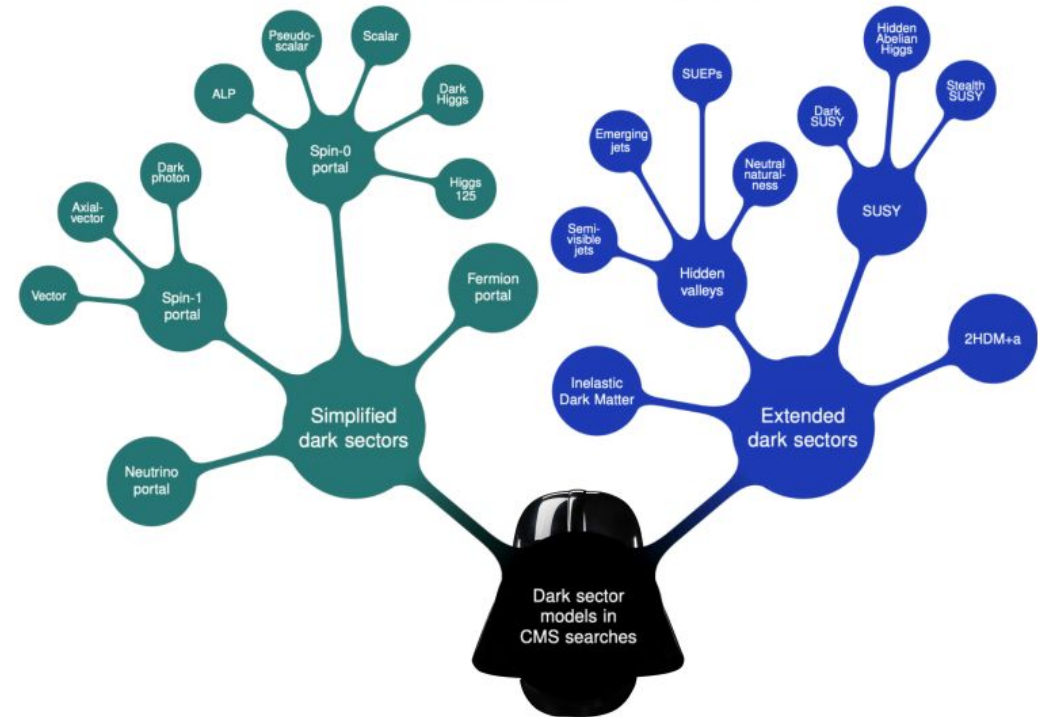
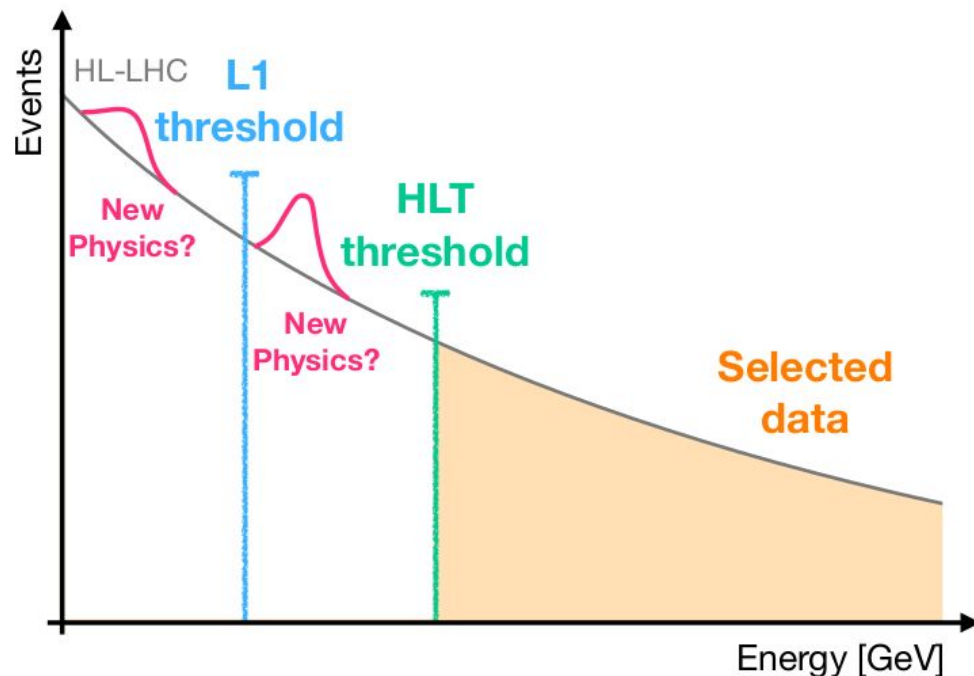
The dark side of the phase space

What if *New Physics* is hidden behind the bulk of background events that are thrown away due to the trigger selections?



The dark side of the phase space

What if *New Physics* is hidden behind the bulk of background events that are thrown away due to the trigger selections?



We could be losing 100% of signal !



NextGen
Next Generation Triggers

Idea stolen from Benedict Maier's talk
@2nd NGT Technical Workshop [\[link\]](#)

EPISODE II

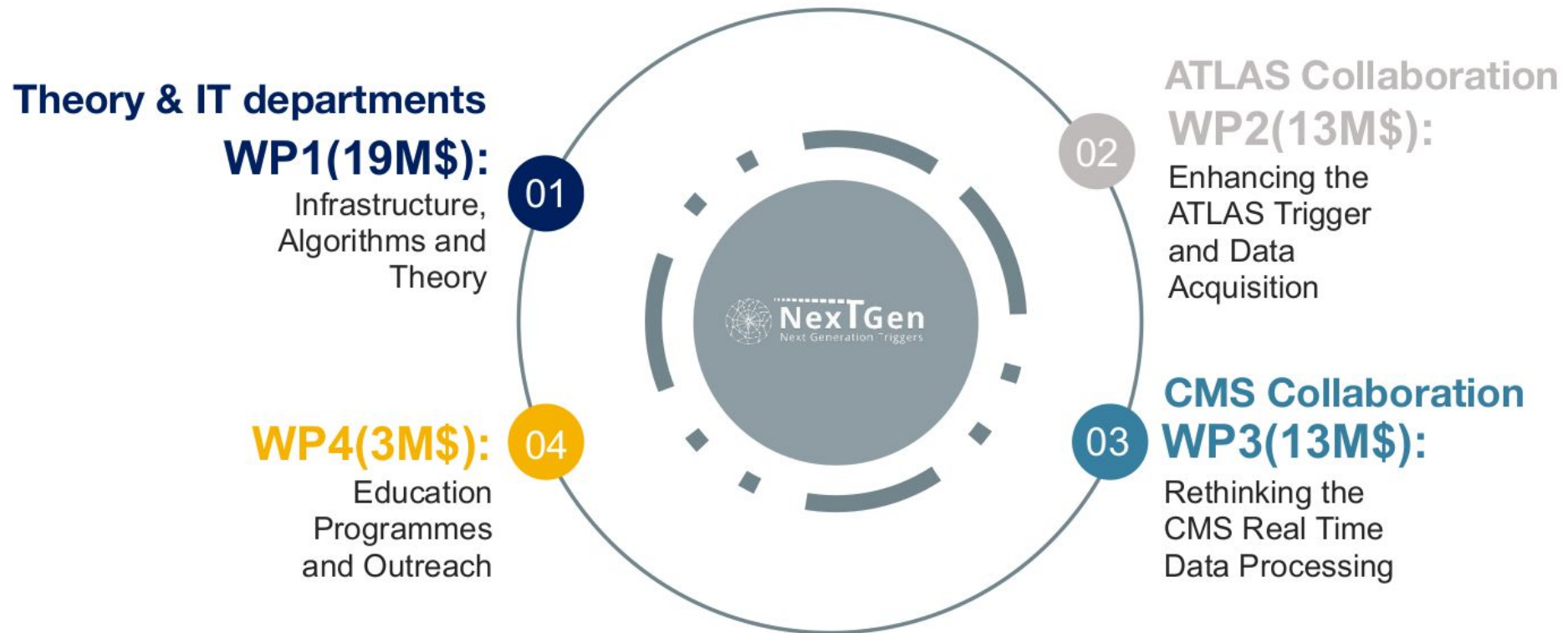
THE RISE OF NEXTGEN

A new era dawns at the High-Luminosity LHC. As collision rates soar and detectors gain unprecedented granularity, a vast storm of data threatens to overwhelm the systems that once kept watch over the frontier of discovery. The triggers of today can no longer stand alone.

In response, the Next-Generation Triggers project has emerged. Born from a unique alliance between CERN and visionary supporters, NextGen reimagines event selection from the ground up, fusing neural-network optimisation, high-performance routing, and cutting-edge FPGAs to sift through the chaos

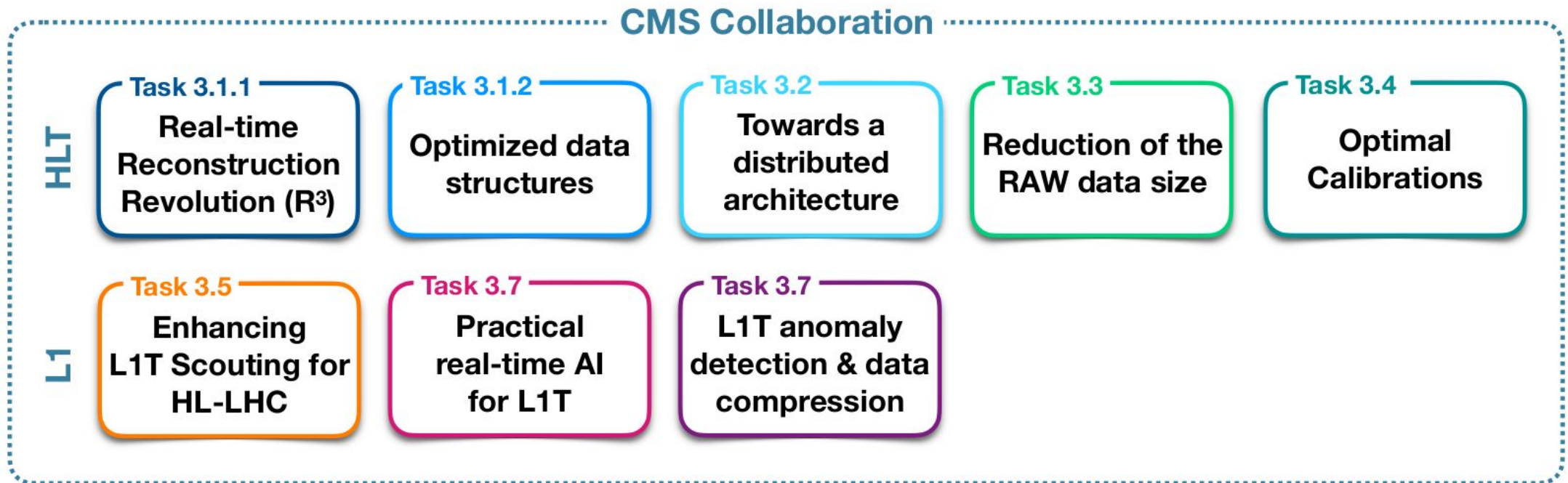
The Next Generation Trigger Project

- The **Next-Generation Trigger** (NGT) project was launched to redesign the real-time data processing
 - extend physics reach **beyond the original trigger** architecture
- Five years project (2024-2028) supported by an external donation

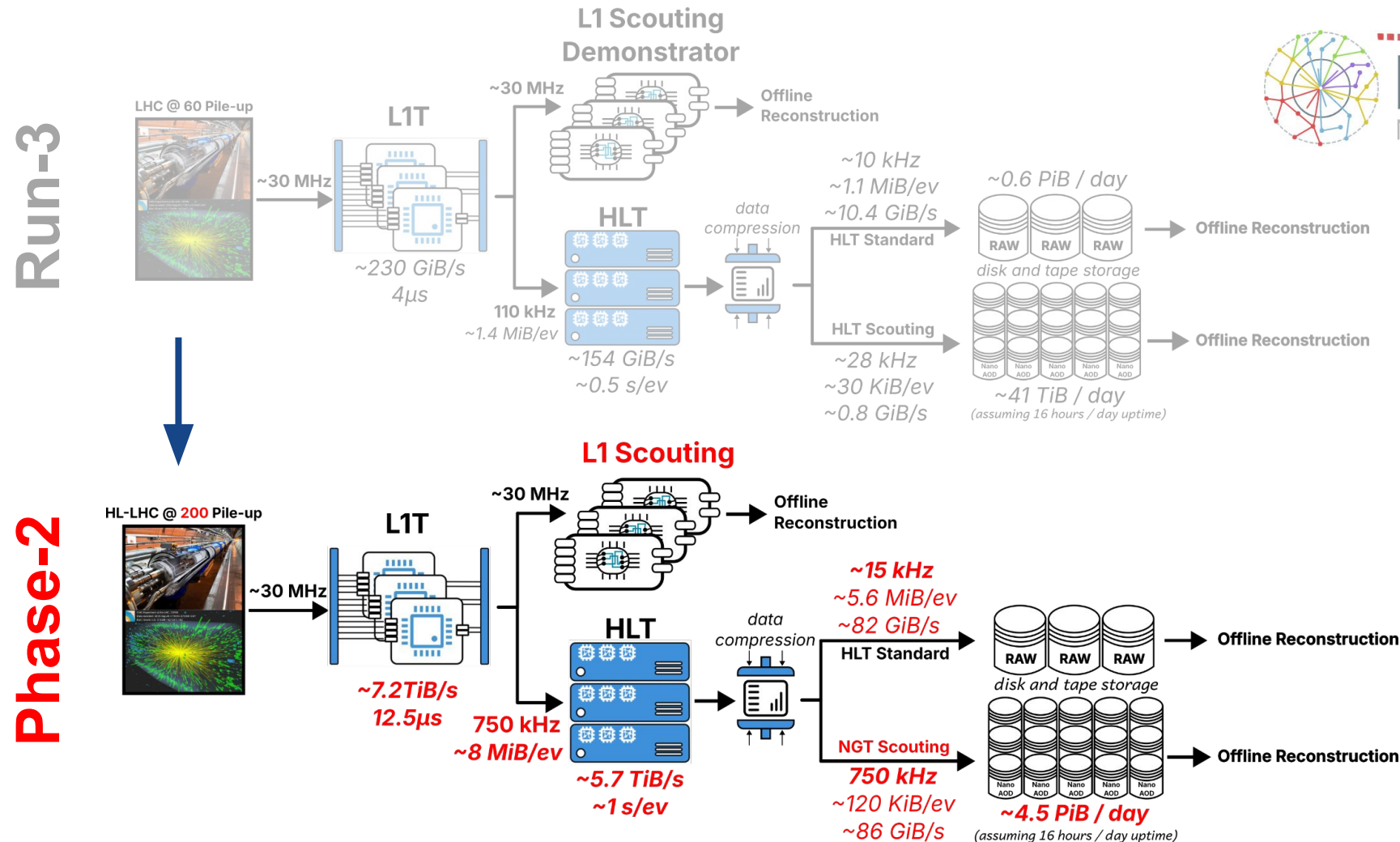


The Next Generation Trigger Project

- The **Next-Generation Trigger** (NGT) project was launched to redesign the real-time data processing
 - extend physics reach **beyond the original trigger** architecture
- Five years project (2024-2028) supported by an external donation



NGT Vision for the Phase-2 CMS Trigger System





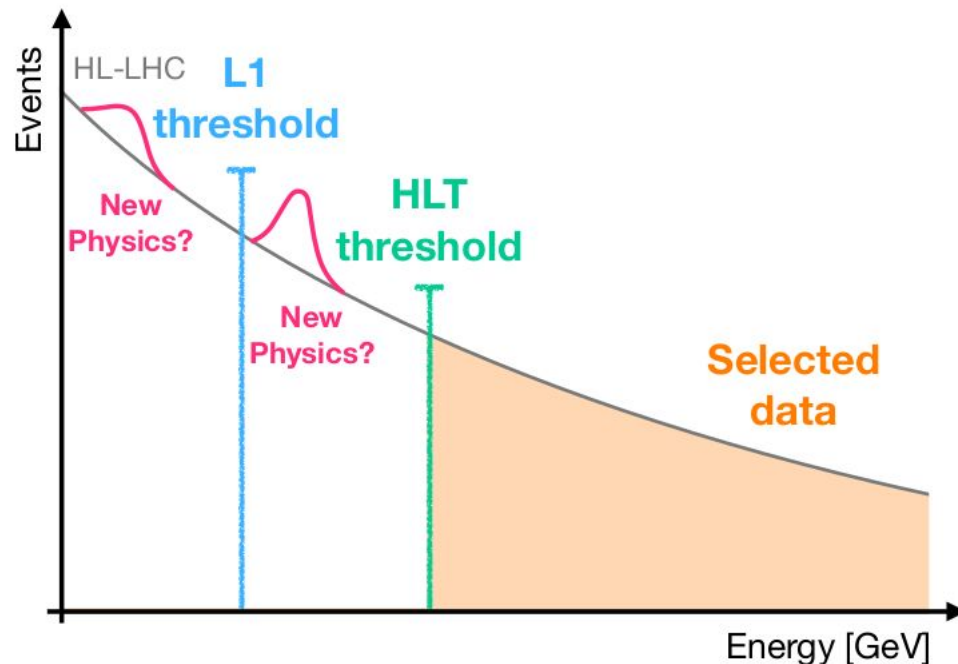
NGT @ L1

“A Machine Learning Revolution”



NextTGen
Next Generation Triggers

Expanding the L1 Physics Acceptance



Enhanced L1 filtering

- Improved signal **acceptance** across all paths
- Access to **exclusive** and **complex** final states



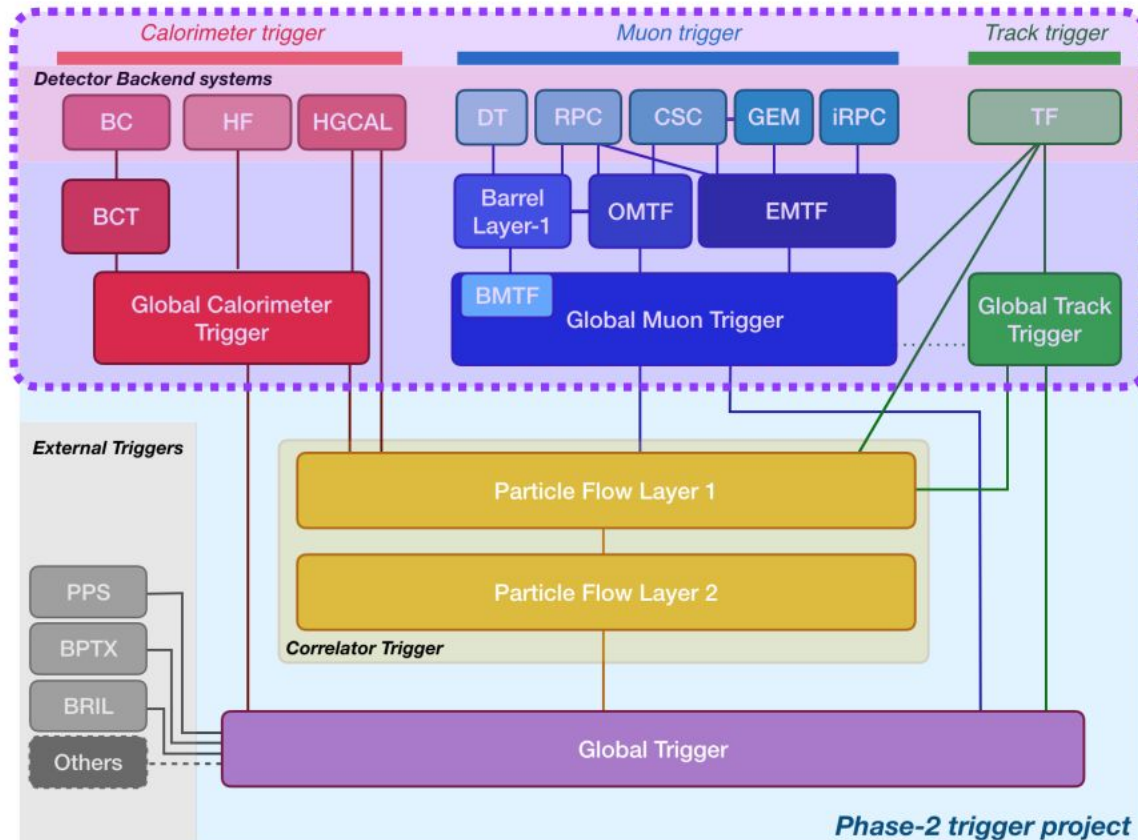
Enhanced L1 Scouting

- Record and analyse **all events** at 40 MHz



The Upgraded L1 Architecture

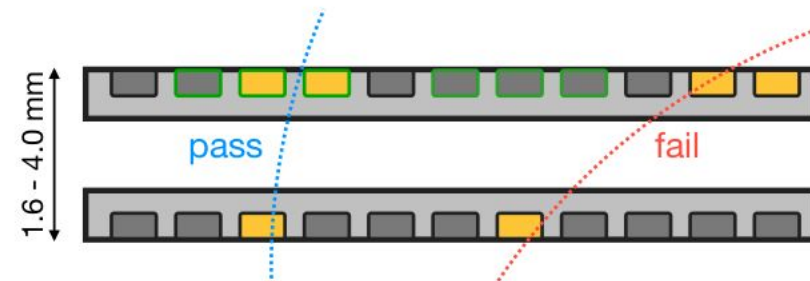
- The L1 system introduces a substantial increase in complexity to cope with HL-LHC conditions
 - Higher latency (12.5 μ s) + more powerful FPGAs (7.5 $\times$ more resources)



[CMS-TDR-021]

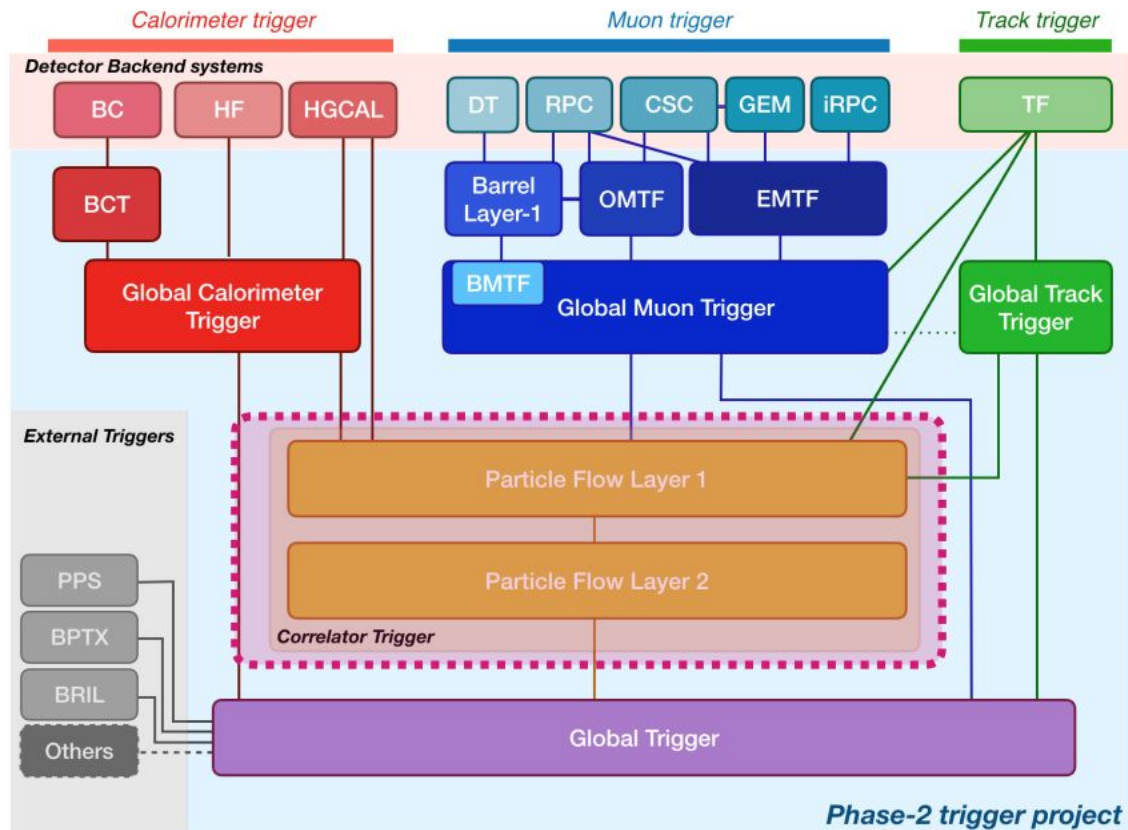
Better inputs

- ▶ **More granular** information
 - Extended **muon** coverage to $|\eta| < 2.8$
 - Shower shapes from **HGCAL**
- ▶ **Tracking** for the first time at L1
 - **Stubs** from Outer Tracker with $|\eta| < 2.4$ and $p_T > 2$ GeV



The Upgraded L1 Architecture

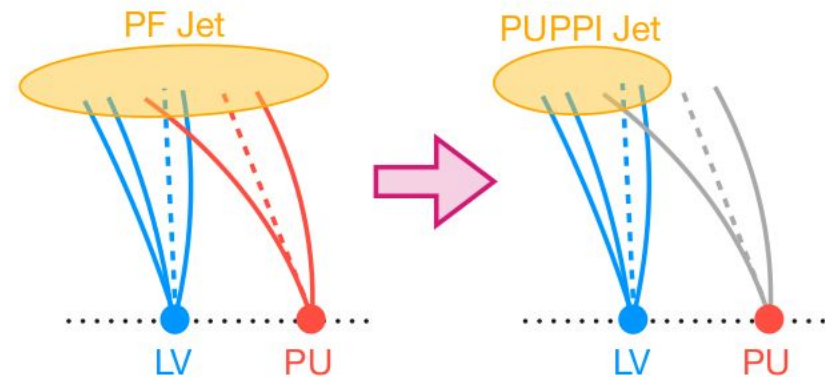
- The L1 system introduces a substantial increase in complexity to cope with HL-LHC conditions
 - Higher latency (12.5 μ s) + more powerful FPGAs (7.5 $\times$ more resources)



[CMS-TDR-021]

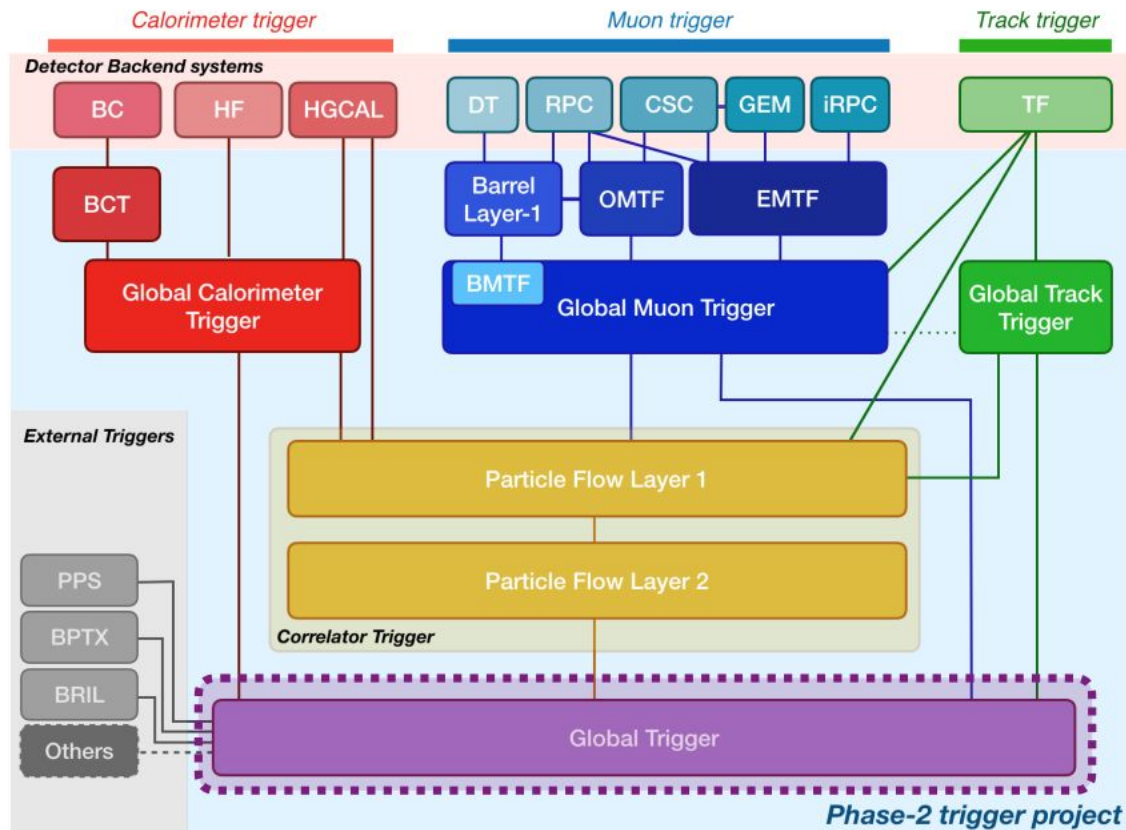
More complex architecture

- Introduction of **Correlator Layer**
 - **Particle Flow** reconstruction logic linking track and calorimeter deposit
 - Better **pile-up rejection** with PUPPI (PileUp-Per-Particle-Identification)



The Upgraded L1 Architecture

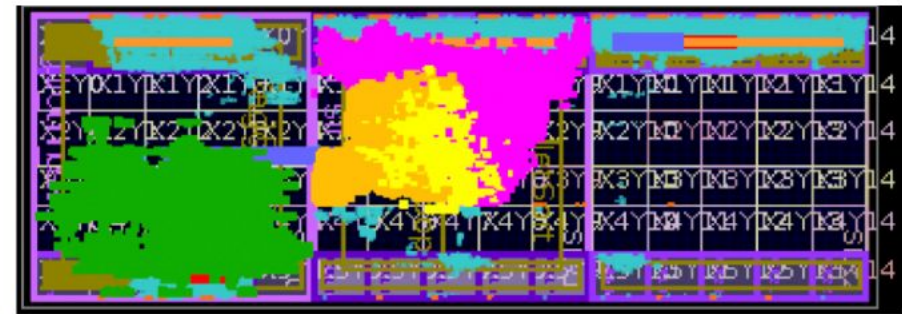
- The L1 system introduces a substantial increase in complexity to cope with HL-LHC conditions
 - Higher latency (12.5 μ s) + more powerful FPGAs (7.5 $\times$ more resources)



[CMS-TDR-021]

Machine-Learning

- Sophisticated ML-based algorithms
 - Custom firmware design for FPGAs
 - Synthesise NN/BDTs to fit within the available resources



Real-time Machine Learning at L1

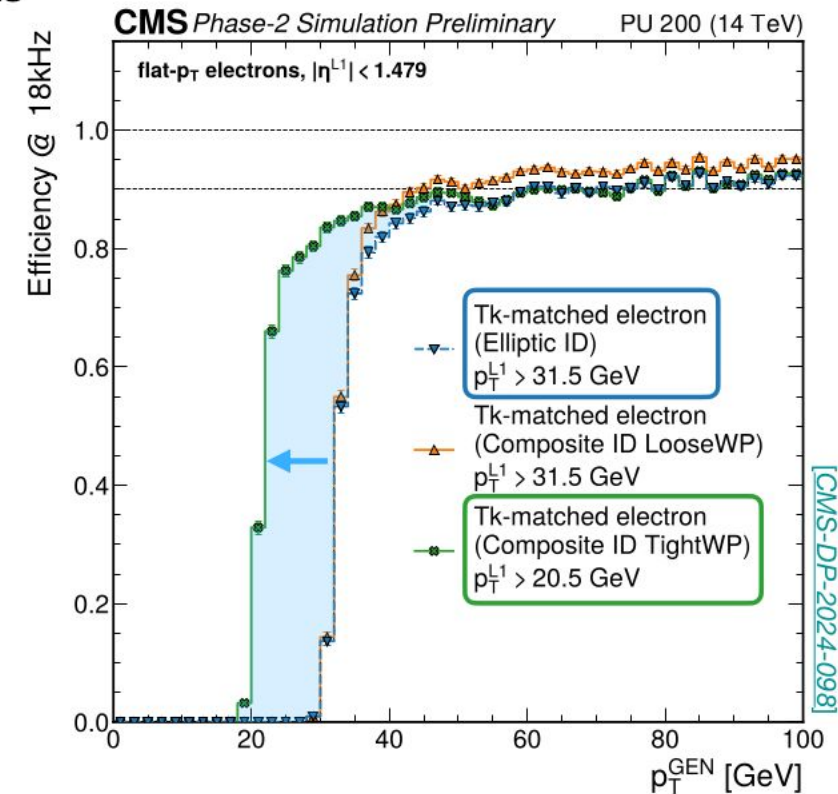
- Enable offline-level intelligence at L1: **object identification**, flavour tagging, energy regression

Object Identification

- Reconstruct **electrons** matching ECAL clusters + tracks

- Challenge: distinguish from π , γ , pile-up
- Baseline approach: **Elliptic ID**
 - Tracks with $p_T > 10$ GeV
 - **Tight** η - ϕ elliptical matching to clusters
- New approach: **Composite ID**
 - Tracks with $p_T > 10$ GeV
 - **Looser** η - ϕ elliptical matching to clusters
 - **ML model** to classify track-cluster pairs

L1 SingleElectron p_T threshold lowered by **more than 10 GeV** for a fixed 18 kHz rate

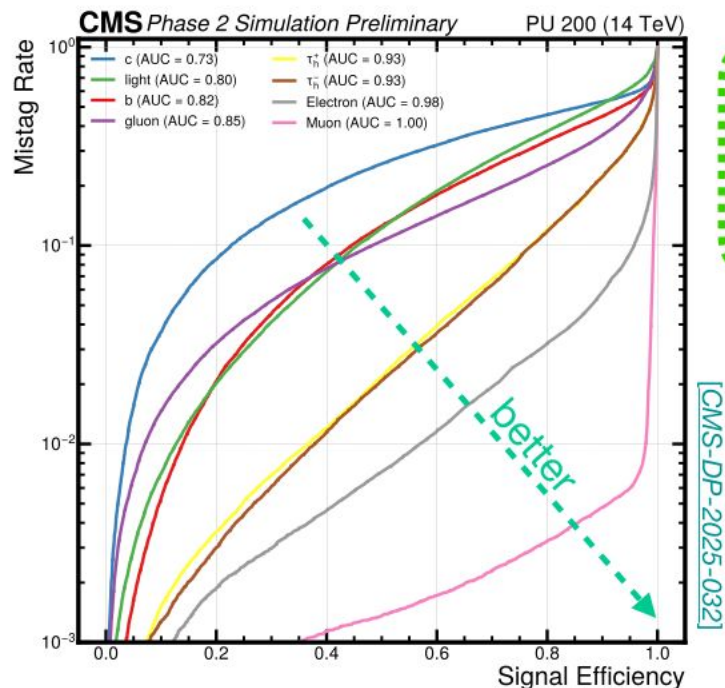


Real-time Machine Learning at L1

- Enable offline-level intelligence at L1: object identification, **flavour tagging**, energy regression

Flavour Tagging

- ▶ Deploy first ultrafast ML-based multi-class **jet tagging** algorithm at L1
 - Jets from Seeded Cone algorithm with radius 0.4 (SC4), based on PUPPI candidates

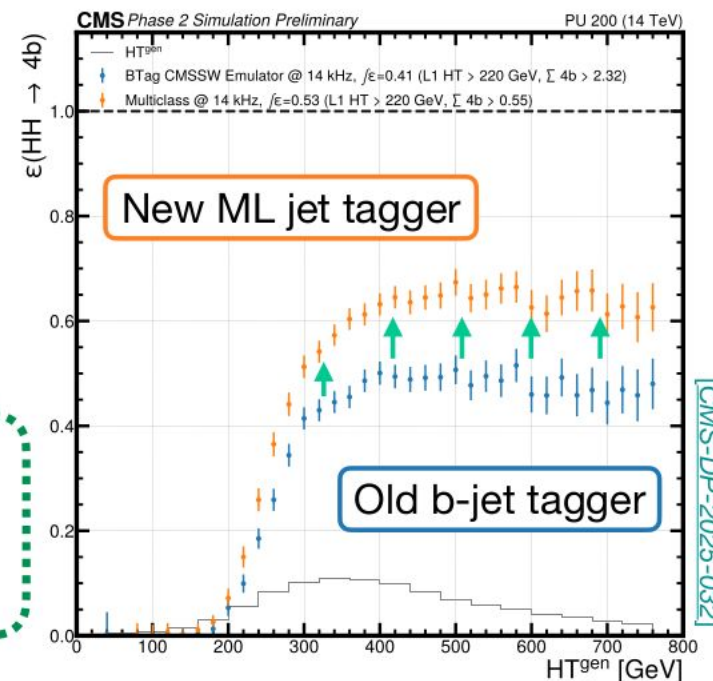


Unified discriminator

- b/c/light jets
- hadronic τ

Evaluation on $HH \rightarrow 4b$

- **15% improvement** in signal efficiency

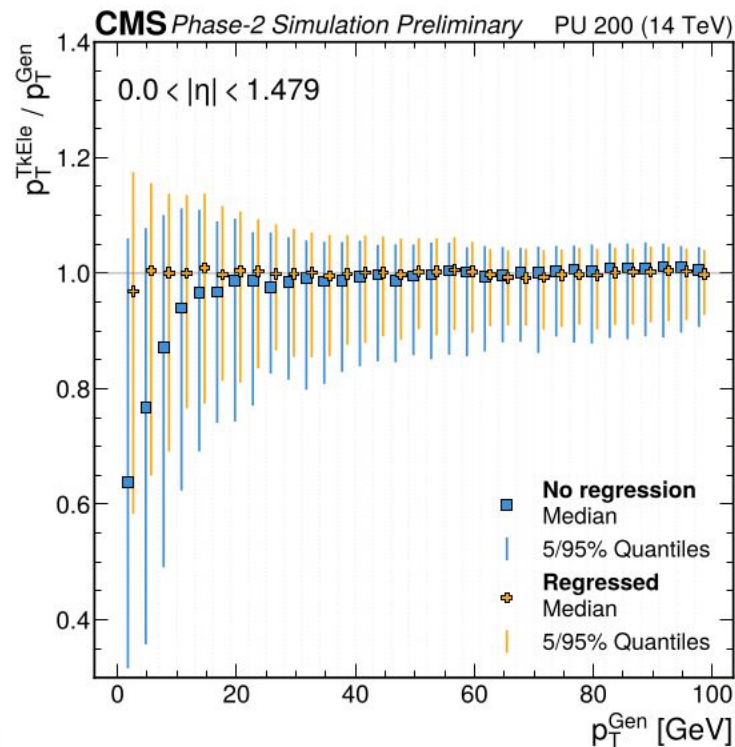


Real-time Machine Learning at L1

- Enable offline-level intelligence at L1: object identification, flavour tagging, **energy regression**

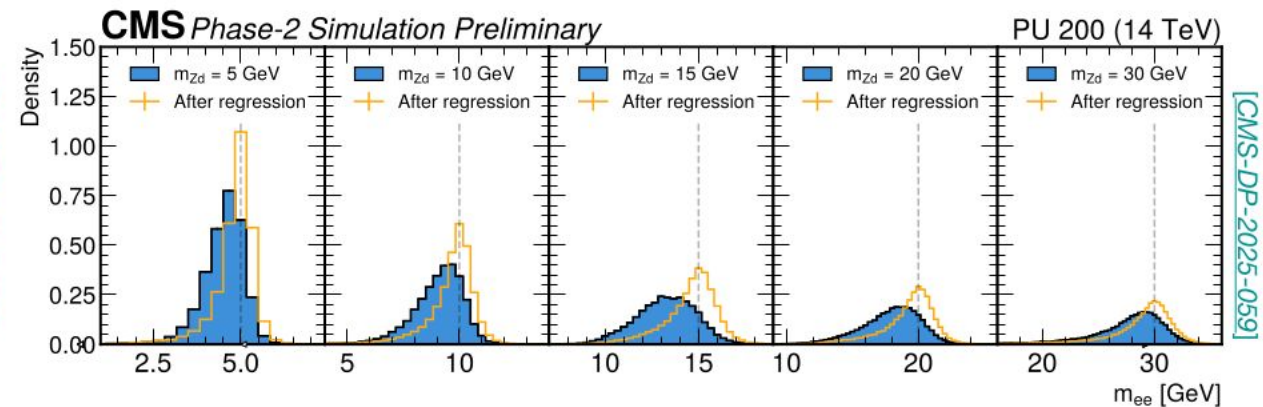
Energy Regression

- ▶ Exploit low-level cluster features in ML models to improve low- p_T **electron resolution**

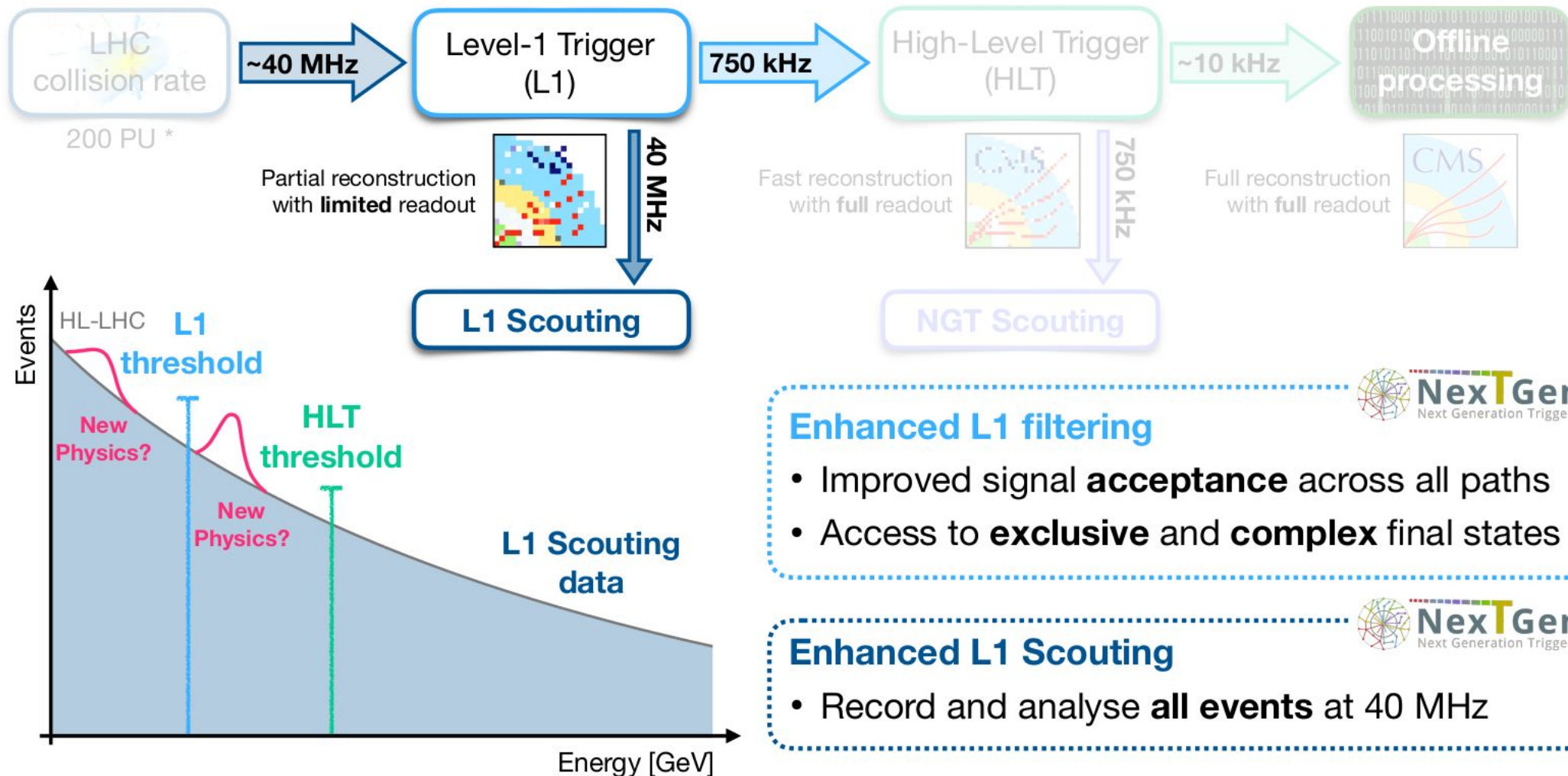


- BDT trained targeting response $p_T^{\text{Gen}} / p_T^{\text{Calo}}$ to 1
 - 7 features: ID score, track quality, shower shape, ...
 - Evaluation on Dark Photon sample $Z' \rightarrow ee$

Regression highly improves **resolution** and **scale**

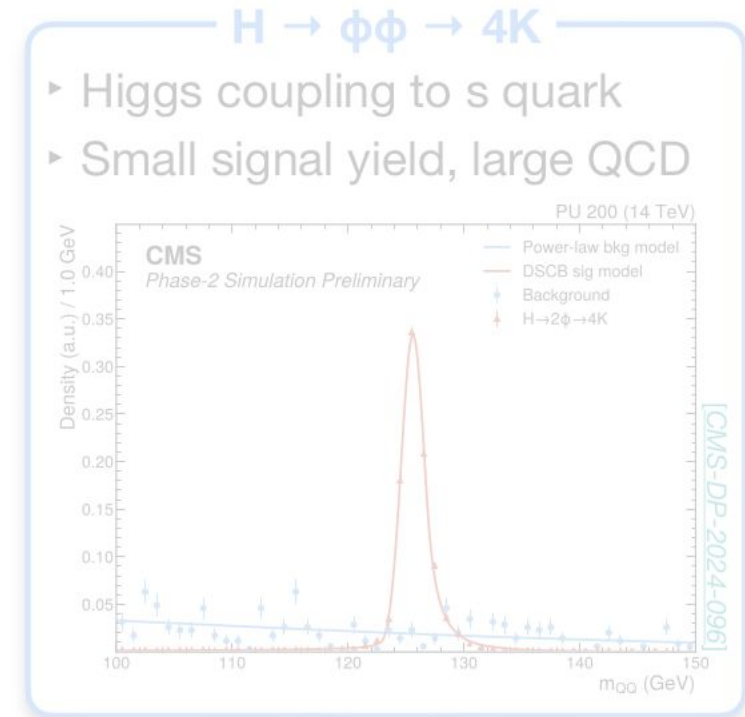
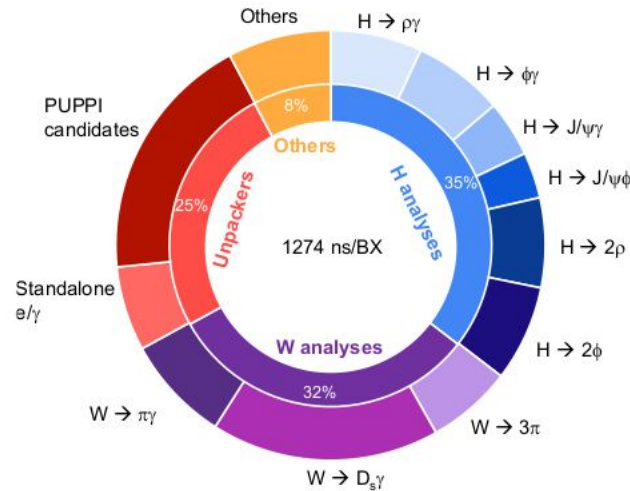
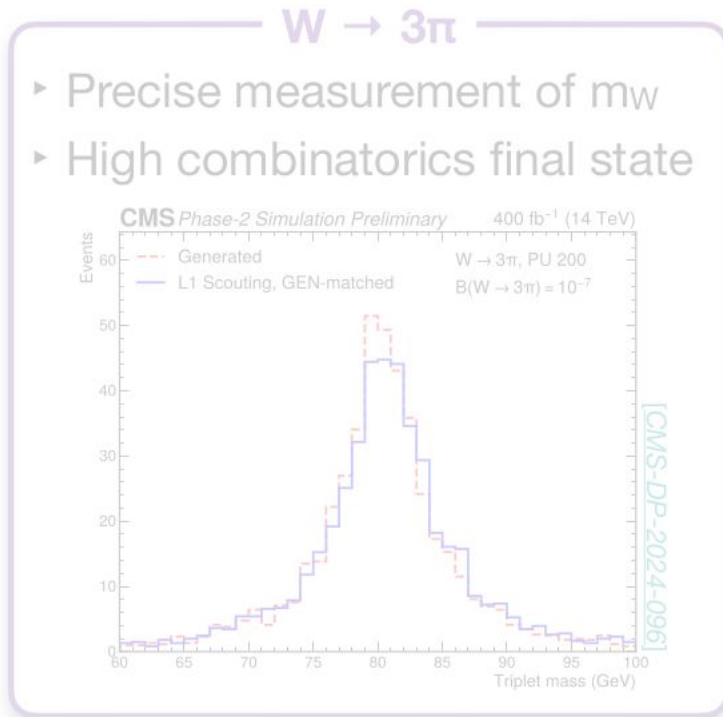


Triggerless Analysis with L1 Scouting



L1 Scouting: What Physics is possible?

- The L1 system evolves from a pure trigger into a **tool for physics analysis**
 - Access **physics signatures** that are not captured by the standard DAQ chain
 - Study time-correlation across **several bunch crossing (BX)** periods for long-lived particles



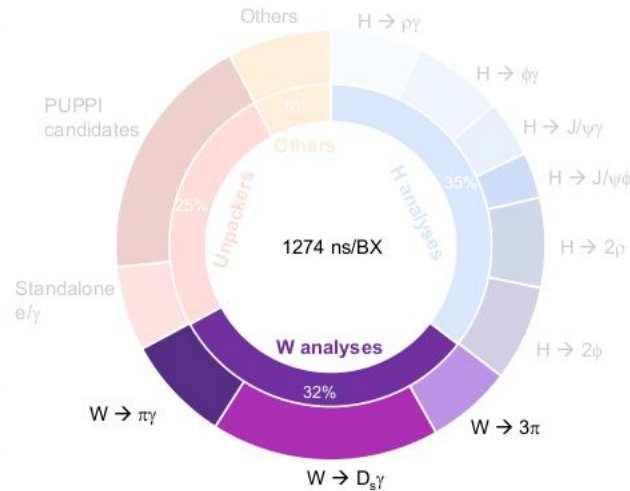
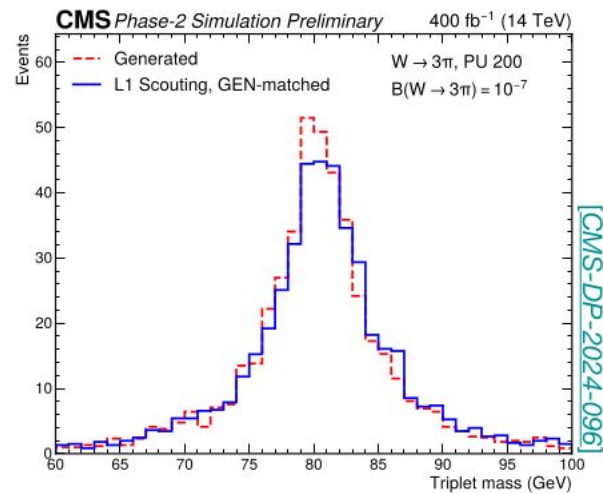
✓ First prototype successfully validated with a demonstrator !

L1 Scouting: What Physics is possible?

- The L1 system evolves from a pure trigger into a **tool for physics analysis**
 - Access **physics signatures** that are not captured by the standard DAQ chain
 - Study time-correlation across **several bunch crossing (BX)** periods for long-lived particles

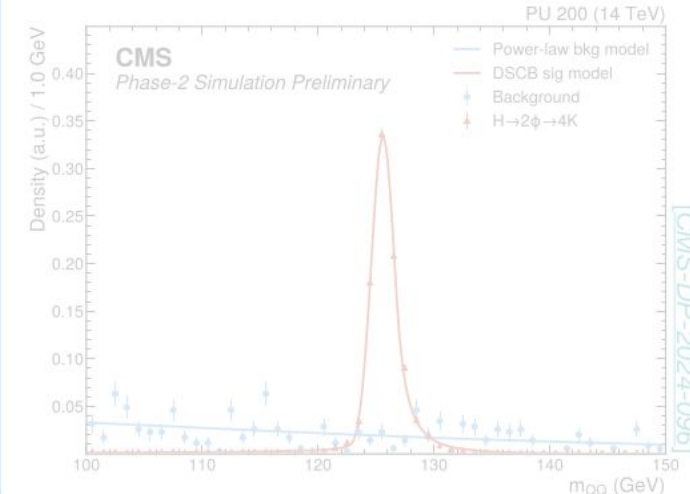
$W \rightarrow 3\pi$

- Precise measurement of m_W
- High combinatorics final state



$H \rightarrow \phi\phi \rightarrow 4K$

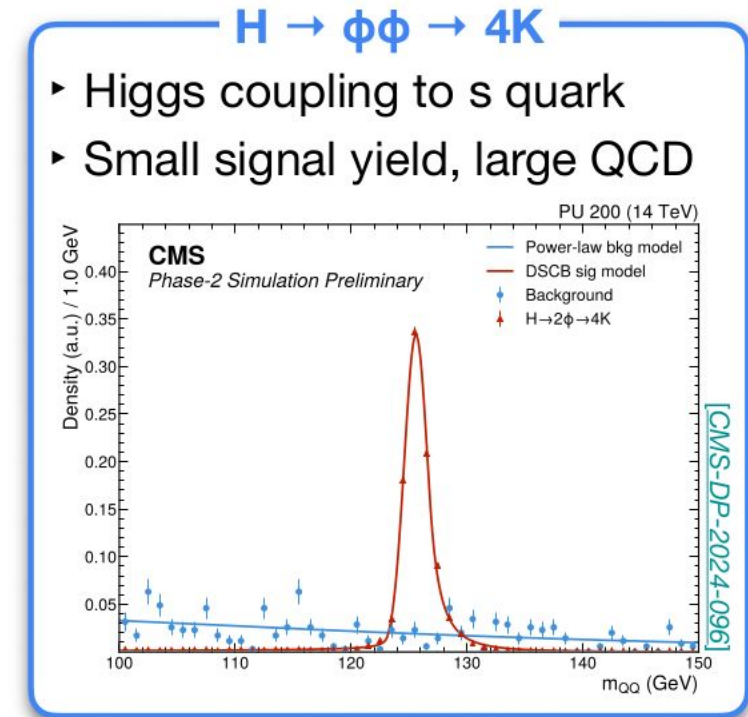
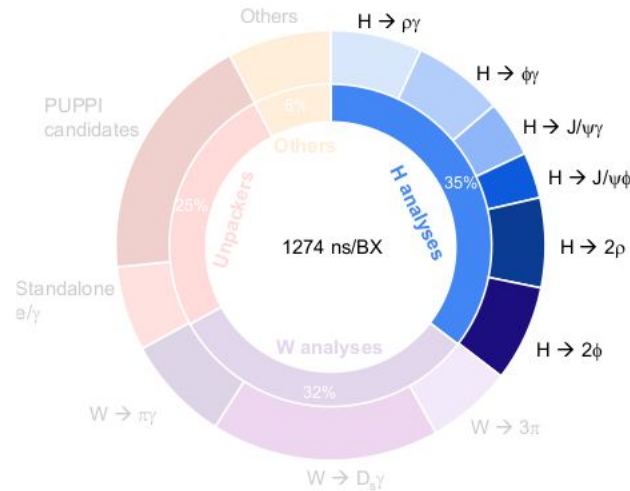
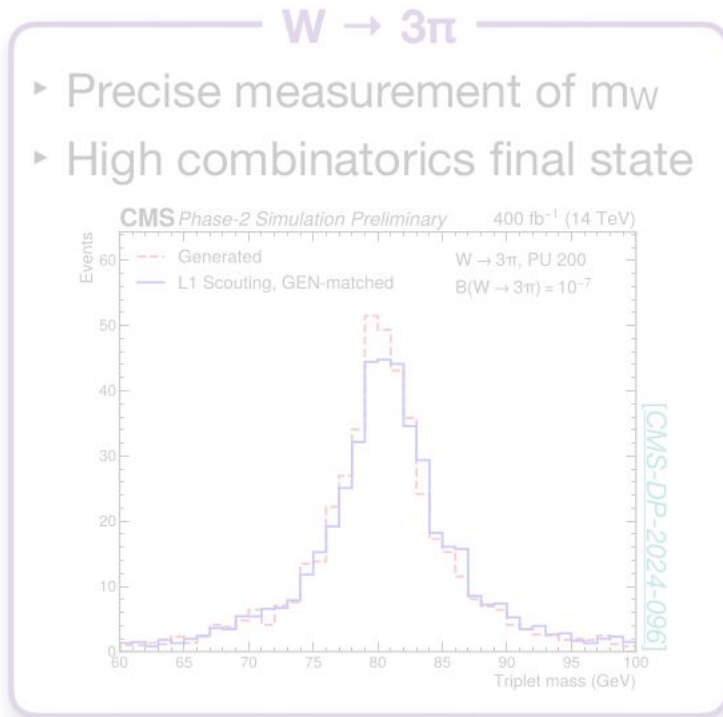
- Higgs coupling to s quark
- Small signal yield, large QCD



✓ First prototype successfully validated with a demonstrator !

L1 Scouting: What Physics is possible?

- The L1 system evolves from a pure trigger into a **tool for physics analysis**
 - Access **physics signatures** that are not captured by the standard DAQ chain
 - Study time-correlation across **several bunch crossing (BX)** periods for long-lived particles



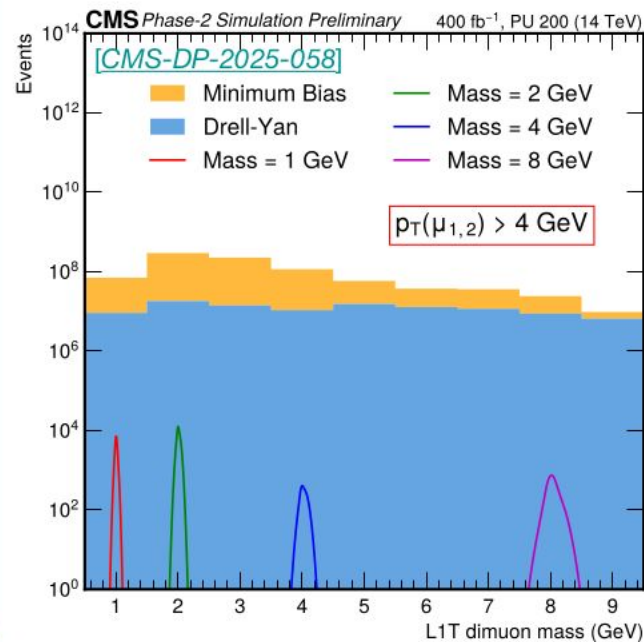
✓ First prototype successfully validated with a demonstrator !

L1 Scouting: Extending the Program

- Extend search reach to **very low-mass** resonances and increase variety of final states

$X \rightarrow \mu\mu$

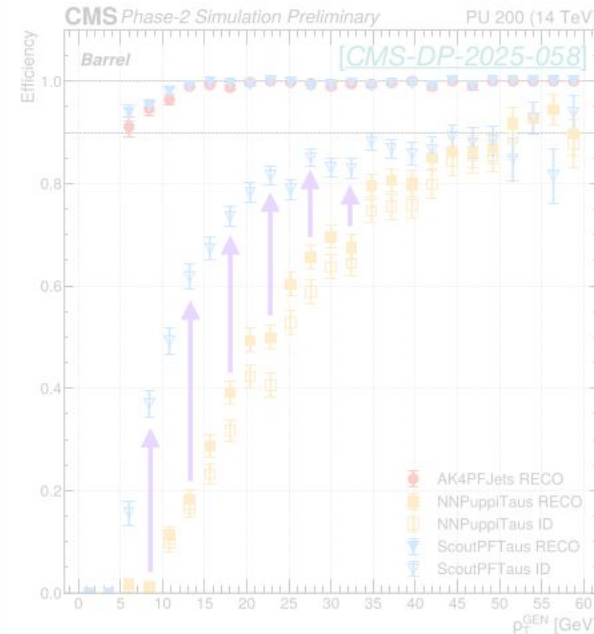
- Very good di-muon mass resolution
 - High **signal** efficiency $> 98\%$
 - Good **background** rejection 5×10^{-6}



- $m_X = 1 \text{ GeV}$
FWHM = **60 MeV**
- $m_X = 2 \text{ GeV}$
FWHM = **82 MeV**
- $m_X = 4 \text{ GeV}$
FWHM = **143 MeV**
- $m_X = 8 \text{ GeV}$
FWHM = **292 MeV**

$\phi \rightarrow \tau\tau$

- Low-mass $\tau\tau$ searches with *ScoutPFTaus*
 - Clustering** with GPU-friendly CLUE algo
 - Transformer-based NN for **identification**



Scouting
PUPPI candidates



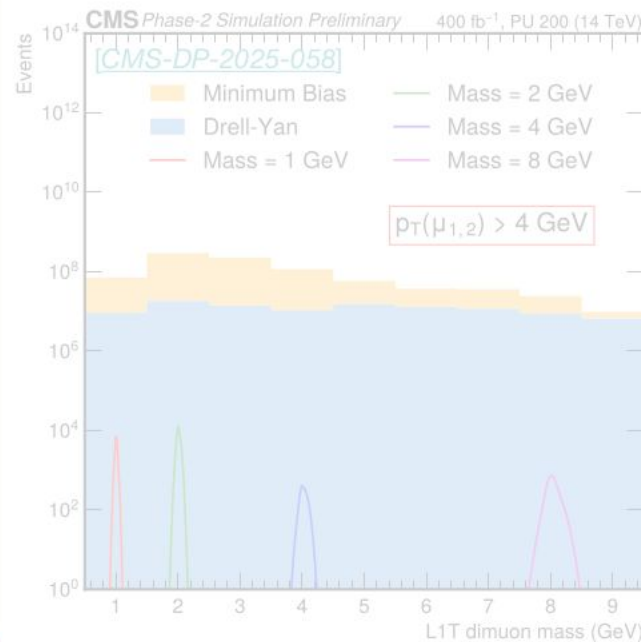
Scouting
PF candidates
(10x multiplicity)

L1 Scouting: Extending the Program

- Extend search reach to **very low-mass** resonances and increase variety of final states

$X \rightarrow \mu\mu$

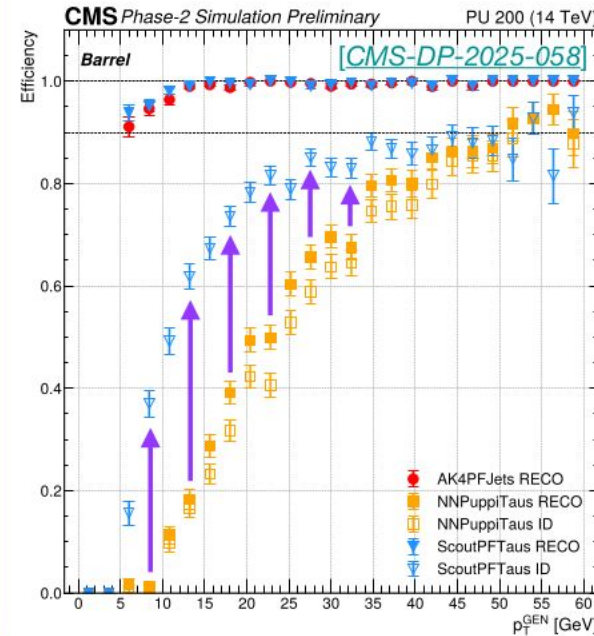
- Very good di-muon mass resolution
 - High **signal** efficiency > 98%
 - Good **background** rejection 5×10^{-6}



- $m_X = 1 \text{ GeV}$
FWHM = 60 MeV
- $m_X = 2 \text{ GeV}$
FWHM = 82 MeV
- $m_X = 4 \text{ GeV}$
FWHM = 143 MeV
- $m_X = 8 \text{ GeV}$
FWHM = 292 MeV

$\phi \rightarrow \tau\tau$

- Low-mass τ searches with *ScoutPFTaus*
 - Clustering** with GPU-friendly CLUE algo
 - Transformer-based NN for **identification**



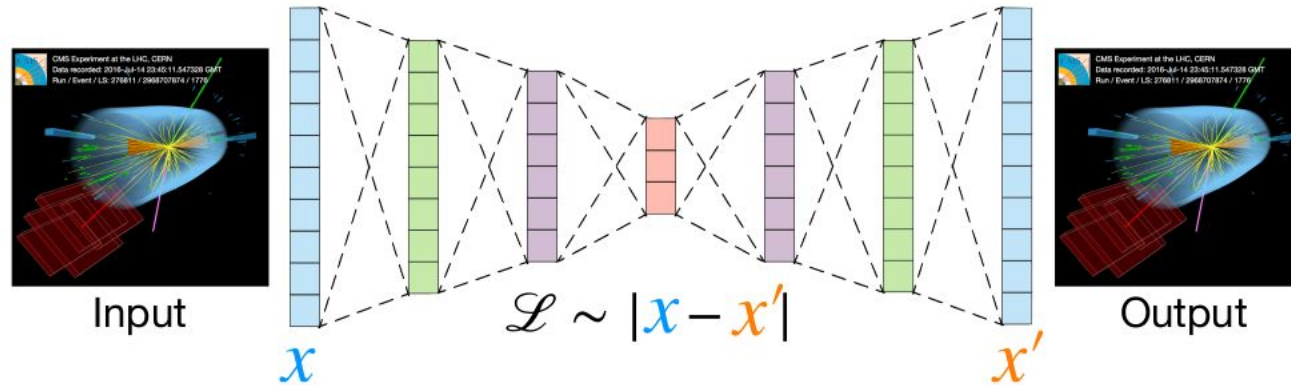
Scouting
PUPPI candidates



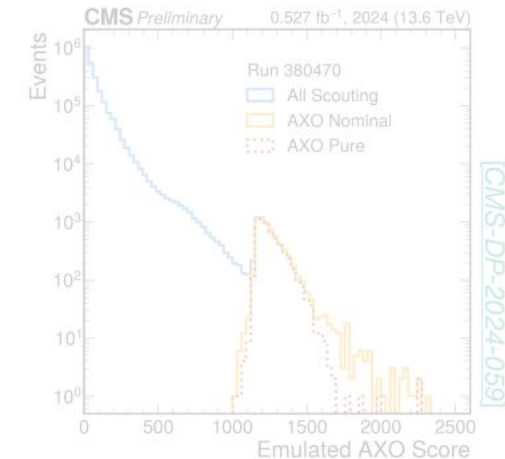
Scouting
PF candidates
(10x multiplicity)

Anomaly Detection at L1

- New physics may fall outside standard triggers, **unsupervised ML** model to identify anomaly events

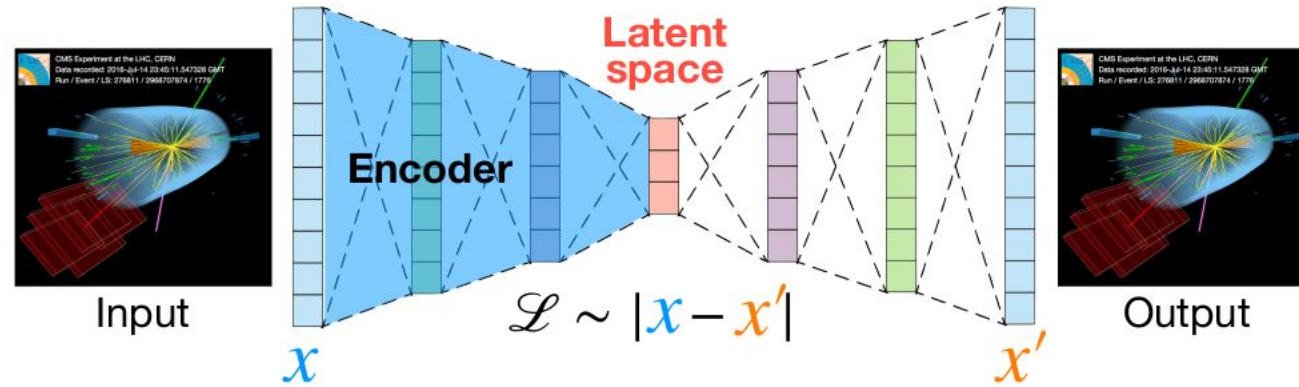


- AXOL1TL is a signal-agnostic Variational AutoEncoder (VAE)
 - ▶ **Typical event** → minimal error (low score)
 - ▶ **Atypical event** → large error (high score)
- Deployment and characterisation on data during Run-3
- Transition to **contrastive learning**, model only based on encoder
 - ▶ Simpler inference, lower latency, better discrimination

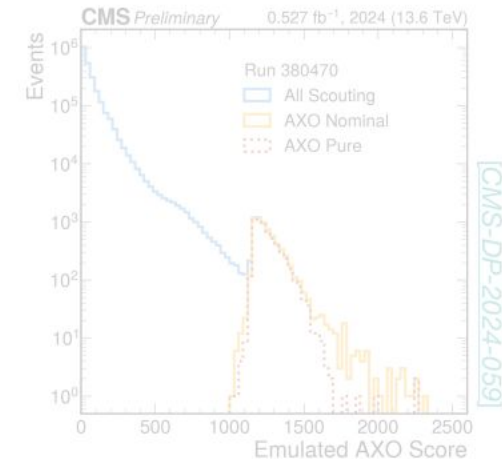


Anomaly Detection at L1

- New physics may fall outside standard triggers, **unsupervised ML** model to identify anomaly events

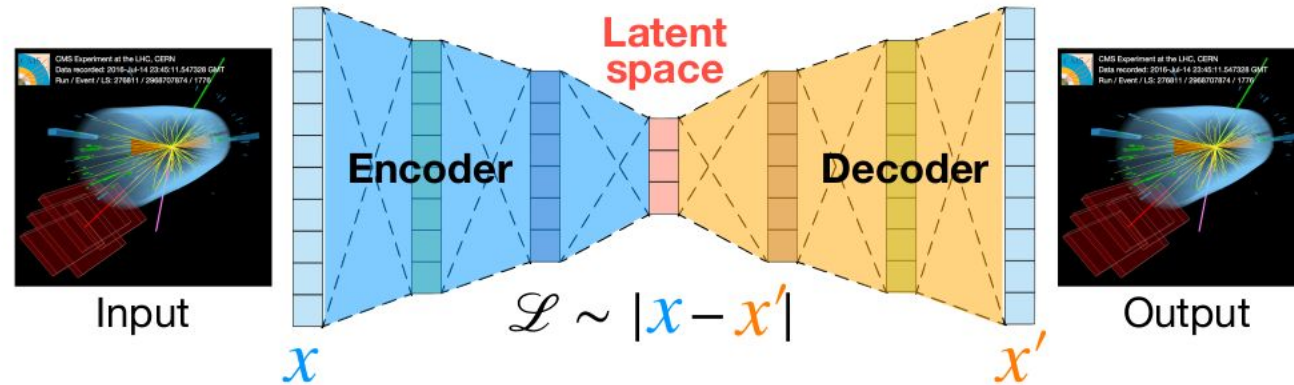


- AXOL1TL is a signal-agnostic Variational AutoEncoder (VAE)
 - ▶ **Typical event** → minimal error (low score)
 - ▶ **Atypical event** → large error (high score)
- Deployment and characterisation on data during Run-3
- Transition to **contrastive learning**, model only based on encoder
 - ▶ Simpler inference, lower latency, better discrimination

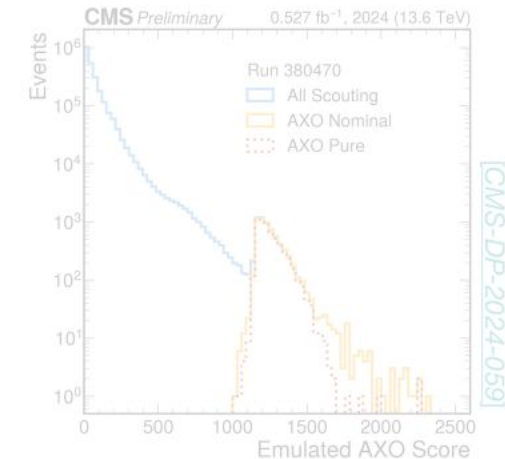


Anomaly Detection at L1

- New physics may fall outside standard triggers, **unsupervised ML** model to identify anomaly events

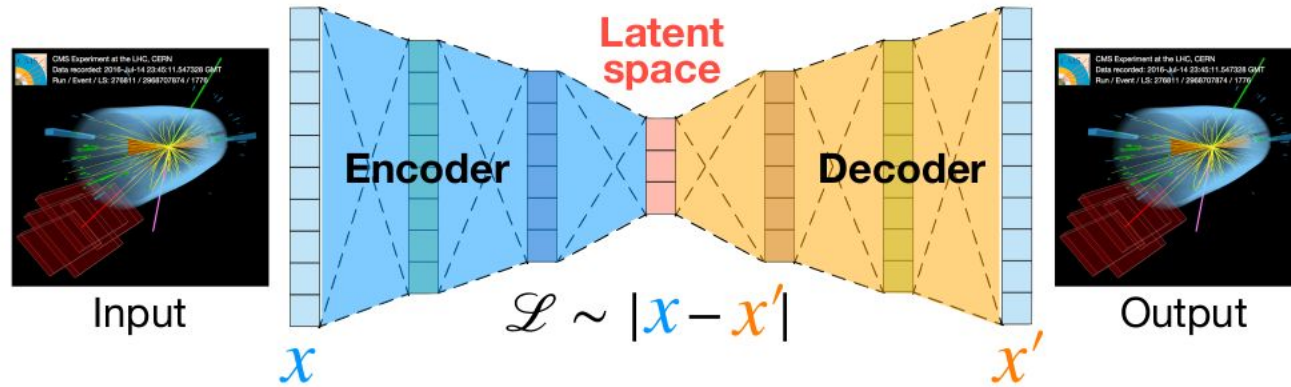


- AXOL1TL is a signal-agnostic Variational AutoEncoder (VAE)
 - **Typical event** → minimal error (low score)
 - **Atypical event** → large error (high score)
- Deployment and characterisation on data during Run-3
- Transition to **contrastive learning**, model only based on encoder
 - Simpler inference, lower latency, better discrimination

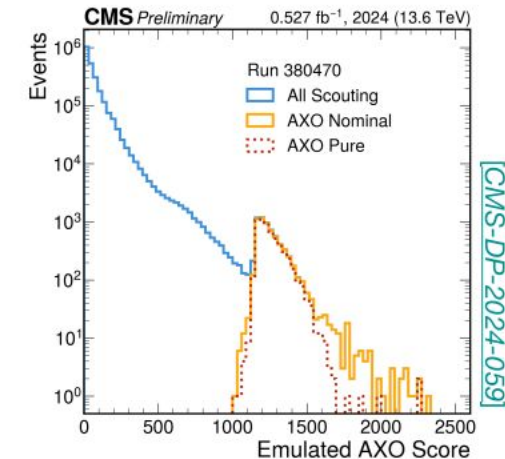


Anomaly Detection at L1

- New physics may fall outside standard triggers, **unsupervised ML** model to identify anomaly events



- AXOL1TL is a signal-agnostic Variational AutoEncoder (VAE)
 - **Typical event** → minimal error (low score)
 - **Atypical event** → large error (high score)
- Deployment and characterisation on data during Run-3
- Transition to **contrastive learning**, model only based on encoder
 - Simpler inference, lower latency, better discrimination



Summary of NGT @ L1

- The Next-Generation Trigger project at L1 aims to leverage the upgraded system architecture to include **real-time Machine Learning** inference directly in FPGAs.
- Improve purity and acceptance through a **smarter trigger decision**:
 - **Object identification** → increased efficiency for complex objects;
 - **Energy regression** → enhanced sensitivity to resonant searches;
 - **Anomaly detection** → new handles on non-SM signatures;
- Complement the standard physics program with a **40 MHz L1 Scouting** approach
 - Access rare SM decays or exotic signatures beyond the reach of the traditional triggers
 - Enable time-correlation studies across several bunch crossings

The L1 system is evolving from a trigger system into a real-time exploration platform expanding the experiment sensitivity to new phenomena



NGT @ HLT

“Scouting at Full Rate”



NexTGen
Next Generation Triggers

CMS Data Scouting: a bit of history

- **Main Idea** : “Do more with less” to overcome HLT bandwidth limitations

- ✓ **Increase trigger rate**

Looser (more inclusive) selections

- ✓ **Decrease event size**

Keep only high-level objects from HLT

- **Run-1:**

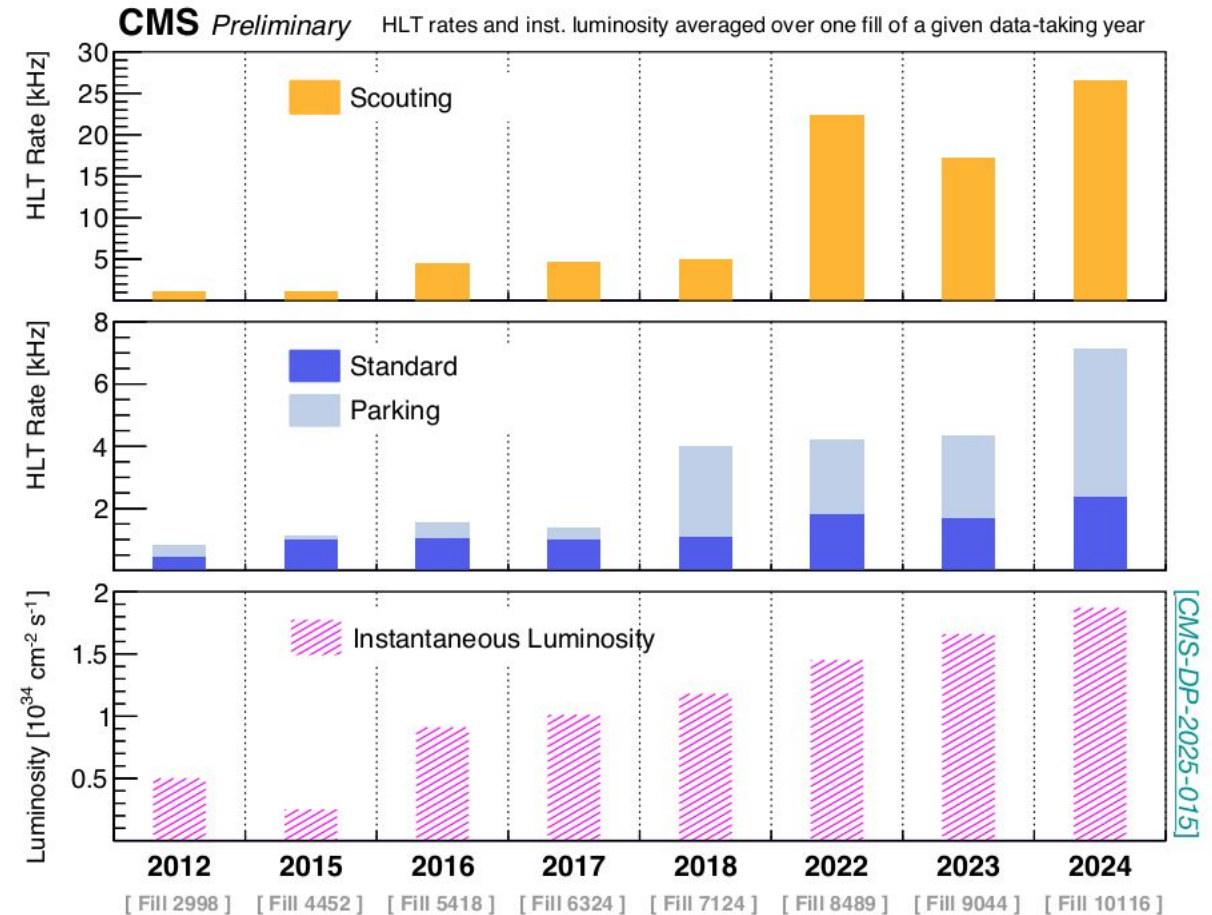
- ▶ recover low-mass multi-jet sensitivity

- **Run-2:**

- ▶ including jets, muons, electrons
- ▶ di-jet and di-muon resonances
- ▶ SM rare decays

- **Run-3:**

- ▶ lower thresholds and new objects
- ▶ 10× standard HLT rate



CMS Data Scouting: a bit of history

- **Main Idea :** “Do more with less” to overcome HLT bandwidth limitations

- ✓ **Increase trigger rate**

Looser (more inclusive) selections

- ✓ **Decrease event size**

Keep only high-level objects from HLT

- **Run-1:**

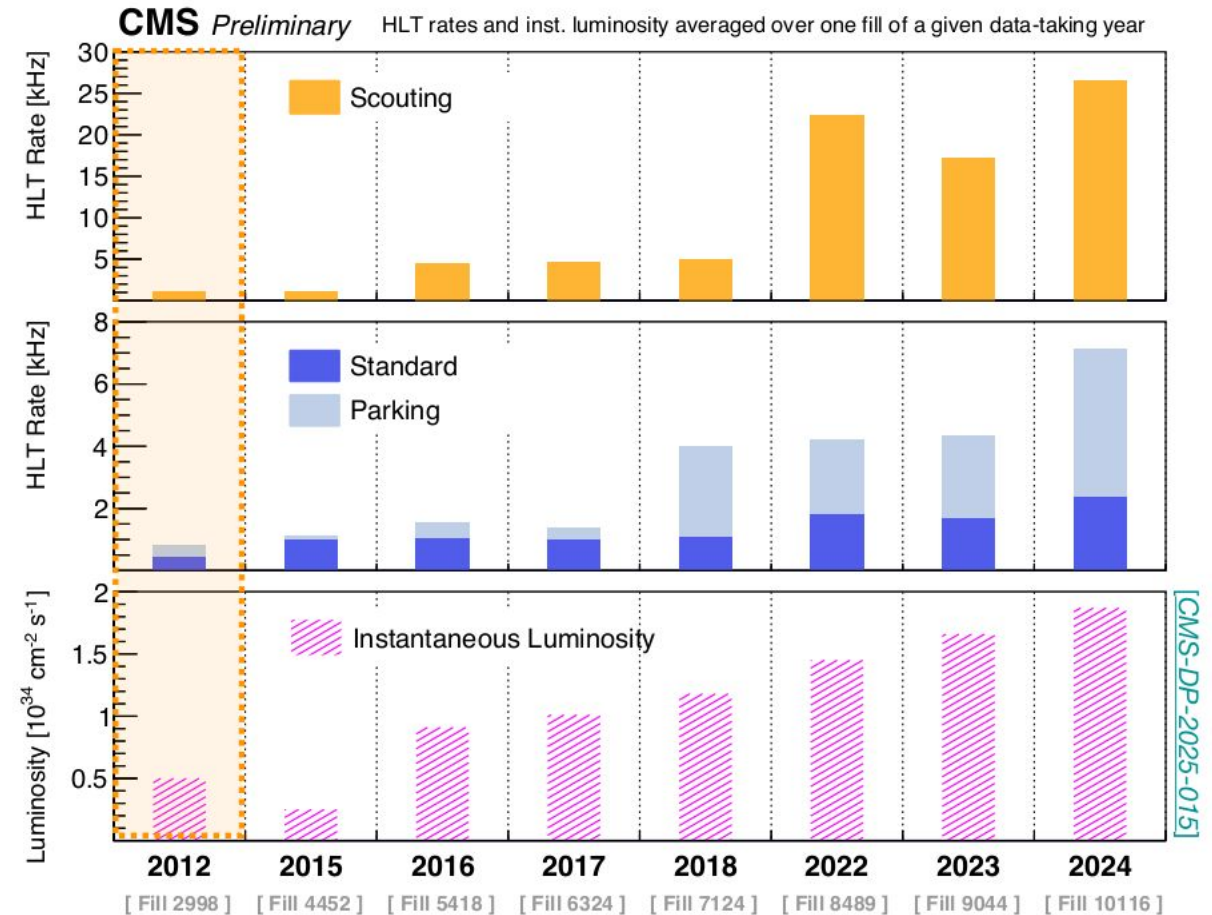
- ▶ recover low-mass multi-jet sensitivity

- **Run-2:**

- ▶ including jets, muons, electrons
- ▶ di-jet and di-muon resonances
- ▶ SM rare decays

- **Run-3:**

- ▶ lower thresholds and new objects
- ▶ 10× standard HLT rate



CMS Data Scouting: a bit of history

- **Main Idea :** “Do more with less” to overcome HLT bandwidth limitations

- ✓ **Increase trigger rate**

Looser (more inclusive) selections

- ✓ **Decrease event size**

Keep only high-level objects from HLT

- **Run-1:**

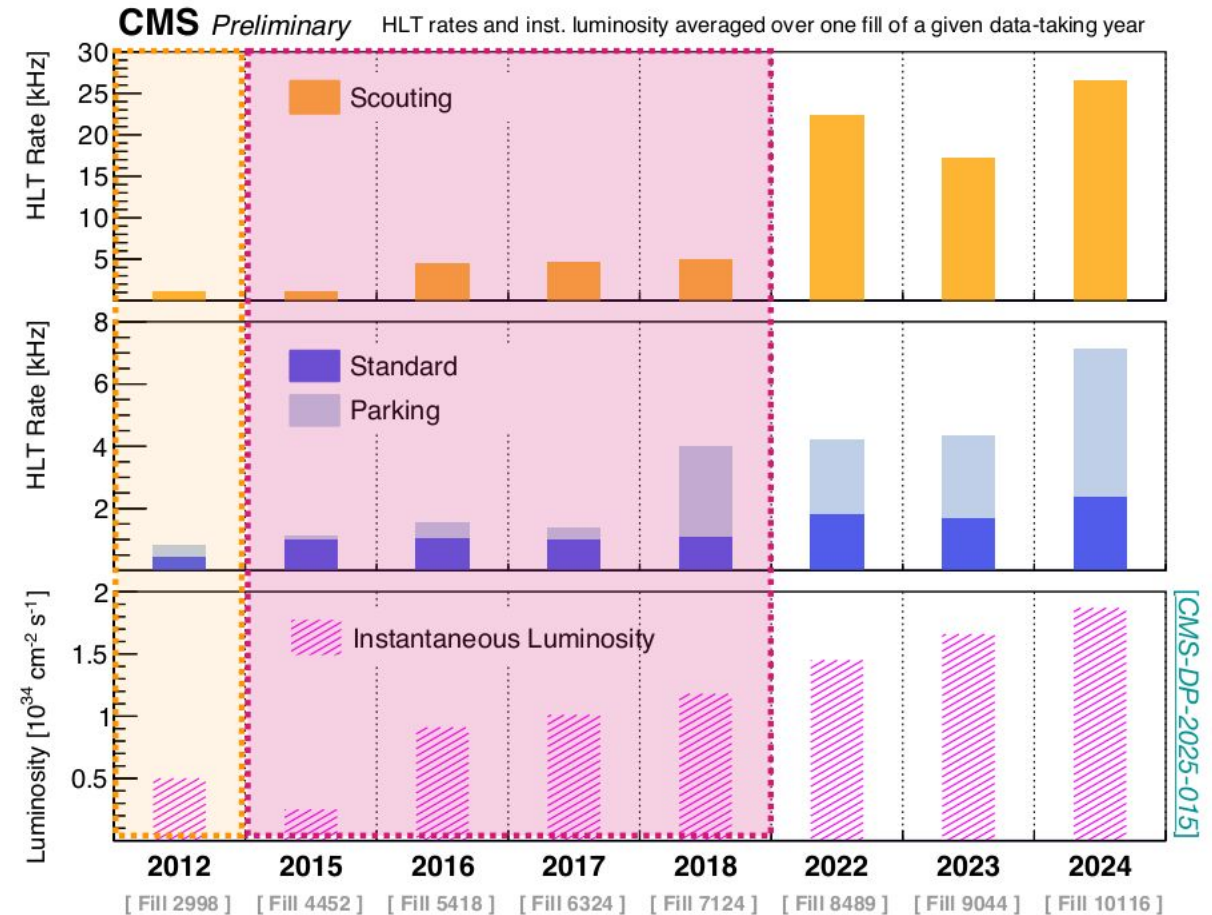
- ▶ recover low-mass multi-jet sensitivity

- **Run-2:**

- ▶ including jets, muons, electrons
- ▶ di-jet and di-muon resonances
- ▶ SM rare decays

- **Run-3:**

- ▶ lower thresholds and new objects
- ▶ 10× standard HLT rate



CMS Data Scouting: a bit of history

- **Main Idea** : “Do more with less” to overcome HLT bandwidth limitations

- ✓ **Increase trigger rate**

Looser (more inclusive) selections

- ✓ **Decrease event size**

Keep only high-level objects from HLT

- **Run-1:**

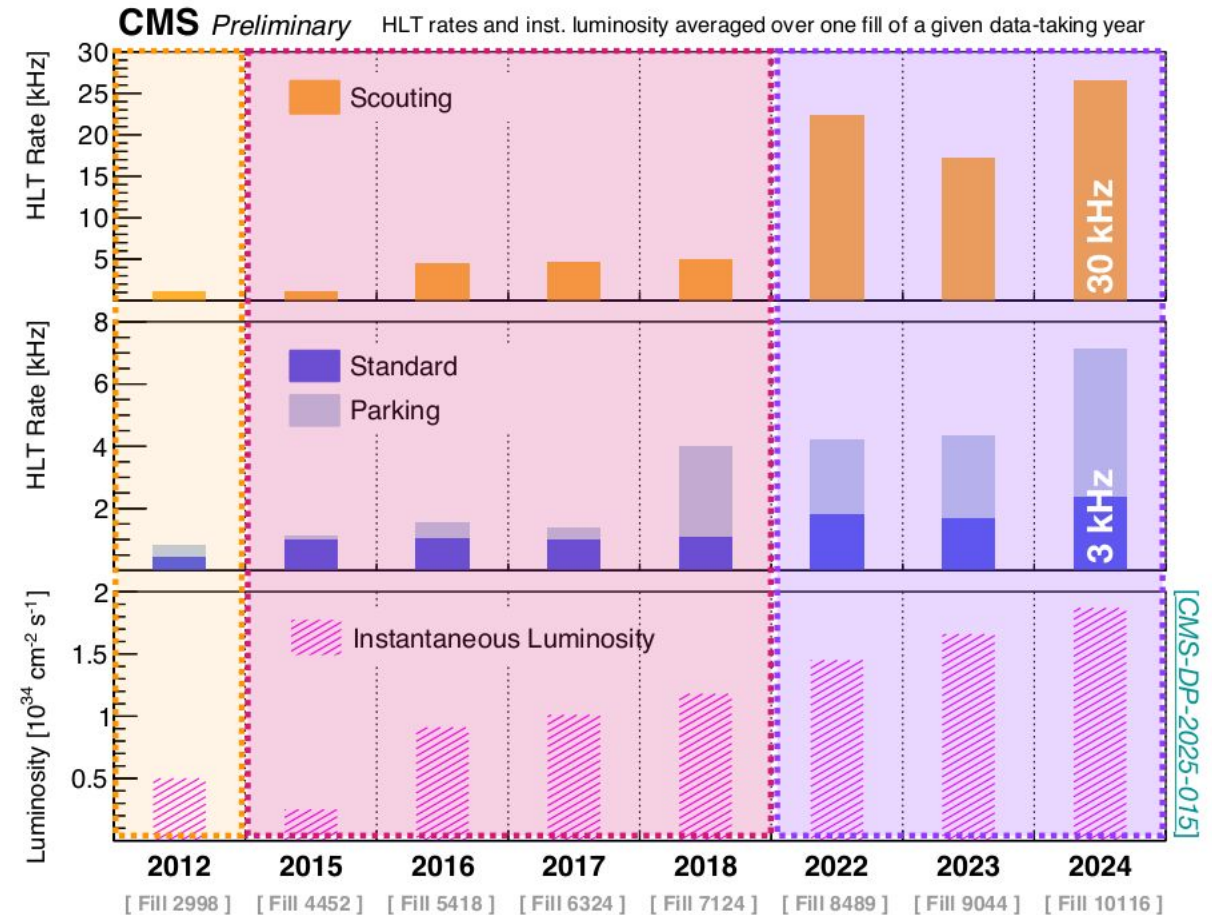
- recover low-mass multi-jet sensitivity

- **Run-2:**

- including jets, muons, electrons
- di-jet and di-muon resonances
- SM rare decays

- **Run-3:**

- lower thresholds and new objects
- 10× standard HLT rate

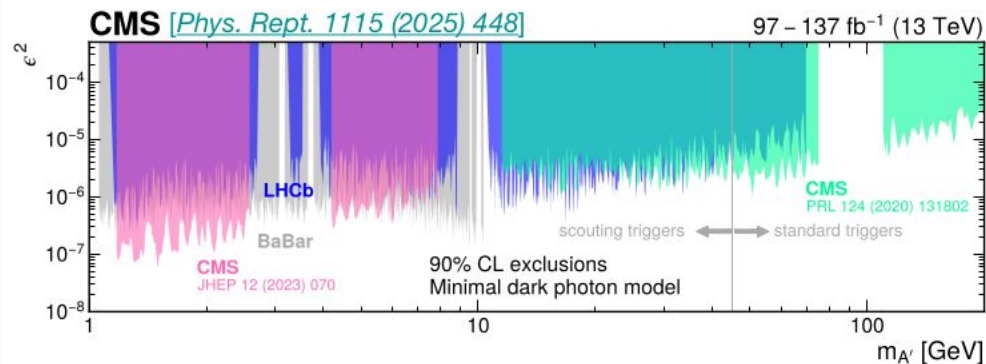


Extending the CMS Physics Reach: Run 2 Scouting

Di-muon searches

• DiMuon Scouting Stream

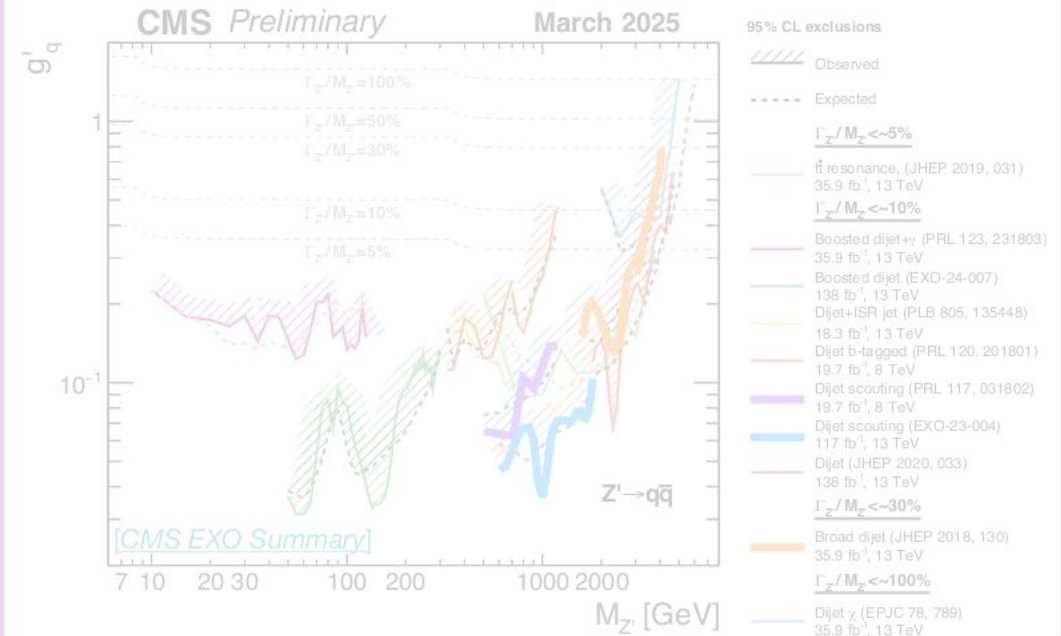
- SM rare decays at low mass:
 - Observation of $\eta \rightarrow 4\mu$
- Searches for prompt and displaced $X_{\mu\mu}$
 - Optimal sensitivity below 10 GeV
 - Access to displaced muons
- CP-violation and B-physics



Di-jet searches

• Jet/HT Scouting Stream

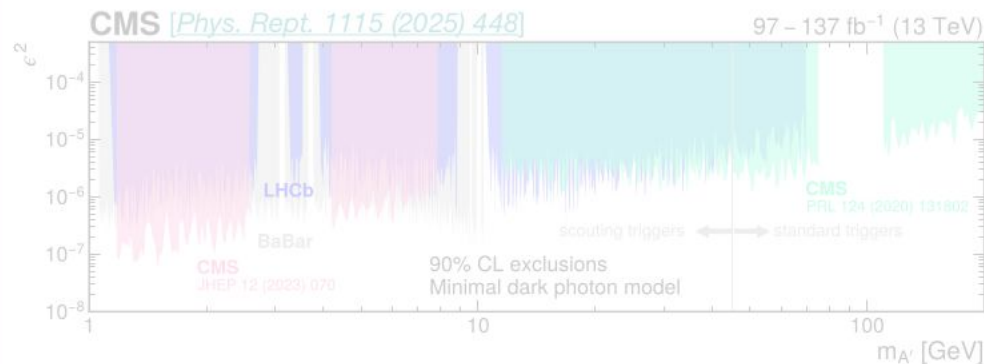
- Resonant searches in hadronic final states
- Sensitivity gain wrt standard triggers



Extending the CMS Physics Reach: Run 2 Scouting

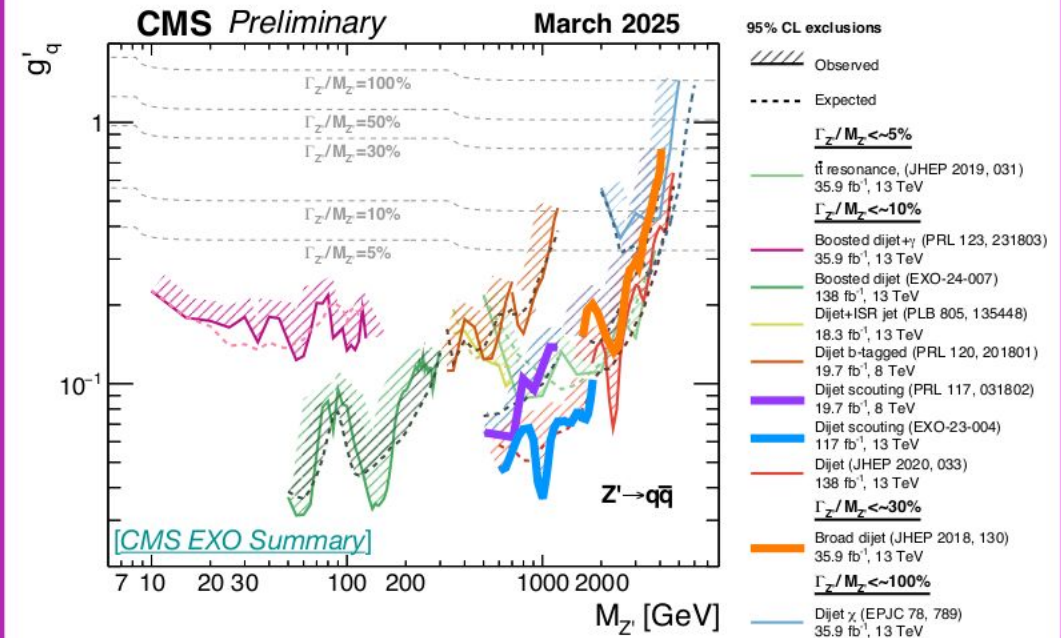
Di-muon searches

- **DiMuon Scouting Stream**
 - SM rare decays at low mass:
 - Observation of $\eta \rightarrow 4\mu$
 - Searches for prompt and displaced $X_{\mu\mu}$
 - Optimal sensitivity below 10 GeV
 - Access to displaced muons
 - CP-violation and B-physics



Di-jet searches

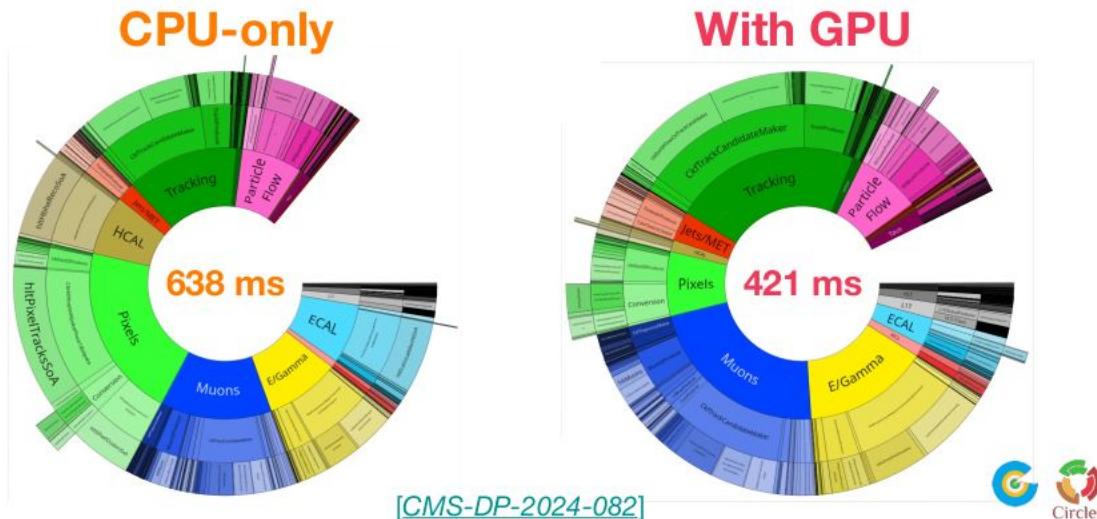
- **Jet/HT Scouting Stream**
 - Resonant searches in hadronic final states
 - Sensitivity gain wrt standard triggers



Extending the CMS Physics Reach: **Run 3** Scouting

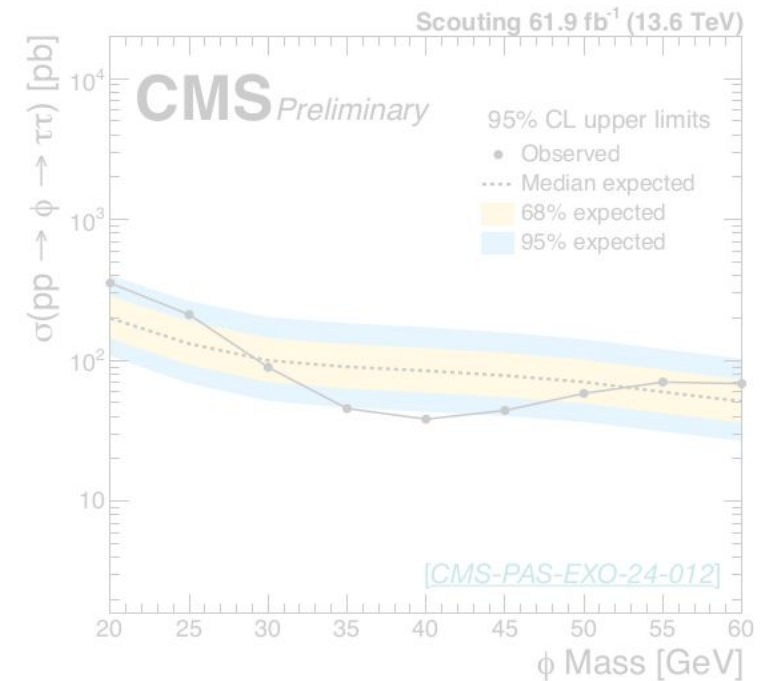
Heterogeneous computing

- **New HLT farm equipped with GPUs**
 - ▶ substantial reduction in HLT processing time
→ over **40%** faster
 - ▶ increased maximum rate that can be processed by the scouting stream
→ **30 kHz** (~30% of HLT input)



Additional objects

- **Complex objects (τ) built offline**
 - ▶ Extend BSM searches in the low-mass regime below 60 GeV



Extending the CMS Physics Reach: **Run 3** Scouting

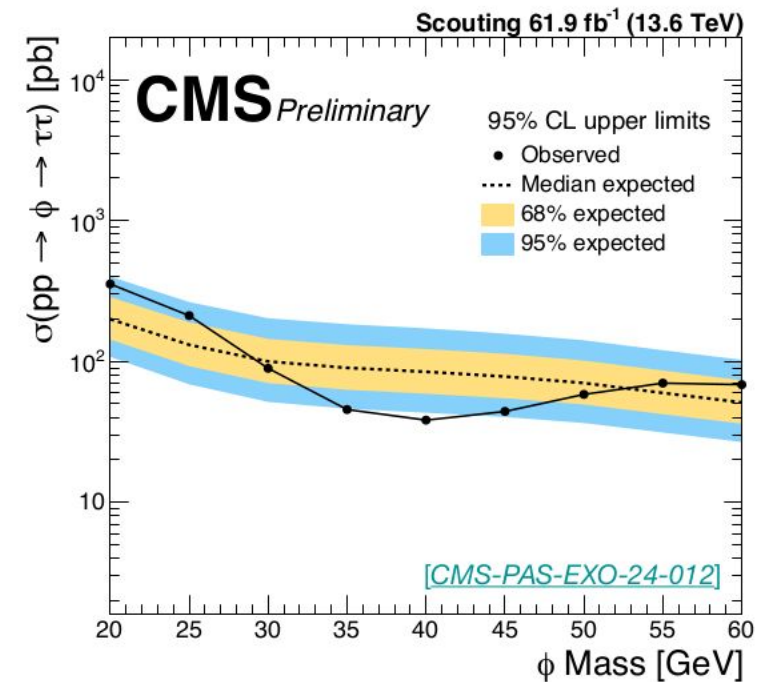
Heterogeneous computing

- New HLT farm equipped with GPUs
 - substantial reduction in HLT processing time
→ over **40%** faster
 - increased maximum rate that can be processed by the scouting stream
→ **30 kHz** (~30% of HLT input)

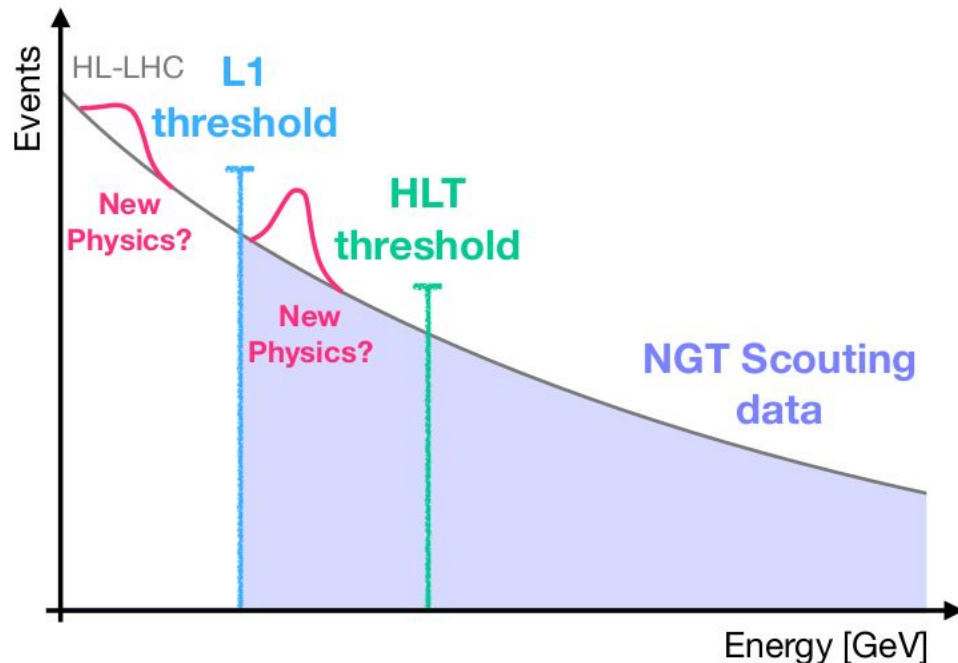
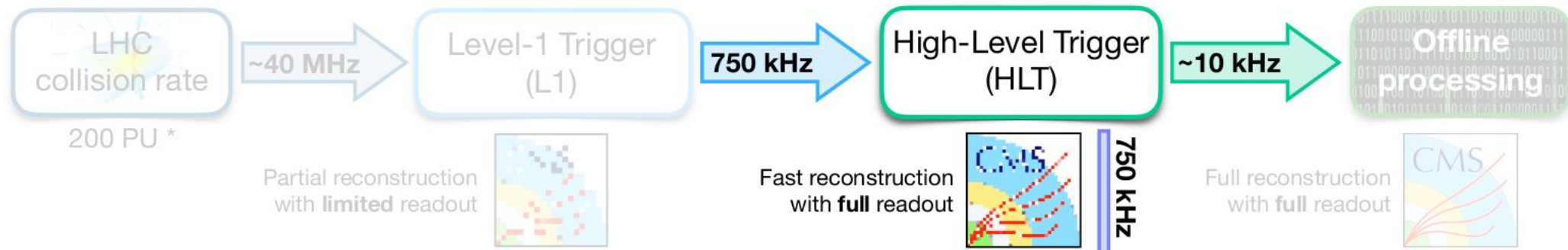


Additional objects

- **Complex objects (τ) built offline**
 - Extend BSM searches in the low-mass regime below 60 GeV



Extending the CMS Physics Reach: **NGT** Scouting



NGT Scouting



NGT Scouting

- Reconstruct @HLT **100%** of L1A events
- Event output rate at **750 kHz**
- Provide **calibrated, offline-like** quality objects
- Analysis-driven output with **reduced event size**



NGT Scouting: How to Make It Happen?

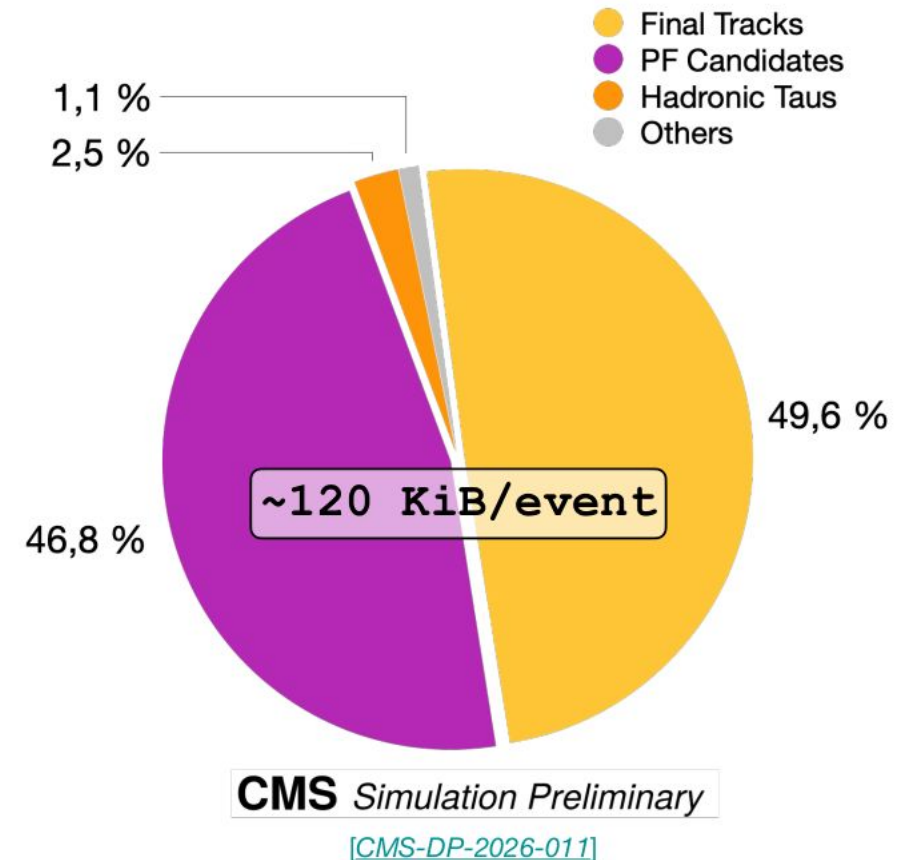
- Main technical **challenges** towards the NGT Scouting approach: **storage**, **calibration**, **computing**

Storage

- Provide lightweight **output format** for analysis
- All relevant HLT objects
 - primary vertices, tracks, PF candidates
 - electrons, photons, muons, taus, jets

120 KiB/event @750 kHz → 88 GiB/s

- Currently **above** the total disk and bandwidth **budget** of ~50 GiB/s
- Size **mitigation** strategies under study
 - Discard low-energy objects
 - Reduce information for PU objects
 - Improve compression factor

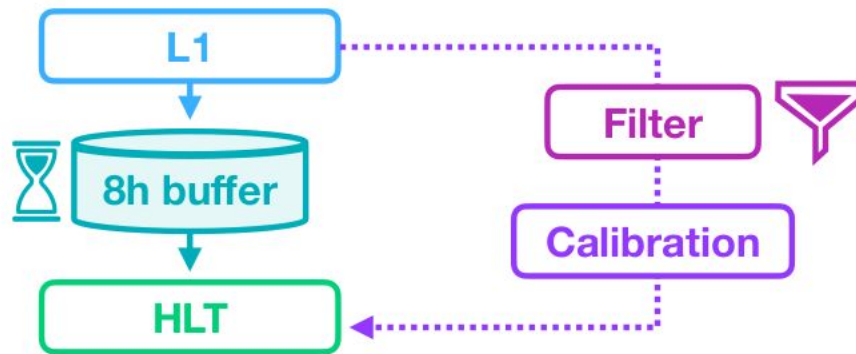


NGT Scouting: How to Make It Happen?

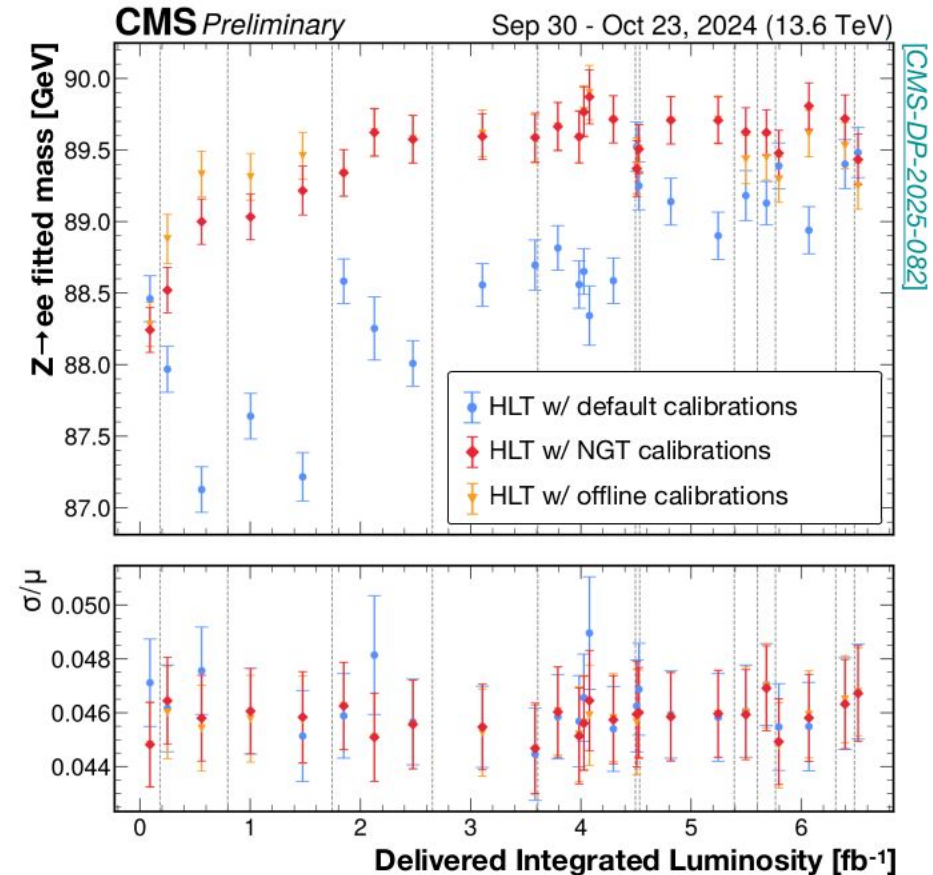
- Main technical **challenges** towards the NGT Scouting approach: **storage**, **calibration**, **computing**

Calibration

- Ensure stable and accurate object calibration
 - More **frequent** update of conditions
 - Derivation on higher **statistics** datasets
- Infrastructure for **8h data buffer** while real-time calibrations are derived and applied on the fly



✓ Validated in Run-3 with **NGT demonstrator** !

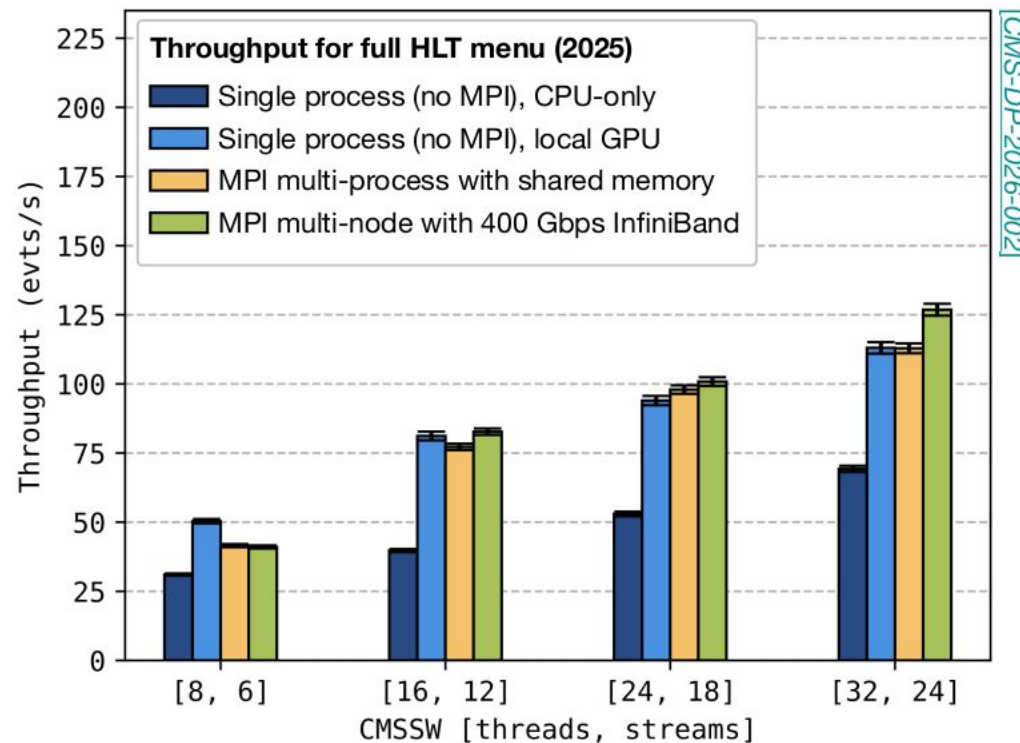


NGT Scouting: How to Make It Happen?

- Main technical **challenges** towards the NGT Scouting approach: **storage**, **calibration**, **computing**

Computing

- Optimize **resource management** for a continuously evolving HLT computing farm



- Run intensive parts of the event processing on heterogeneous architectures
 - **50%** of code run on GPUs in Run-4
 - **80%** of code run on GPUs in in Run-5
- Current GPU use limited to **local** resources
 - Run-3 node: 2 × AMD Milan 64-core CPUs + 2 × NVIDIA Tesla T4 GPUs
- Enable offloading to **remote** resources with *Message Passing Interface (MPI)*
 - Same or higher event throughput with **scalable and flexible** computing model

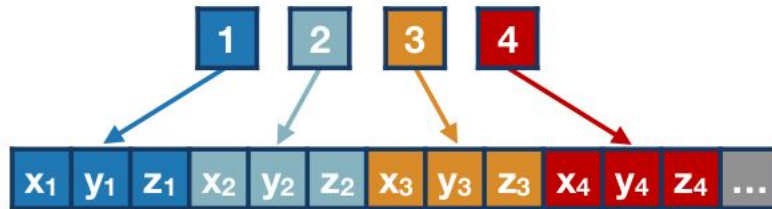
NGT Scouting: How to Make It Happen?

- Main technical **challenges** towards the NGT Scouting approach: **storage**, **calibration**, **computing**

Computing

- Develop **data-oriented structures** for better memory access in parallel processing

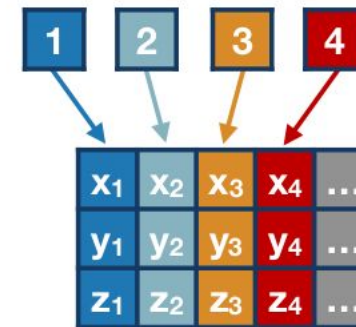
Array of Struct (AoS)



- Most natural way to represent data in most languages is object-by-object
- AoS become **inefficient** with GPUs
 - Poor parallel memory access
 - Long processing time

VS

Structure of Array (SoA)



- SoA stores data quantity-by-quantity
- Less intuitive, but better cache locality
- Efficient** solution for GPU processing

NGT Scouting: How to Make It Happen?

- Main technical **challenges** towards the NGT Scouting approach: **storage**, **calibration**, **computing**

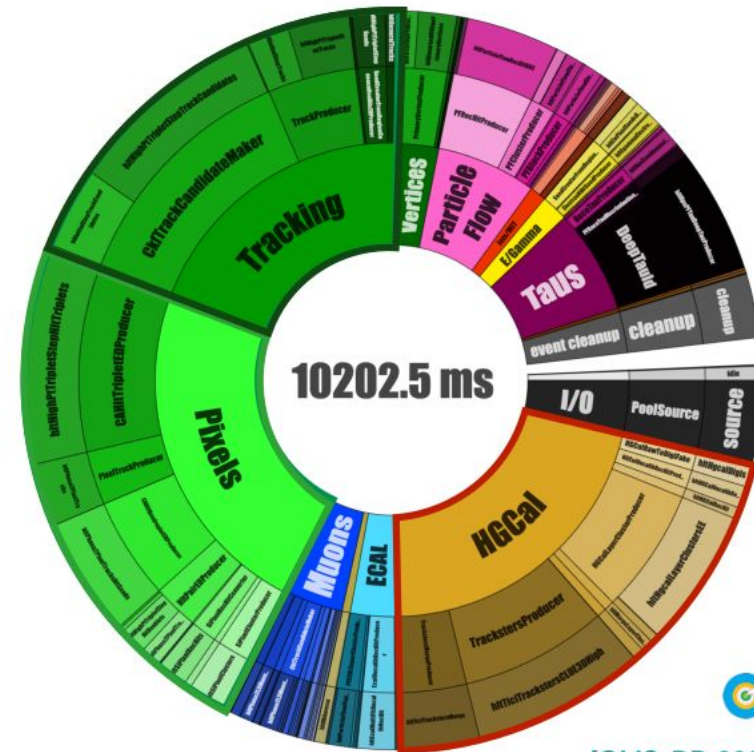
Computing

- Fit full-event reconstruction algorithms within available time/computing **budget**

- **HLT timing** for $t\bar{t}$ events in 200 PU (CPU-only)

Average Time = ~10 s

- **Pixel** : 23%
- **Tracking** : 19%
- **HGCal** : 21%
- Dominated by algorithms operating with
 - high detector granularity
 - high object multiplicity
- **Redesign** and **accelerate** most expensive steps
 - Track reconstruction = **4.4 s**



[CMS-DP-2026-011]

NGT Scouting: How to Make It Happen?

- Main technical **challenges** towards the NGT Scouting approach: **storage**, **calibration**, **computing**

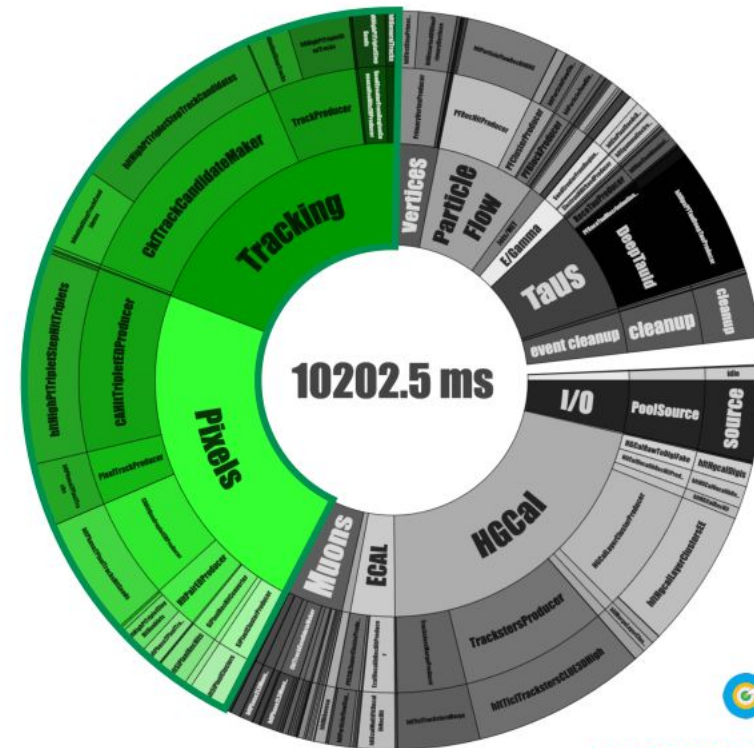
Computing

- Fit full-event reconstruction algorithms within available time/computing **budget**

- **HLT timing** for $t\bar{t}$ events in 200 PU (CPU-only)

Average Time = ~10 s

- **Pixel** : 23%
- **Tracking** : 19%
- **HGCal** : 21%
- Dominated by algorithms operating with
 - high detector granularity
 - high object multiplicity
- **Redesign** and **accelerate** most expensive steps
 - Track reconstruction = **4.4 s**



[CMS-DP-2026-011]

The CMS Track Reconstruction

- To resolve the high hit **combinatorics**, tracks are reconstructed sequentially via:
 - **Track seeding**: initial track candidates are built using hits from **inner layers**
→ *Cellular-Automaton (CA)*
 - **Track building**: track candidates are extended into final tracks adding hits from **outer layers**
→ *Combinatorial Kalman Filter (CKF)*

Outer Tracker (OT)

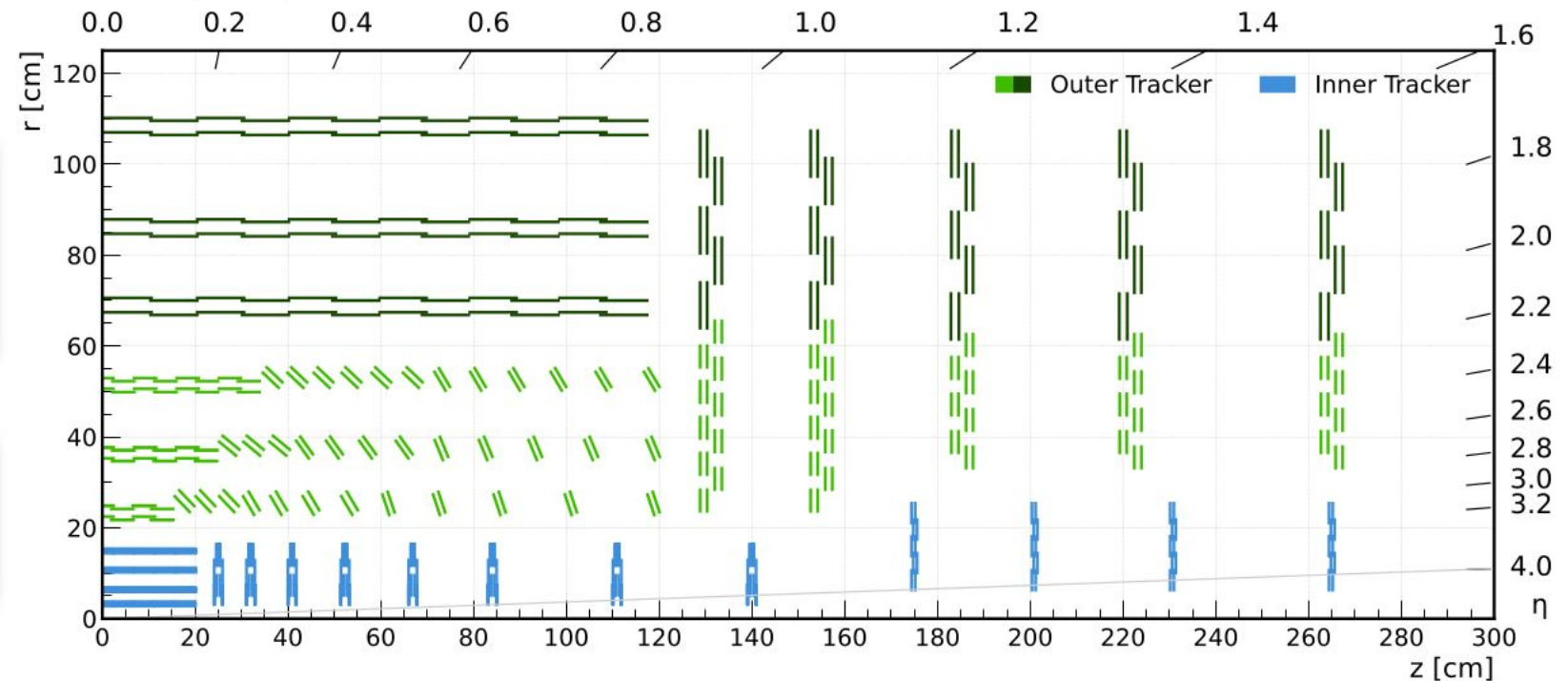
(6 barrel + 2×5 endcap disks)

- ▶ Pixel-strip (PS) modules
- ▶ Two-strips (2S) modules

Inner Tracker (IT)

(4 barrel + 2×12 endcap disks)

- ▶ Pixel modules



The CMS Track Reconstruction

- To resolve the high hit **combinatorics**, tracks are reconstructed sequentially via:
 - **Track seeding**: initial track candidates are built using hits from **inner layers**
→ *Cellular-Automaton (CA)*
 - **Track building**: track candidates are extended into final tracks adding hits from **outer layers**
→ *Combinatorial Kalman Filter (CKF)*

Outer Tracker (OT)

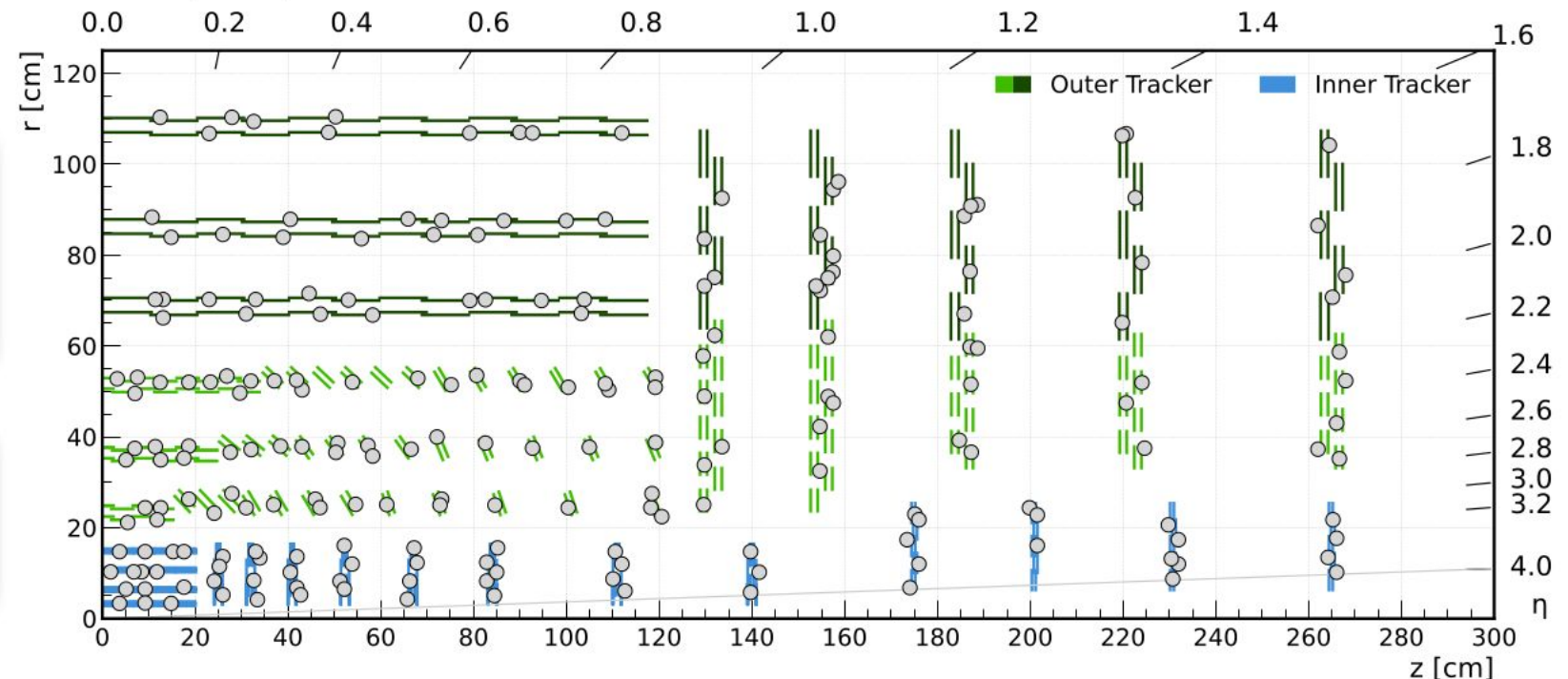
(6 barrel + 2×5 endcap disks)

- ▶ Pixel-strip (PS) modules
- ▶ Two-strips (2S) modules

Inner Tracker (IT)

(4 barrel + 2×12 endcap disks)

- ▶ Pixel modules



The CMS Track Reconstruction

- To resolve the high hit **combinatorics**, tracks are reconstructed sequentially via:
 - **Track seeding**: initial track candidates are built using hits from **inner layers**
→ *Cellular-Automaton (CA)*
 - **Track building**: track candidates are extended into final tracks adding hits from **outer layers**
→ *Combinatorial Kalman Filter (CKF)*

Outer Tracker (OT)

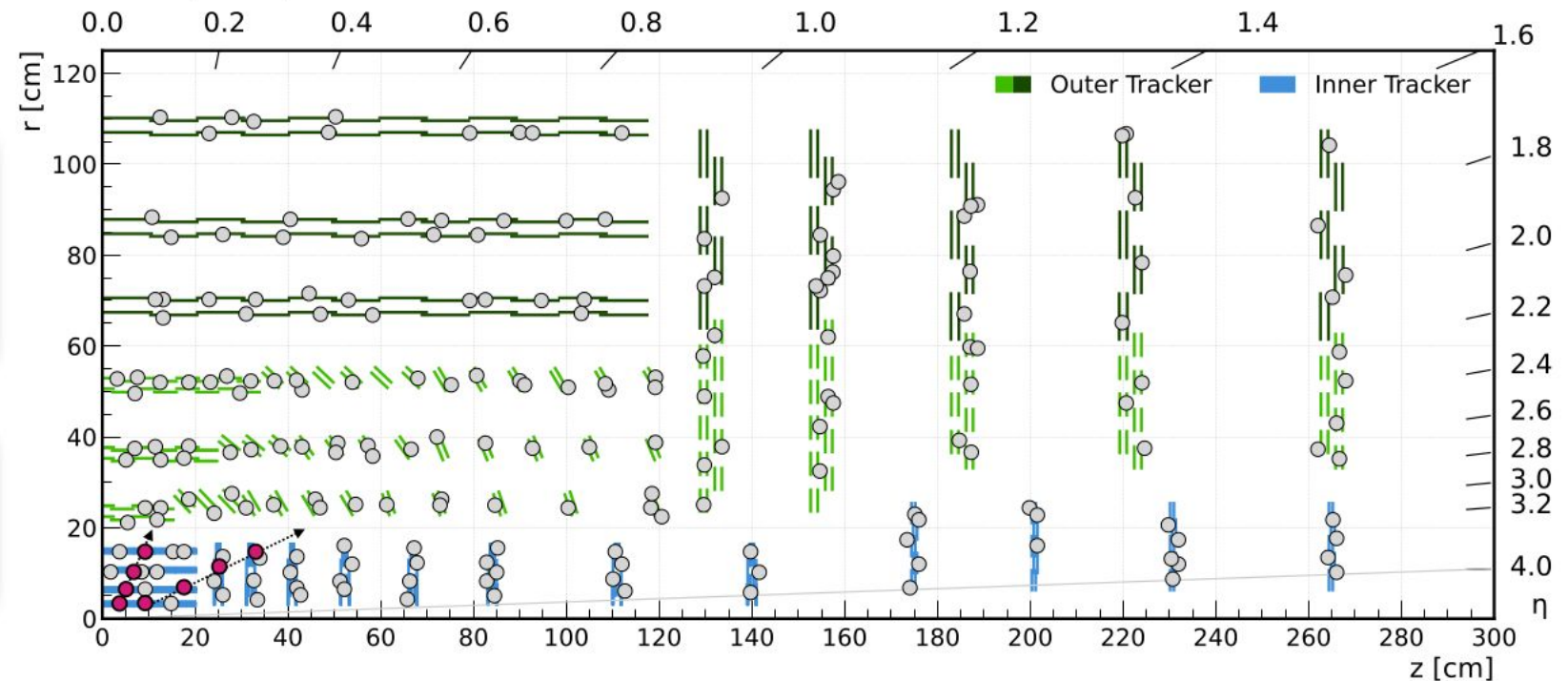
(6 barrel + 2×5 endcap disks)

- ▶ Pixel-strip (PS) modules
- ▶ Two-strips (2S) modules

Inner Tracker (IT)

(4 barrel + 2×12 endcap disks)

- ▶ Pixel modules



The CMS Track Reconstruction

- To resolve the high hit **combinatorics**, tracks are reconstructed sequentially via:
 - **Track seeding**: initial track candidates are built using hits from **inner layers**
→ *Cellular-Automaton (CA)*
 - **Track building**: track candidates are extended into final tracks adding hits from **outer layers**
→ *Combinatorial Kalman Filter (CKF)*

Outer Tracker (OT)

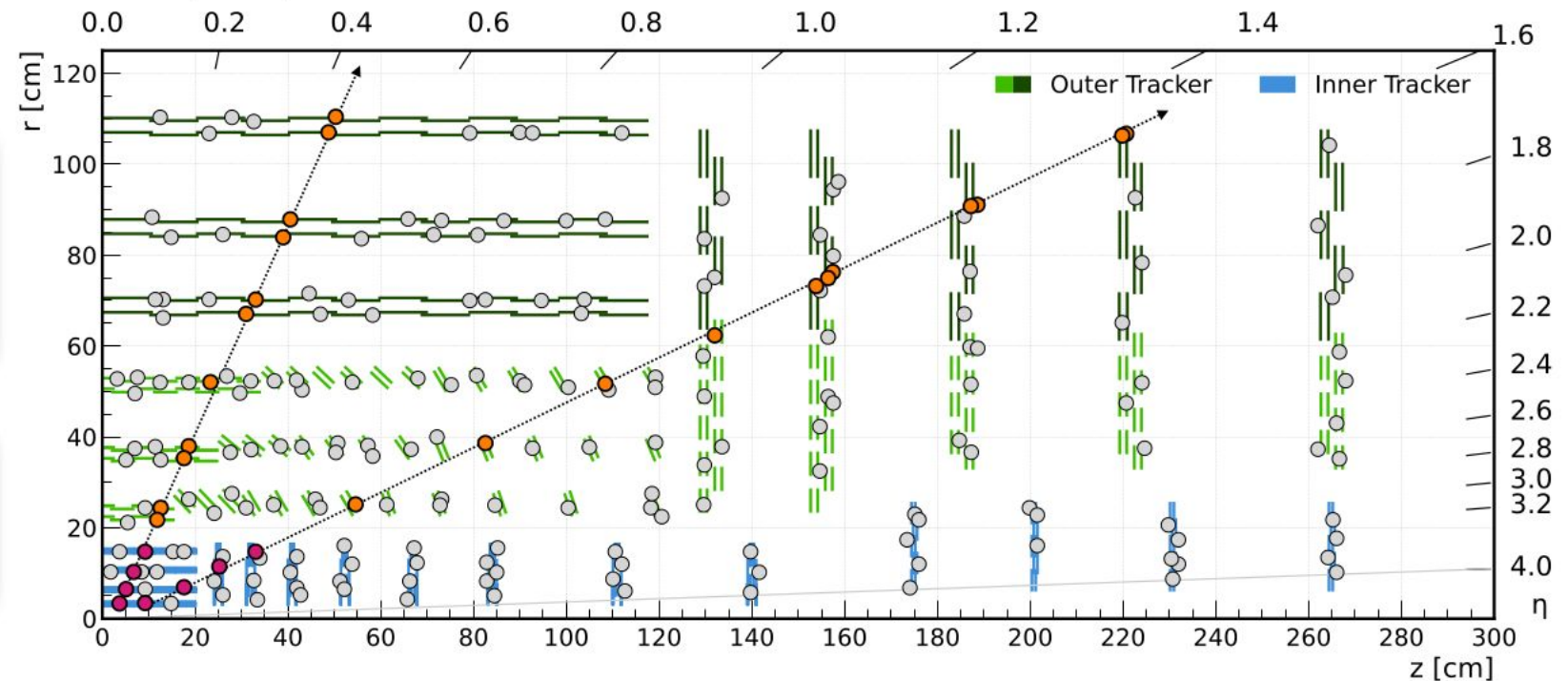
(6 barrel + 2×5 endcap disks)

- ▶ Pixel-strip (PS) modules
- ▶ Two-strips (2S) modules

Inner Tracker (IT)

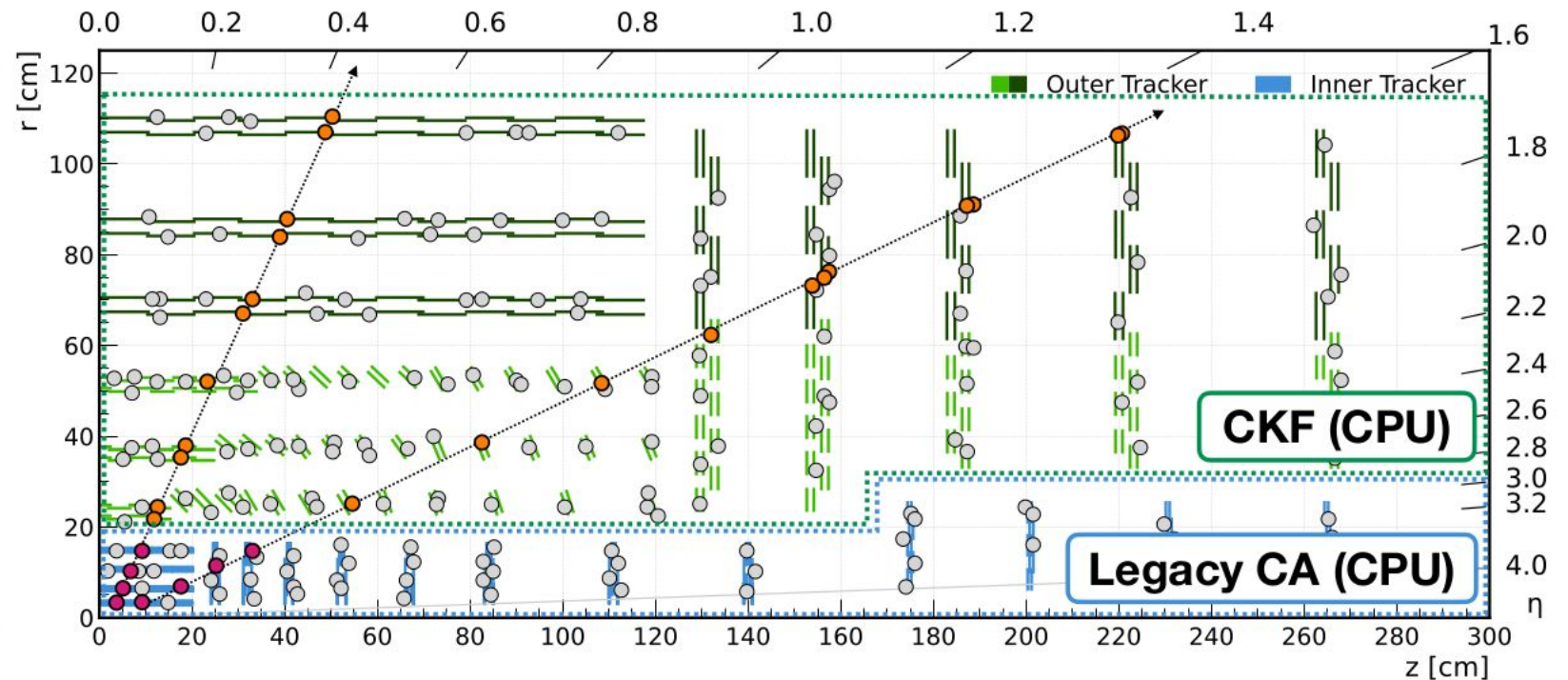
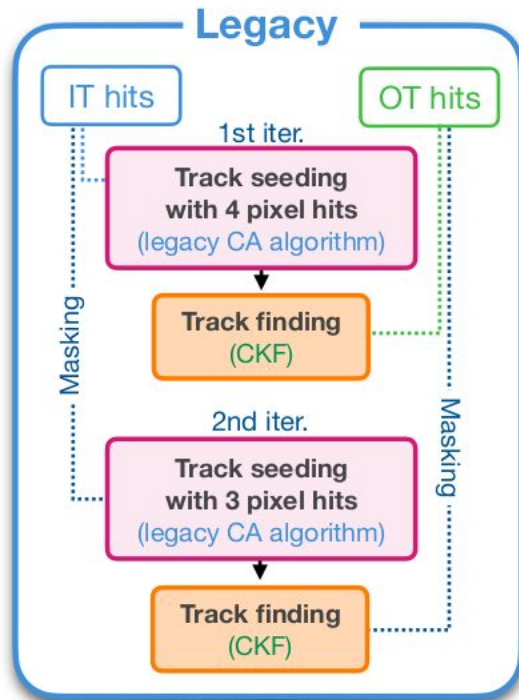
(4 barrel + 2×12 endcap disks)

- ▶ Pixel modules



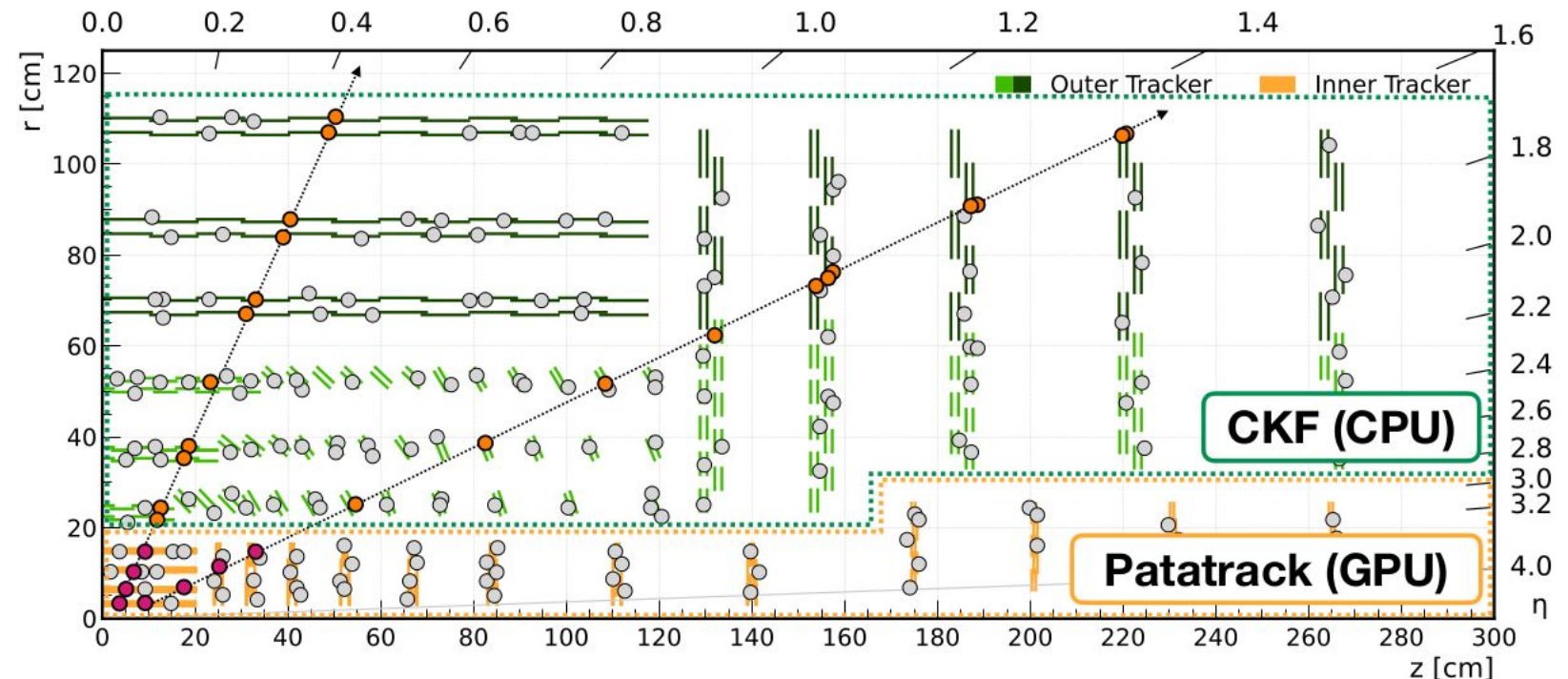
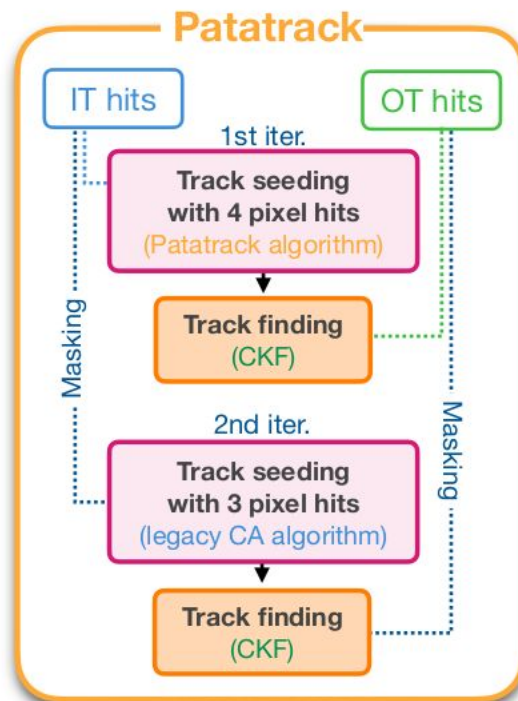
Track Reconstruction at HLT: Legacy

- In the legacy reconstruction of HLT final tracks, **2 iterations** are executed
1st iter : **track seeding** with **4 pixel hits** + **track building**
2nd iter : after masking, **track seeding** with **3 pixel hits** + **track building**
- **BUT** track seeding with CA is inefficient → not sustainable in HL-LHC high PU environment



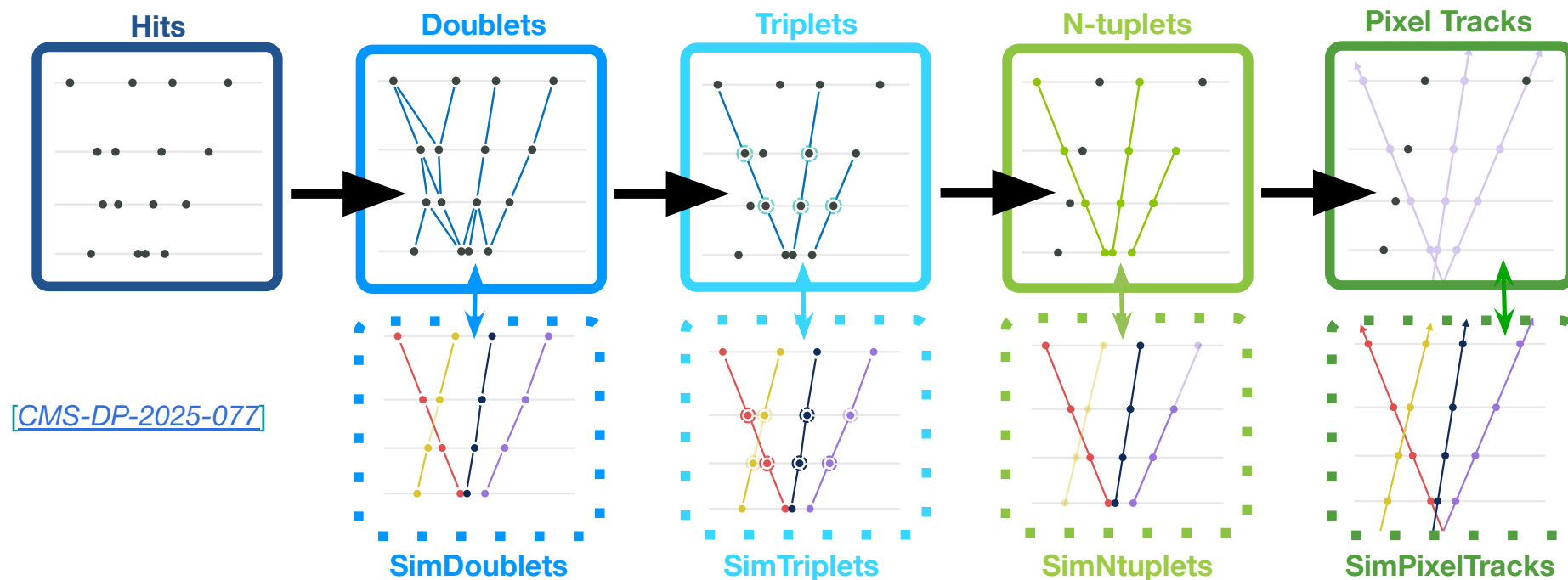
Track Reconstruction at HLT: Patatrack

- The **Patatrack** algorithm is used to accelerate the **track seeding** step (already online in Run-3)
 - GPU parallelisation implemented with performance portability library [alBaka](#)
- A new **Patatrack** configuration is investigated for the NGT scouting:
 - same **2 iterations**, but **1st iter** legacy CA replaced by **Patatrack** algorithm (GPU compatibility)



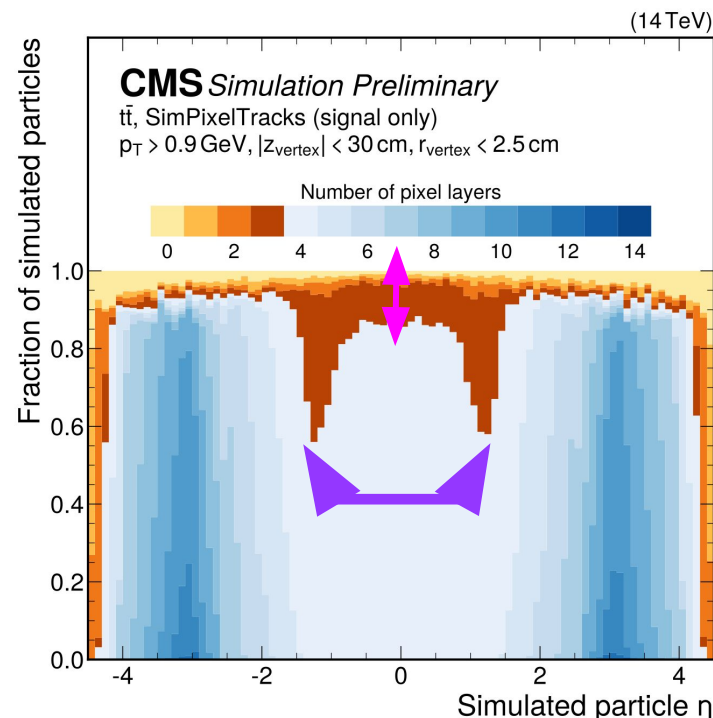
Simulation truth for pixel track reconstruction

- [Patatrack](#) is a Cellular-Automaton (CA) algorithm which builds tracks by **connecting pixel hits**
- In each step compatibility checks are used to resolve combinatorics, depending on **O(100) parameters** to be tuned
 - ? **Automated tuning** targeting pixel track final performance: long convergence time, no insight into the algorithm
 - ✓ **Performance evaluation** at **each step** of reconstruction with simulation truth information:
 - Deeper analysis and control on the parameters, potentially improving the algorithm itself
 - Definition of **new simulation truth objects** corresponding to intermediate steps of reconstruction



Simulation truth for pixel track reconstruction

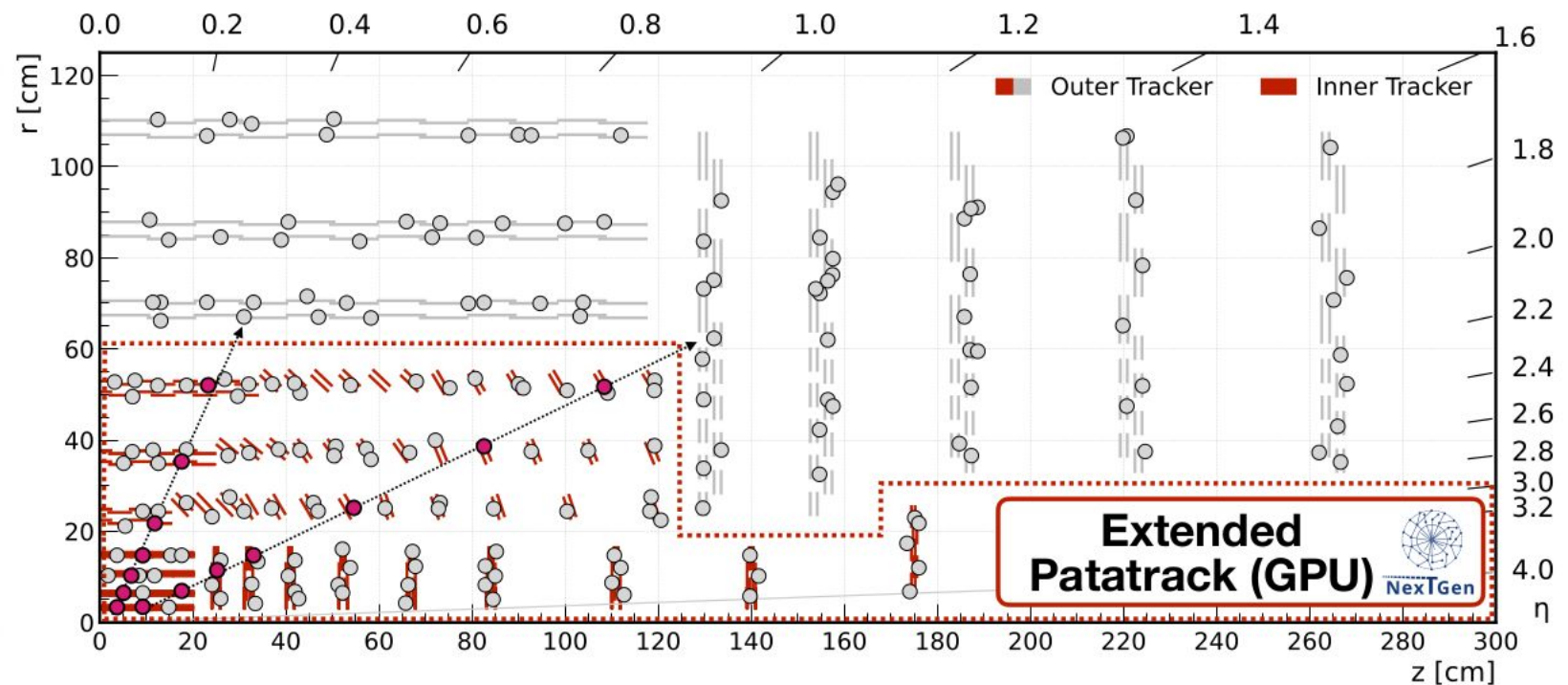
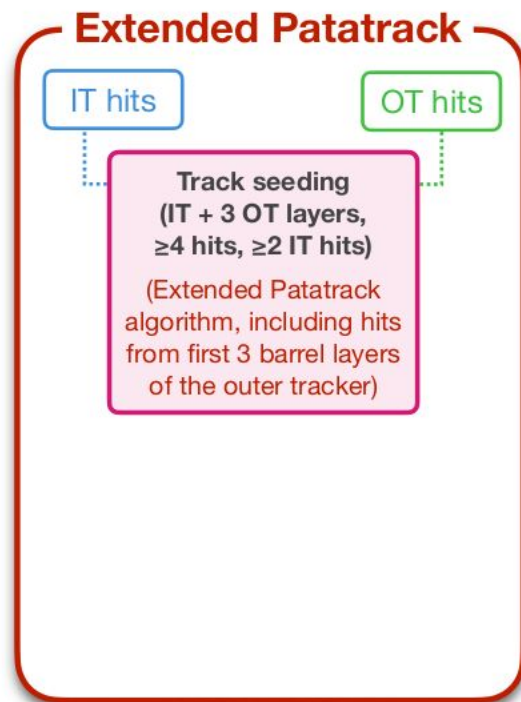
- Development of performance evaluation at each step enabled optimisation of each reconstruction step individually
 - **SimDoublets**: establish correct **layer pairs connection** (missing between barrel and very forward)
 - **SimTriplets**: set **parameters** according to simulation truth distributions
 - **SimNtuplets**: identify **source of inefficiency** in the barrel and transition regions



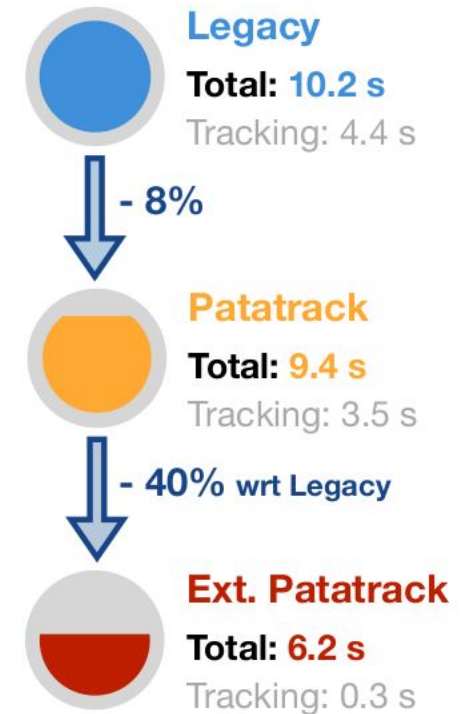
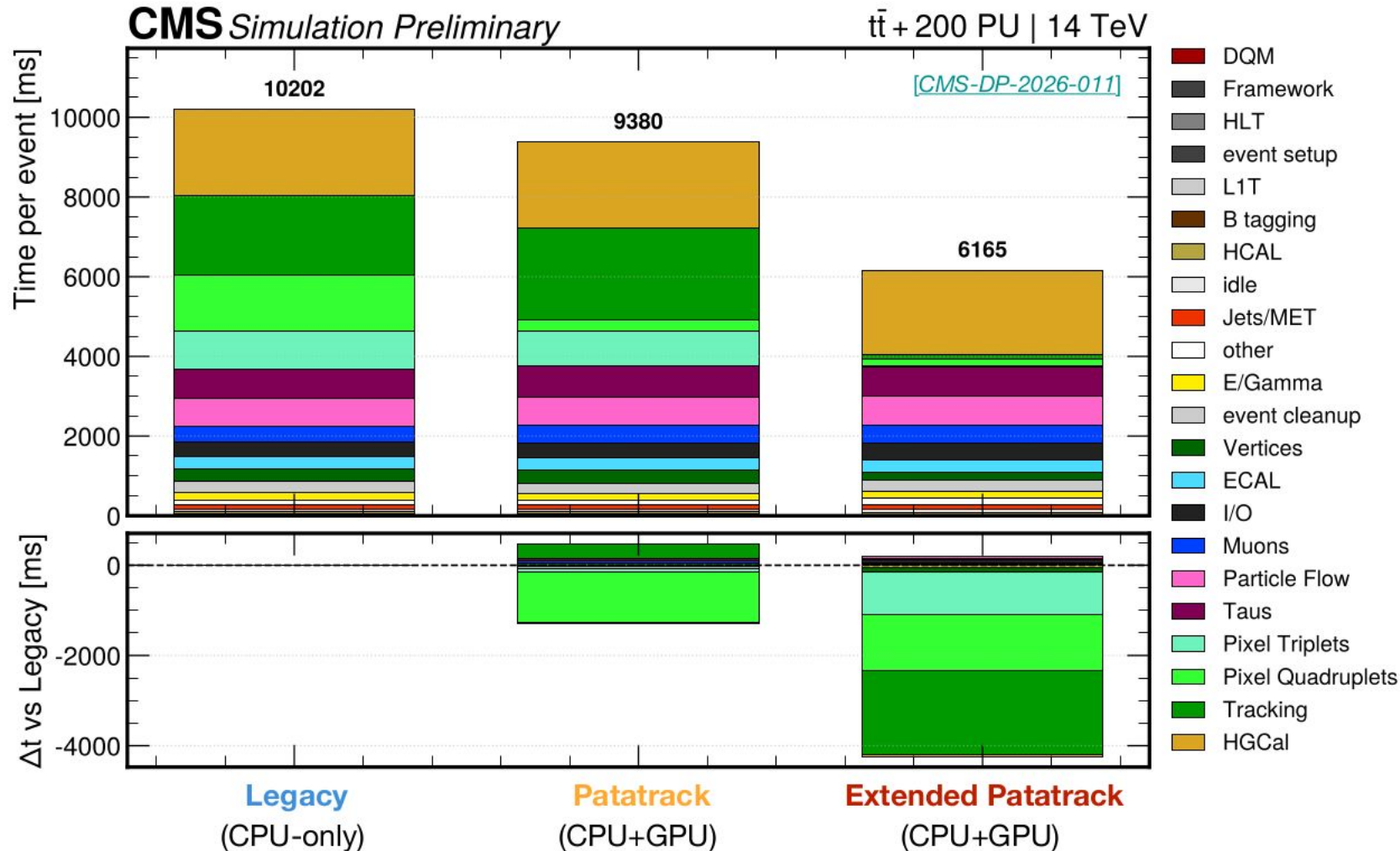
- Pixel Track reconstruction requires **at least 4 hits**, but
 - ▶ **Barrel region**: ~10% of particles generate ≤ 3 hits
 - ▶ **Barrel-Endcap transition**: ~40% of particles generate ≤ 3 hits
- **Intrinsic inefficiency** in the current configuration
- Main reason for **second tracking iteration** with triplet seeds
- Potential solutions for efficiency **recovery**:
 - Rely on second iteration (but additional timing/CPU cost) ✗
 - Require less hits (but exploding fake rate @200 PU) ✗
 - **Collect more hits from outer tracker** ✓

Track Reconstruction at HLT: **Extended Patatrack**

- The Patatrack-based **track seeding** step is extended to **include hits from the OT pixel strips**
 - better detector coverage (3 additional hits in barrel), efficiency recovery and improved resolution
- An alternative fast **Extended Patatrack** configuration is investigated for the NGT scouting:
 - only track seeding** with the Extended Patatrack, **no track building** step (GPU compatibility)

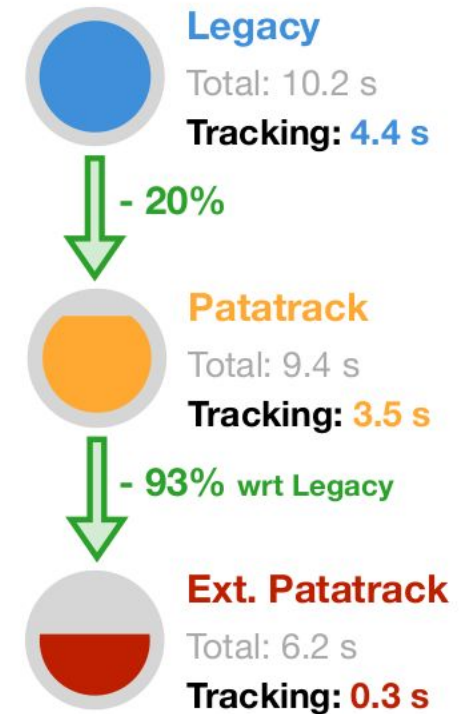
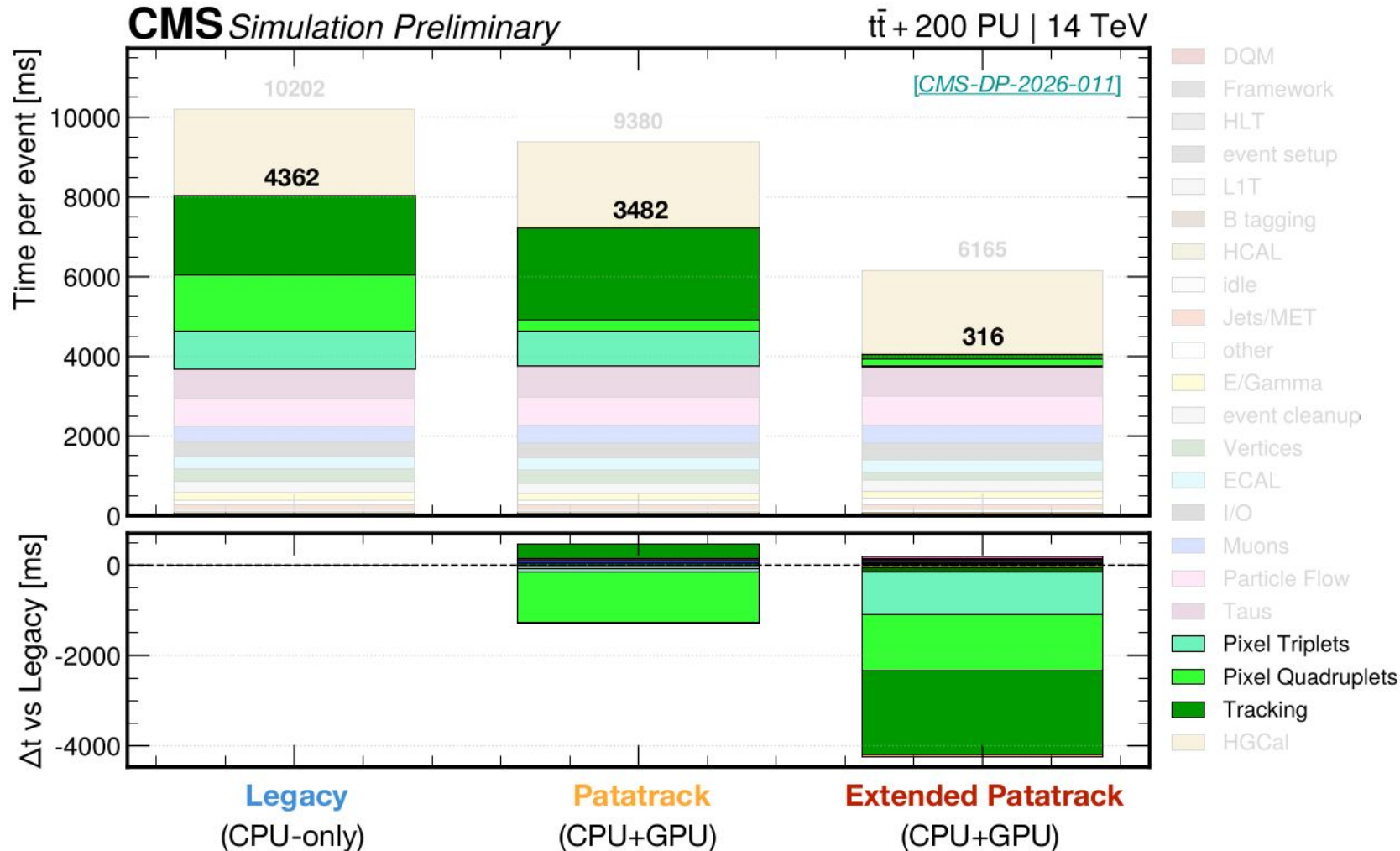


HLT Timing with Extended Patatrack



**40% faster with
Extended Patatrack !!**

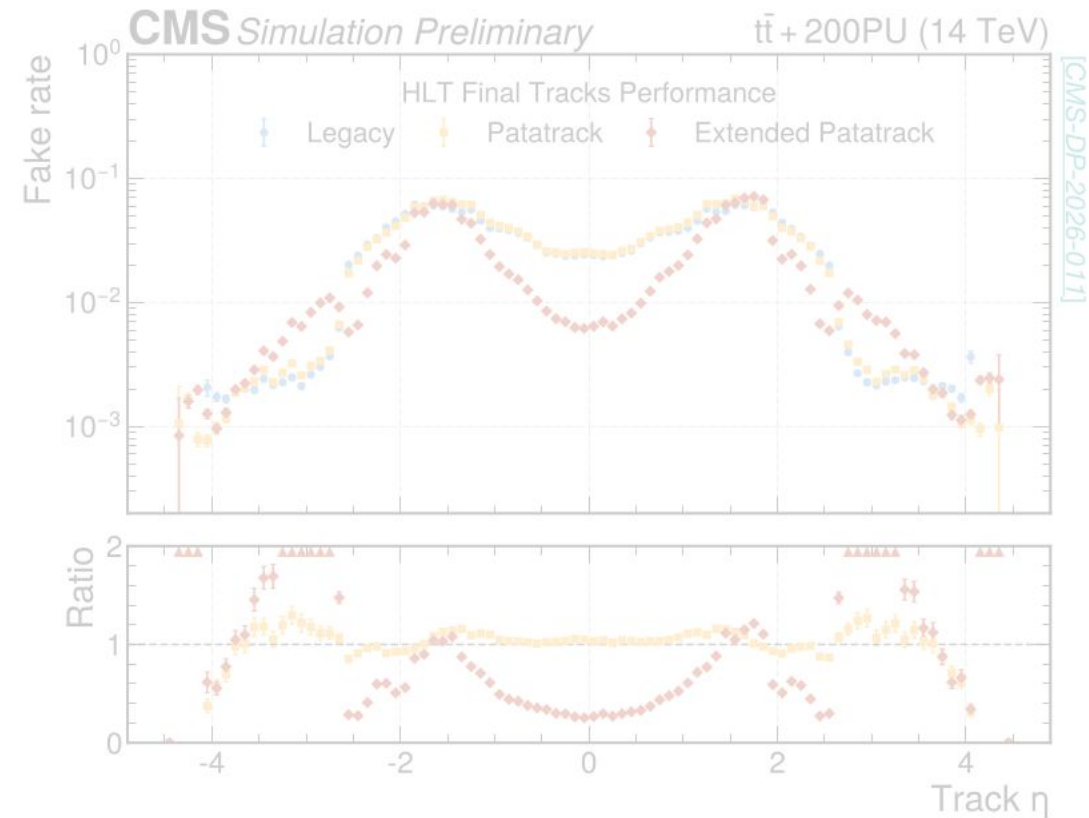
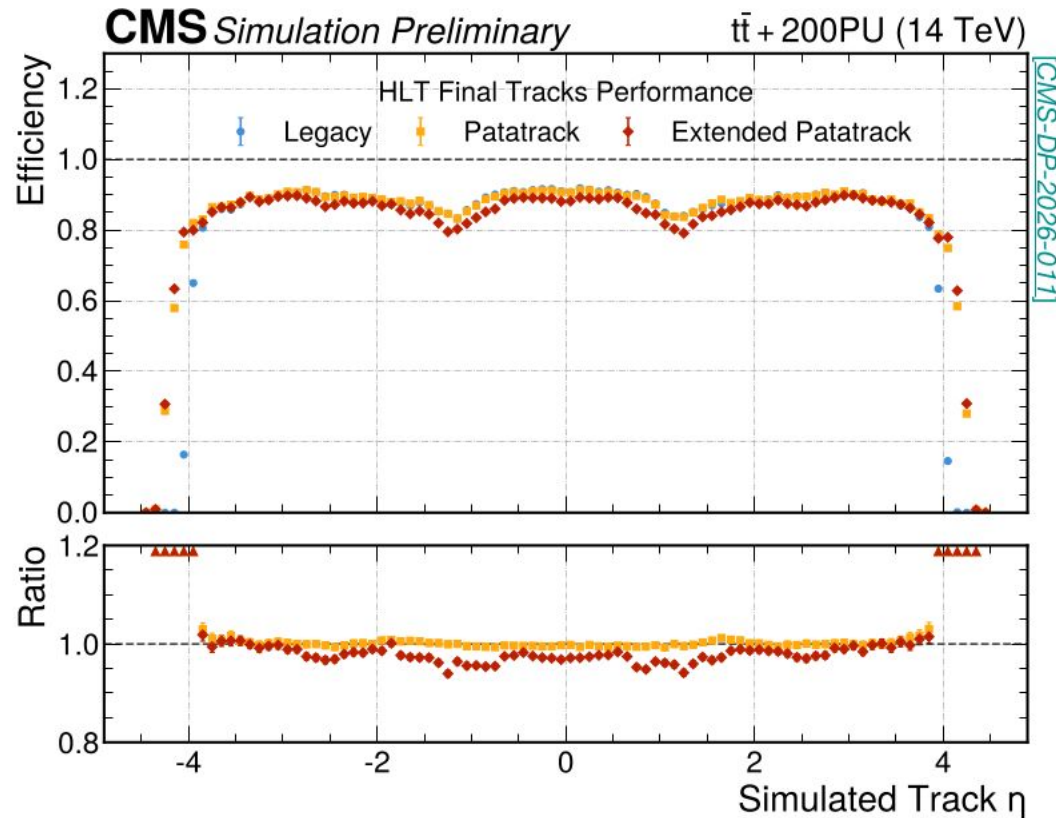
HLT Timing with Extended Patatrack



**40% faster with
Extended Patatrack !!**

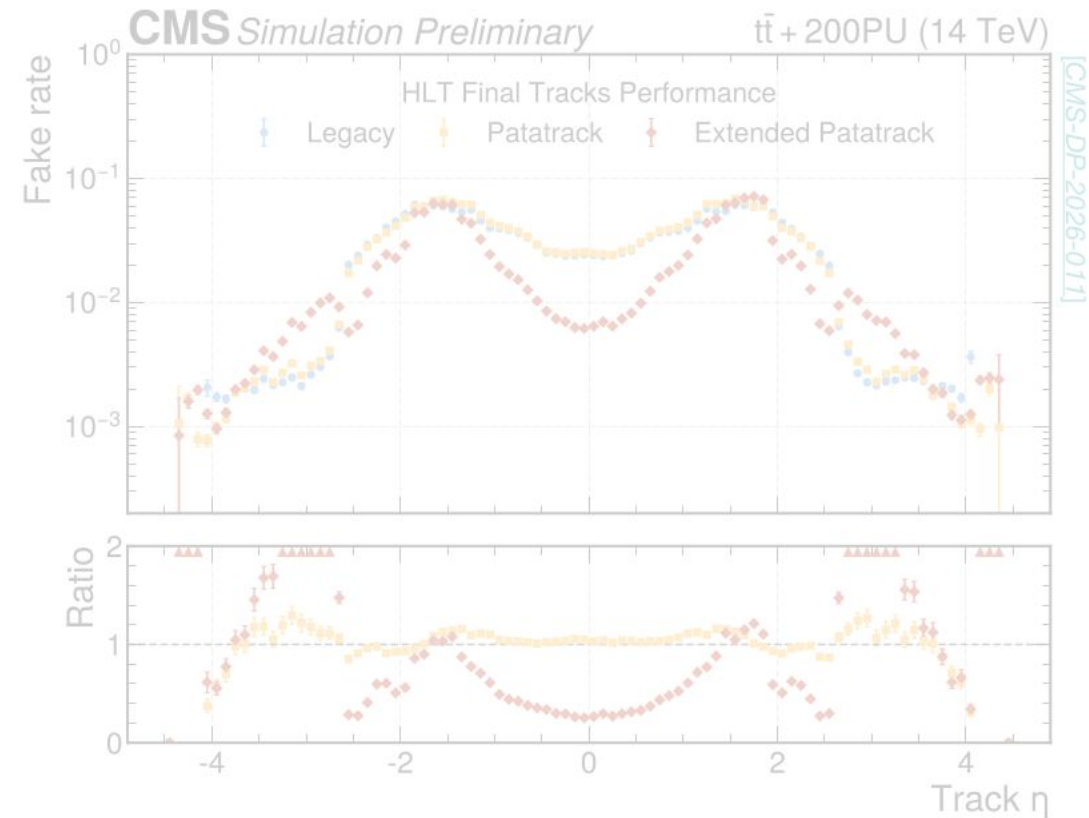
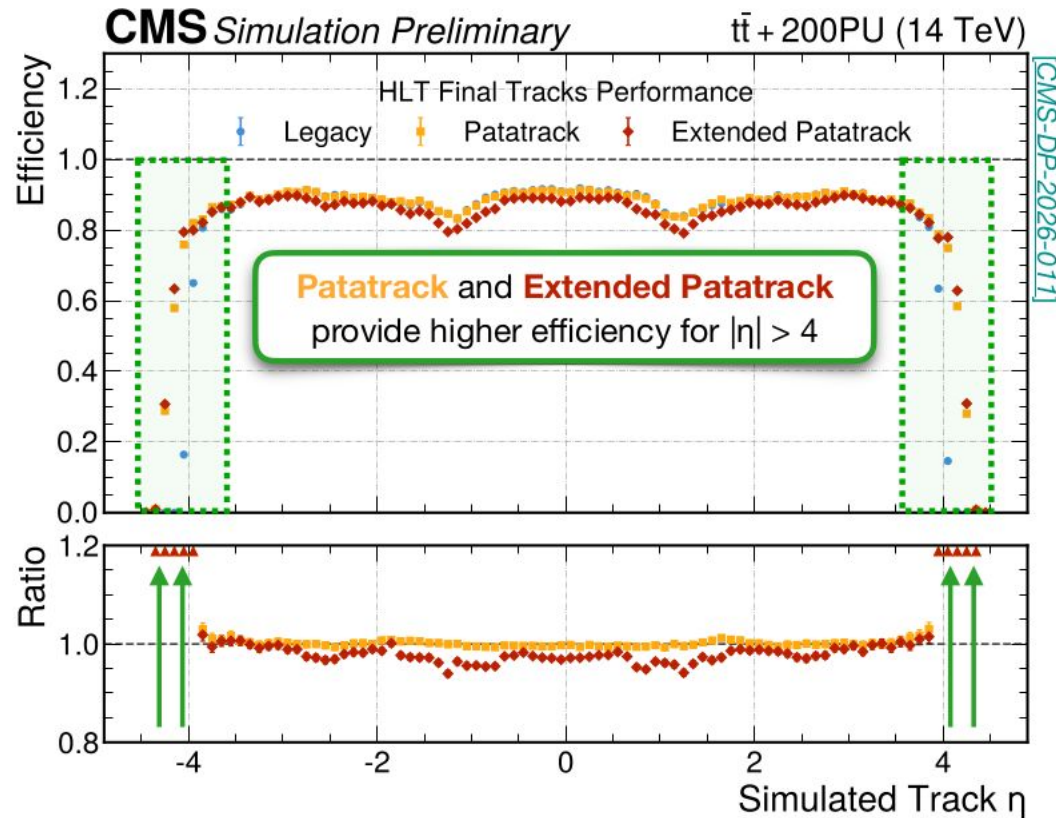
Tracks Performance with **Extended Patatrack**

- The alternative fast **Extended Patatrack** configuration maintains the track physics performance
 - Efficiency:** fraction of simulated tracks matched to at least one reconstructed track
 - Fake rate:** fraction of reconstructed tracks not matched to any simulated track



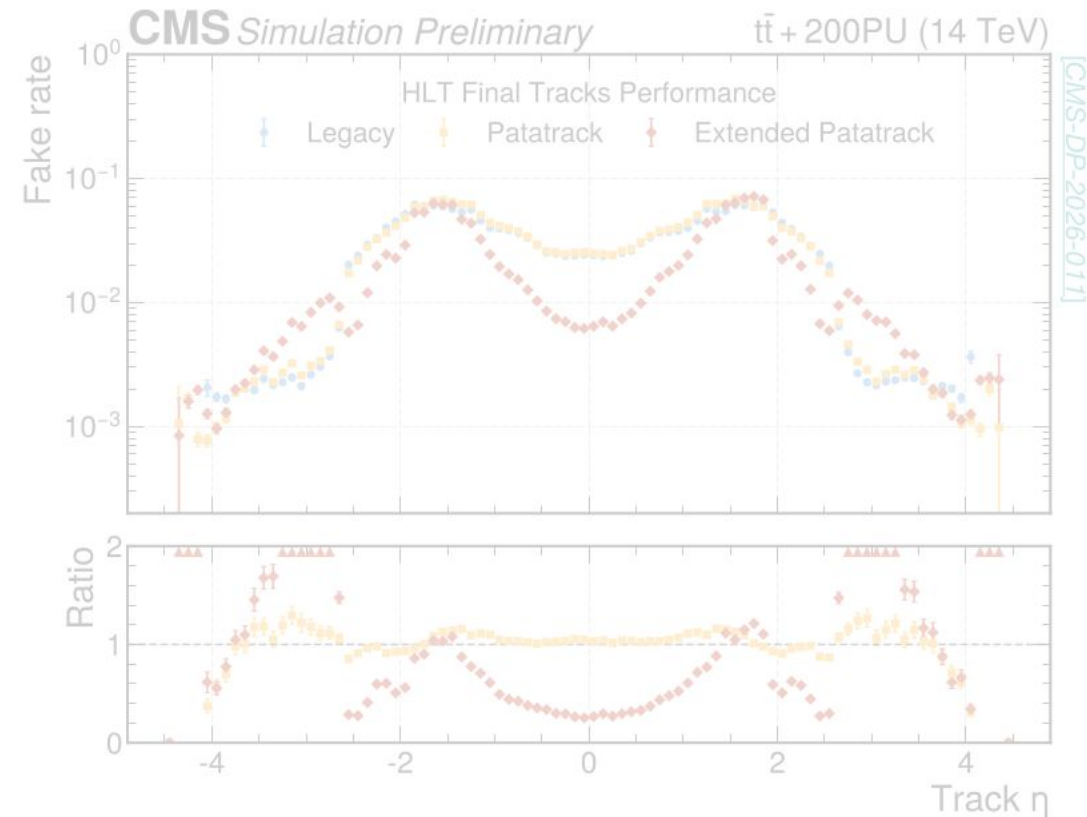
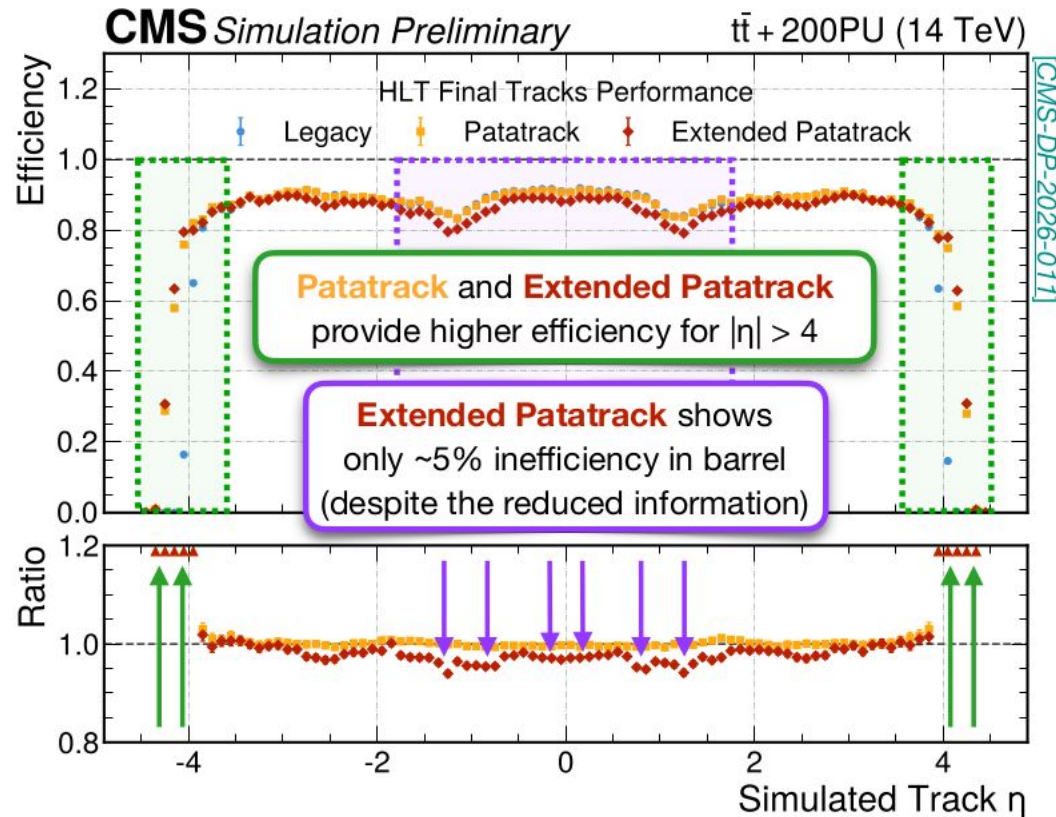
Tracks Performance with **Extended Patatrack**

- The alternative fast **Extended Patatrack** configuration maintains the track physics performance
 - Efficiency**: fraction of simulated tracks matched to at least one reconstructed track
 - Fake rate**: fraction of reconstructed tracks not matched to any simulated track



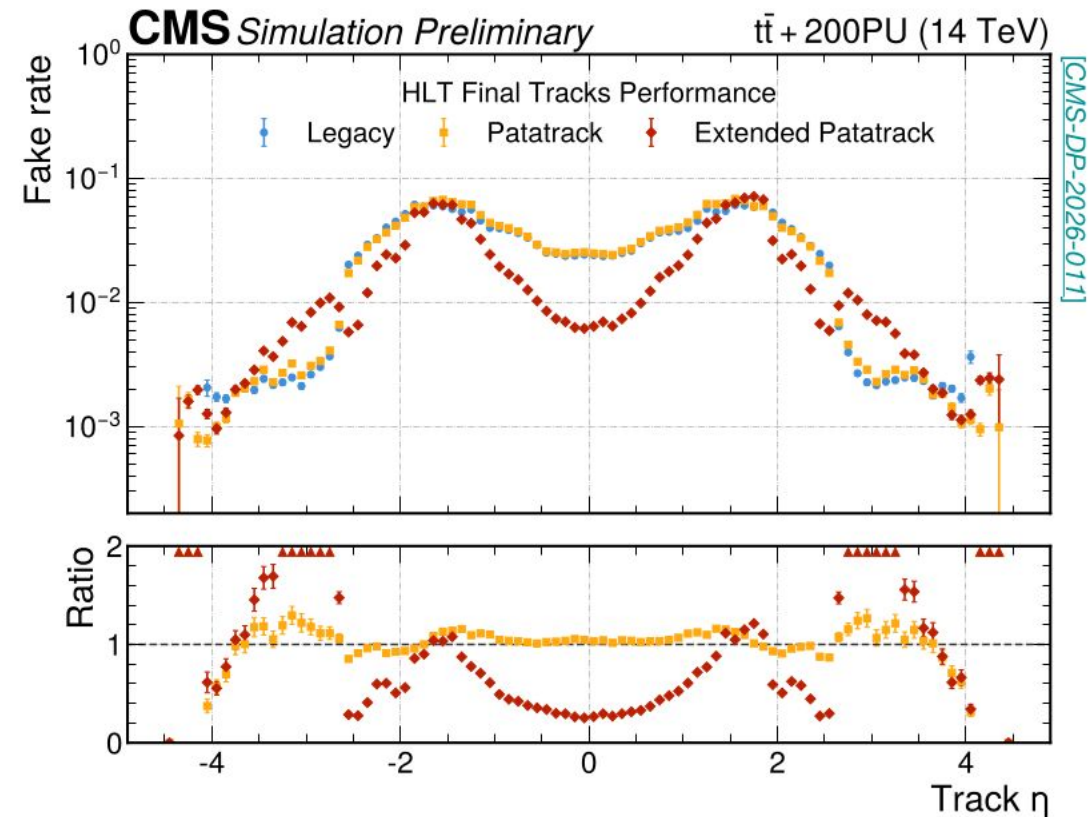
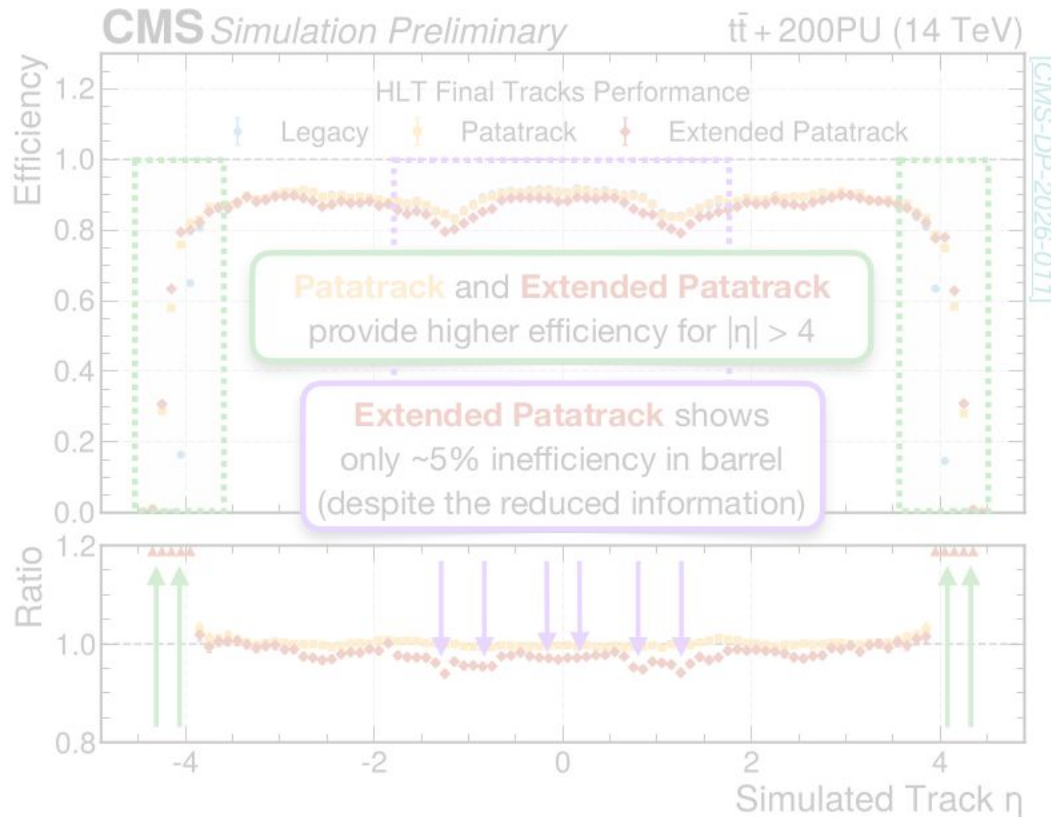
Tracks Performance with Extended Patatrack

- The alternative fast **Extended Patatrack** configuration maintains the track physics performance
 - Efficiency:** fraction of simulated tracks matched to at least one reconstructed track
 - Fake rate:** fraction of reconstructed tracks not matched to any simulated track



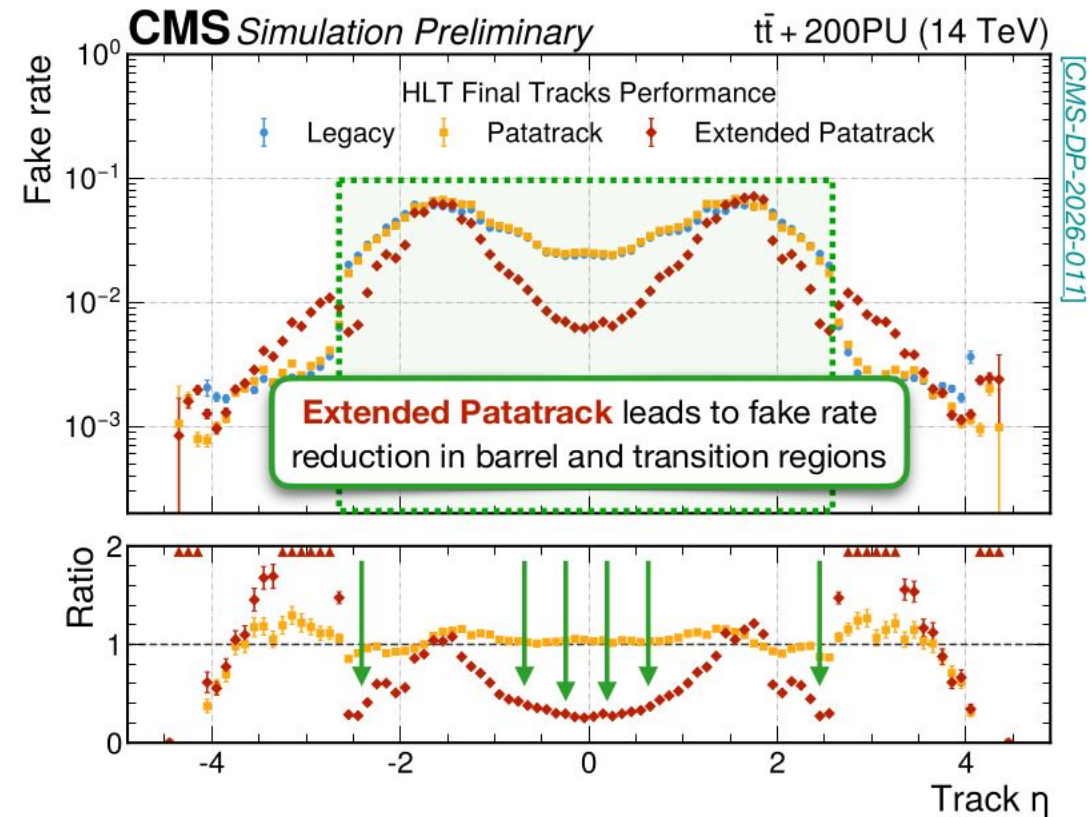
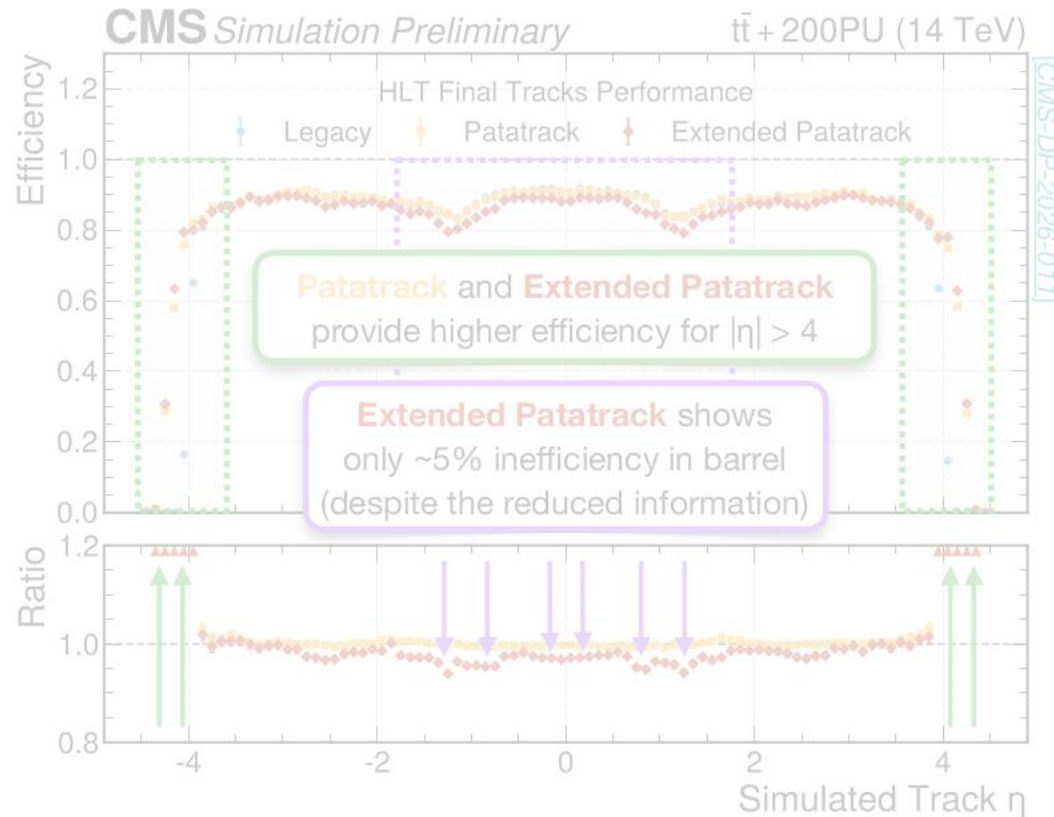
Tracks Performance with **Extended Patatrack**

- The alternative fast **Extended Patatrack** configuration maintains the track physics performance
 - Efficiency:** fraction of simulated tracks matched to at least one reconstructed track
 - Fake rate:** fraction of reconstructed tracks not matched to any simulated track



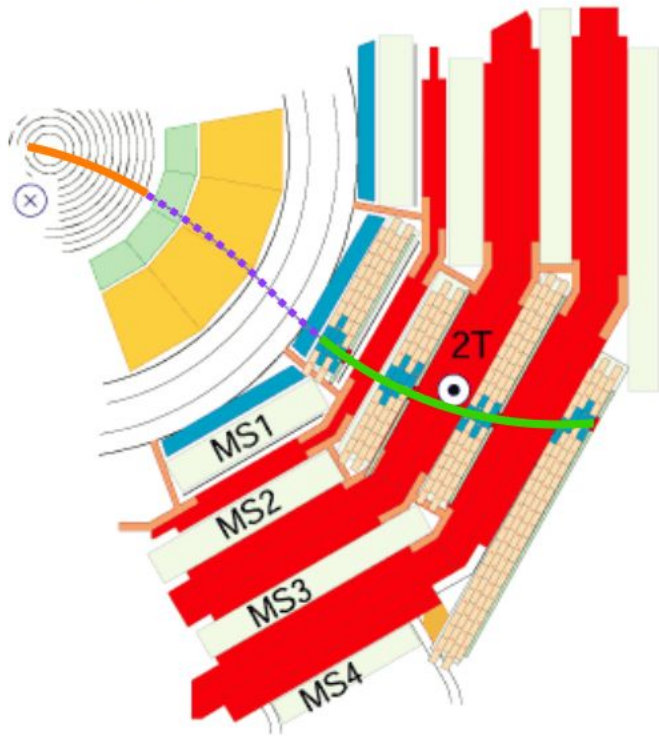
Tracks Performance with **Extended Patatrack**

- The alternative fast **Extended Patatrack** configuration maintains the track physics performance
 - Efficiency:** fraction of simulated tracks matched to at least one reconstructed track
 - Fake rate:** fraction of reconstructed tracks not matched to any simulated track



A Streamlined Muon Tracking

- The legacy HLT muon reconstruction is strongly dependent on tracking performance
 - **Standalone Muons** are built using information from **muon chambers**
 - **Tracker Muons** are built using hits from **Inner/Outer tracker**
 - **Global Muons** link standalone and tracker muons matching trajectories



Tracker Muons

Inside-Out (IO)

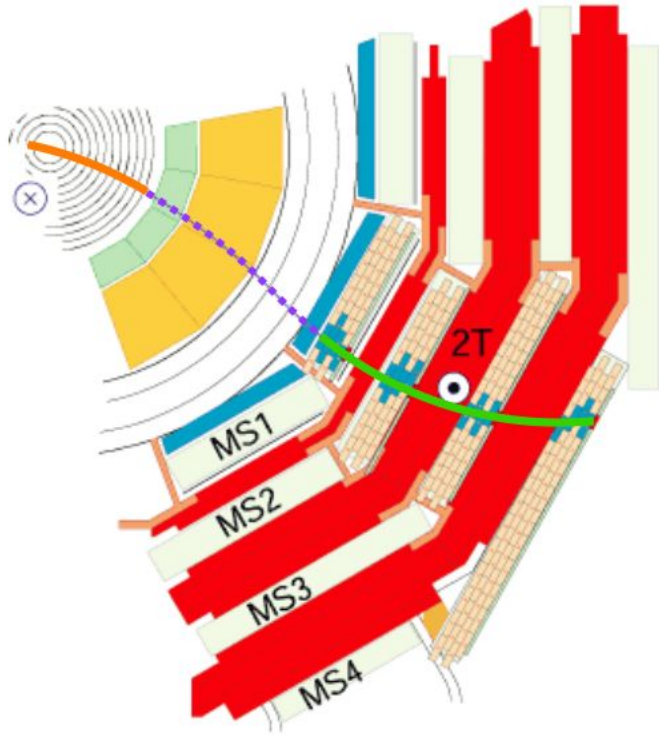
Outside-In (OI)

- ▶ Streamline and accelerate the **Inside-Out (IO)** step:
 - **Legacy** : 2 iterations (seeding + building)
 - **Streamlined** : 1 single iteration (seeding + building), based on Extended Patatrack pixel tracks
- ▶ Processing time for HLT muon reconstruction is reduced

263 ms → 101 ms
60% time reduction !

A Streamlined Muon Tracking

- The legacy HLT muon reconstruction is strongly dependent on tracking performance
 - **Standalone Muons** are built using information from **muon chambers**
 - **Tracker Muons** are built using hits from **Inner/Outer tracker**
 - **Global Muons** link standalone and tracker muons matching trajectories



Tracker Muons

Inside-Out (IO)

Outside-In (OI)

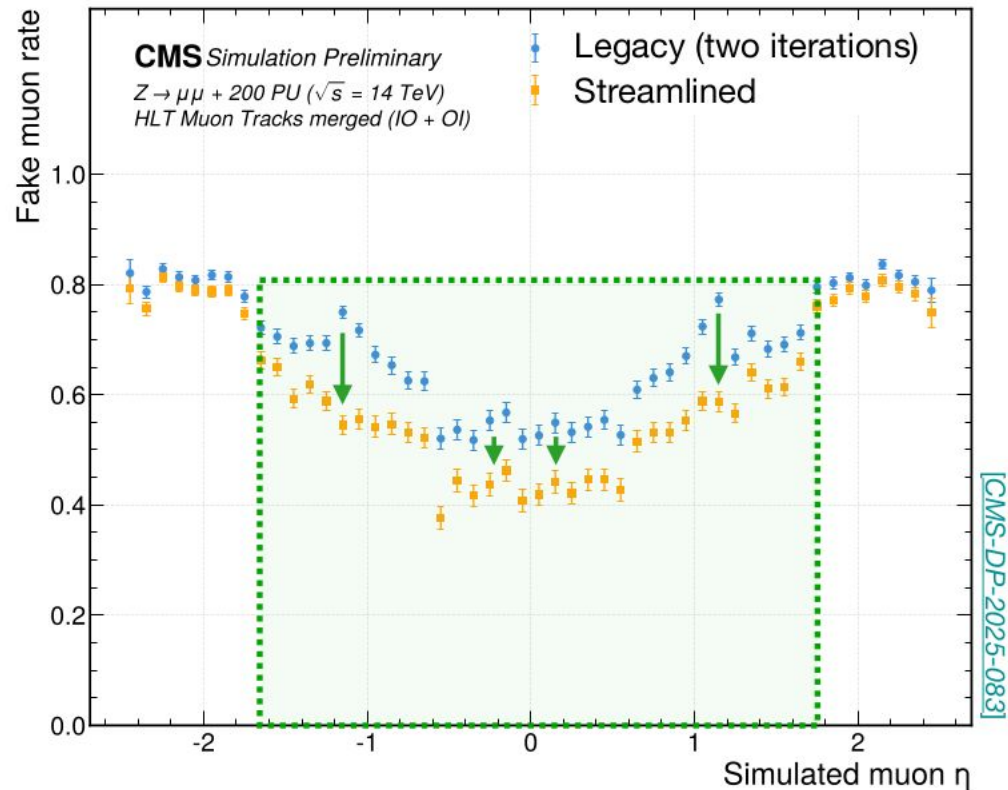
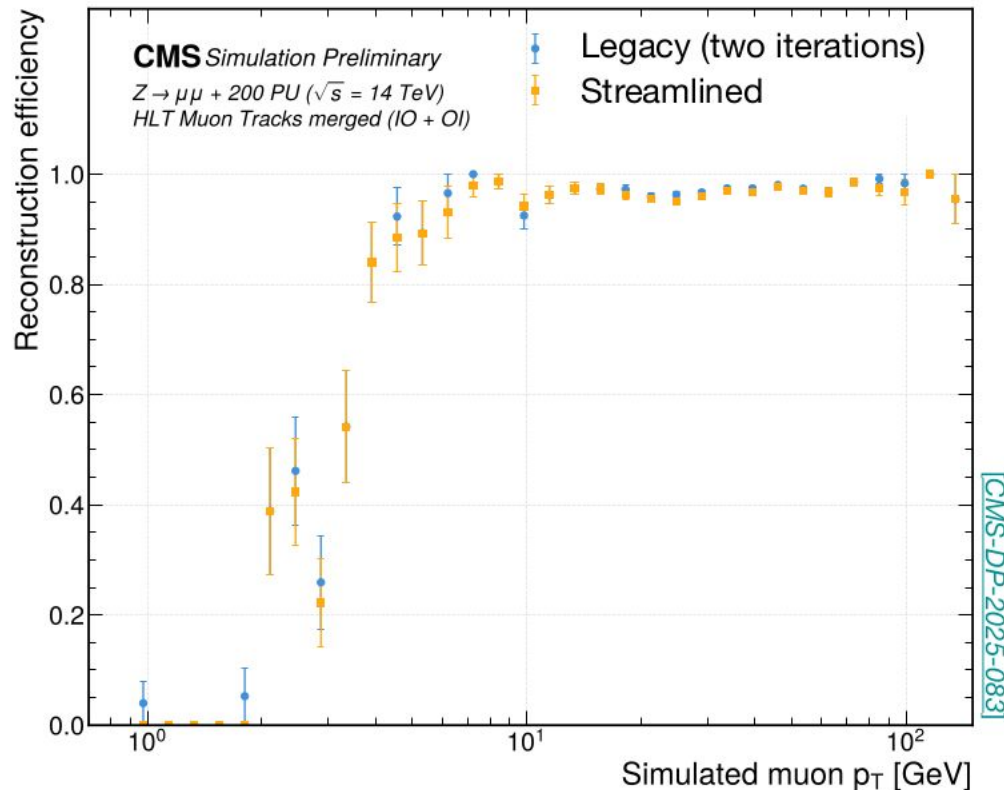
- ▶ Streamline and accelerate the **Inside-Red (IO)** step:
 - **Legacy** : 2 iterations (seeding + building)
 - **Streamlined** : 1 single iteration (seeding + building), based on Extended Patatrack pixel tracks
- ▶ Processing time for HLT muon reconstruction is reduced

263 ms → 101 ms

60% time reduction !

Muon Performance with Streamlined Reconstruction

- Muon physics performance is maintained across the whole p_T and η range:
 - The **Streamlined** algorithm provides **same efficiency** and **lower fake rate** wrt Legacy, while drastically reducing processing time



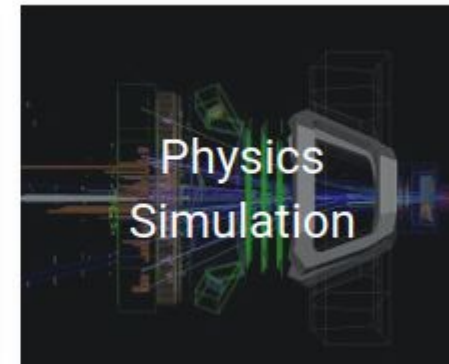
Summary of NGT @ HLT

- The Next-Generation Trigger project at HLT aims to complement the traditional trigger filtering with a full event processing paradigm: **NGT Scouting at full rate**
 - Reconstruct at HLT all events selected by **L1 at 750 kHz** with offline-like quality
 - Store data in a custom lightweight analysis-driven format
- **Important challenges to be addresses:**
 - **Storage**: reduce event size → **data compression**
 - **Calibration**: apply real-time calibrations at HLT → **few hours data buffer**
 - **Computing**:
 - **Optimize resource management** → **flexible computing farm, data-oriented structures**
 - **Redesign expensive algorithms** → **alternative fast algorithms (tracks, muons)**

The NGT project drives the path forward toward addressing these challenges and brings us closer to the full realization of the full rate scouting paradigm

NGT at the Frontier of Innovation

- The HL-LHC challenge demands a fundamental rethinking in the CMS event selection;
- The key to unlocking the full potential of the HL-LHC may lie in the trigger system;
- The Next-Generation Triggers project plays a crucial role in driving this transformation;
 - **NGT @ L1:** ML-driven decisions and 40 MHz scouting
 - extending sensitivity beyond conventional thresholds
 - **NGT @ HLT:** offline quality reconstruction at full L1 accept rate
 - every event accepted becomes available for analysis
- Together, these developments will reshape how observe, select and interpret collision events!





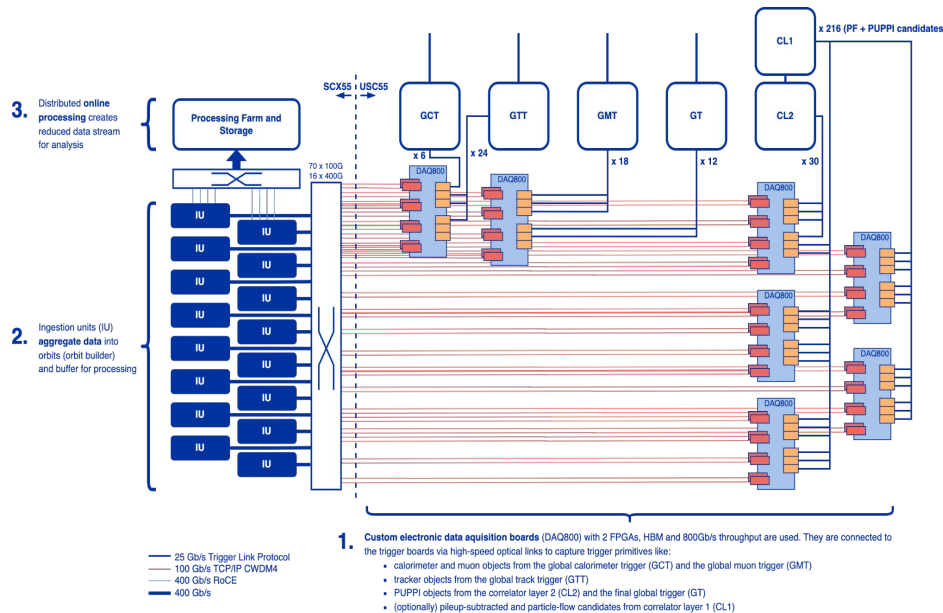
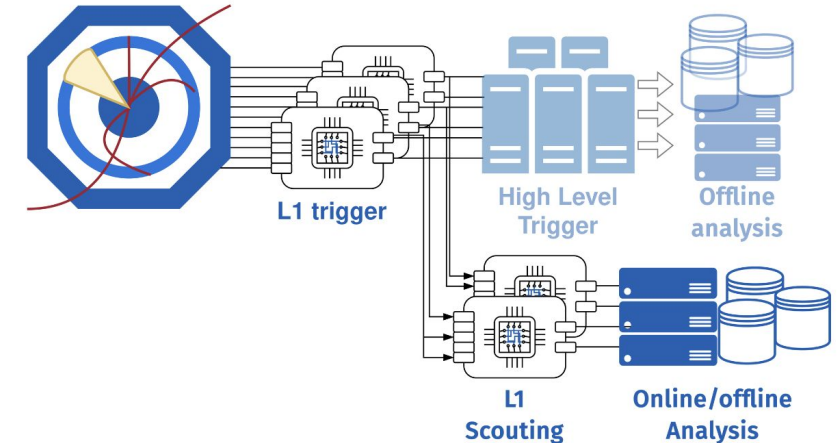
NexTGen
Next Generation Triggers

This work has been funded by the Eric & Wendy Schmidt Fund for Strategic Innovation through the CERN Next Generation Triggers project under grant agreement number SIF-2023-004.

3.5 Enhancing L1T Scouting for HL-LHC: **overview**

Level-1 Data Scouting (L1DS) acquires and analyses **L1T objects** for every event at the full 40 MHz bunch-crossing rate.

- Proof of principle demonstrated during Run 3 data-taking.
 - [DP-2025-056](#), [Talk at CHEP24](#), [Poster at ICHEP24](#), [Talk at TWEPP23](#)
 - First physics results: *Search for heavy long-lived charged particles with level-1 trigger scouting data* [CMS-EXO-25-010](#).



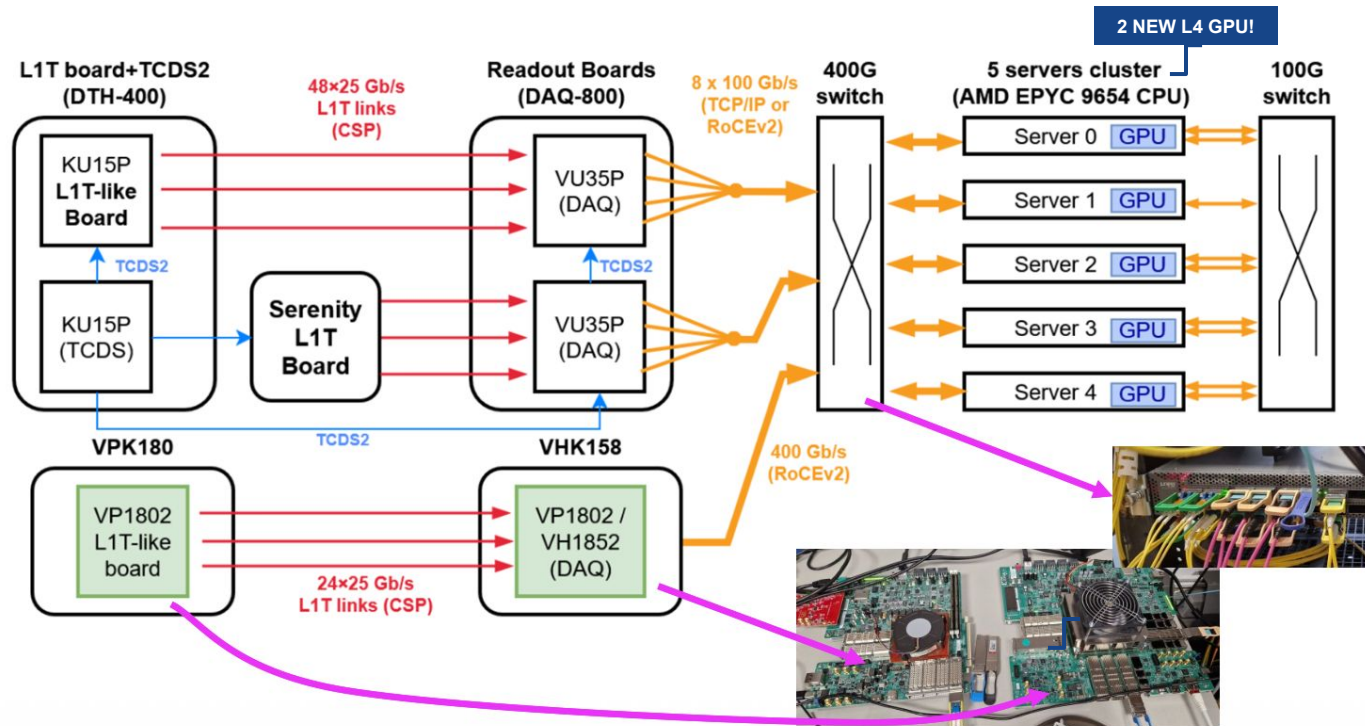
With NGT we explore the physics opportunities and technical feasibility using additional L1 inputs with respect to TDR baseline.

- Explore **increasing complex physics analyses** to enhance sensitivity to all signature that evade the standard L1T → HLT → Offline chain.
- **R&D on the technical feasibility** using different L1 inputs and processing boards with respect to baseline.
 - Investigate running algorithms on CPU / GPU / FPGA / AI engines.
 - Test different approaches for data acquisition (TCP/IP, RoCE, networking).
 - Prototype a data acquisition board to increase links and resources using the same FPGA.

HL-LHC L1DS Scouting baseline: [Emilio's talk at ICHEP](#), [Proceeding](#)

3.5 Enhancing L1T Scouting for HL-LHC: **test system**

Developed a demonstrator for data acquisition and online processing chain (CERN B40, DAQ Lab)



DESIGN OF THE PHASE-2 L1T SCOUTING DEMONSTRATOR

Phase-2 Demonstrator

- System in place: receive from the DAQ800 all the 6+6 TCP streams of PUPPI + e/γ candidates (140 Gb/s total), and run physics analyses at 40 MHz.
- Adopted **Kubernetes for the orchestration** of applications and monitoring.

NextGen Networking

- Upgrade all input data **networking to 400Gb/s**: NICs, switch, optics and fibers connected.
 - **Successful integration and tests of RoCEv2** on two Versal development kits.
- Talk at FPGA 2026 Developer Forum**

NextGen Acquisition Board

- Development of the next generation of data acquisition board (beyond ATCA).
- **Modular Versal ACAP-based system.**
 - Versal Compute Modules + 1x Zynq US+ System Controller. Optical I/O heavy. > 350GB/s aggregate LPDDR4 bandwidth (2U). Linux-enabled System Controller for monitoring & control.

3.5 Enhancing L1T Scouting for HL-LHC: physics

Developing physics analyses to prove the capabilities of the Phase-2 L1 Scouting.

- Identify **suitable signatures** and apply **simple** kinematic and isolation **selections to reduce background rates**, enabling storage of selected events for final analysis.
- First results includes signatures identifiable with PUPPI candidates (Pile-up subtracted PF) and e/ γ objects:
 - $W \rightarrow 3\pi$ (current BR UL 10^{-6} , SM @ $10^{-8/-12}$)
 - $W \rightarrow Ds \gamma \rightarrow KK\pi \gamma$ (current BR UL 10^{-4} , SM @ 10^{-8}), $W \rightarrow \pi \gamma$ (current BR UL 10^{-5} , SM @ $10^{-9/-7}$)
 - $H \rightarrow \rho\rho \rightarrow 4\pi$ ($H \rightarrow \phi\phi \rightarrow 4K$, $H \rightarrow \phi J/\psi \rightarrow KK\mu\mu$), $H \rightarrow \rho\gamma \rightarrow \pi\pi\gamma$ ($H \rightarrow \phi\gamma \rightarrow KK\gamma$, $H \rightarrow J/\psi\gamma \rightarrow \mu\mu\gamma$)

Newer signatures:

- Low mass $X \rightarrow \mu\mu$, Low mass $X \rightarrow ee$
- Rare B meson decays: $B_s \rightarrow \phi\phi \rightarrow K^+K^-K^+K^-$, $B \rightarrow eeK$, $B_s \rightarrow \tau\tau \rightarrow 6\pi$
- Multiple soft jets

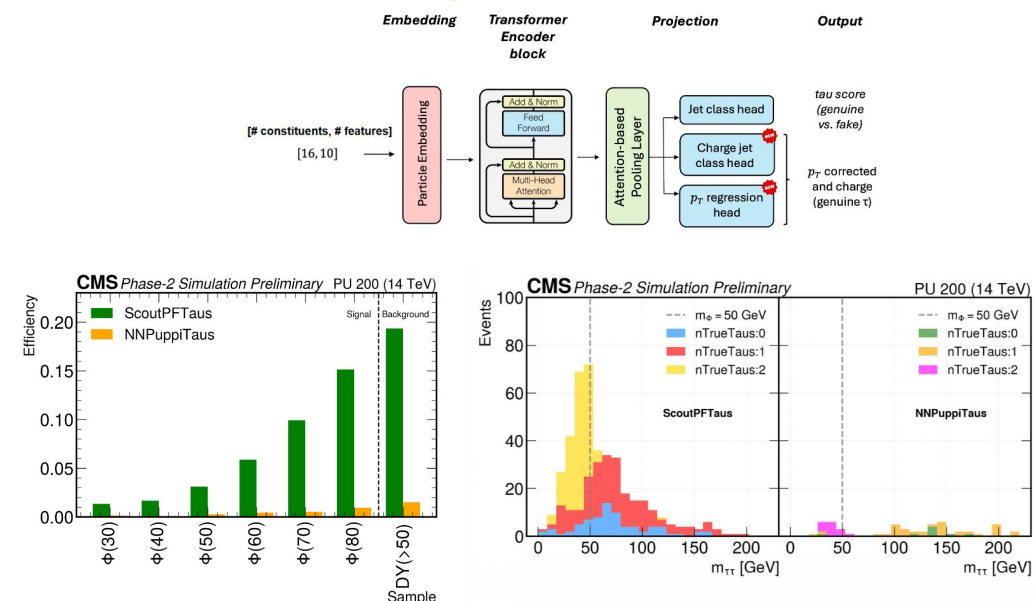
Summary paper on Physics Prospects targeting publication in 2027!

Material: [DP-2025-058](#), [DP-2026-043](#), [Talk at ACAT25](#), [Poster at CHEP24](#)

NEW RESULTS

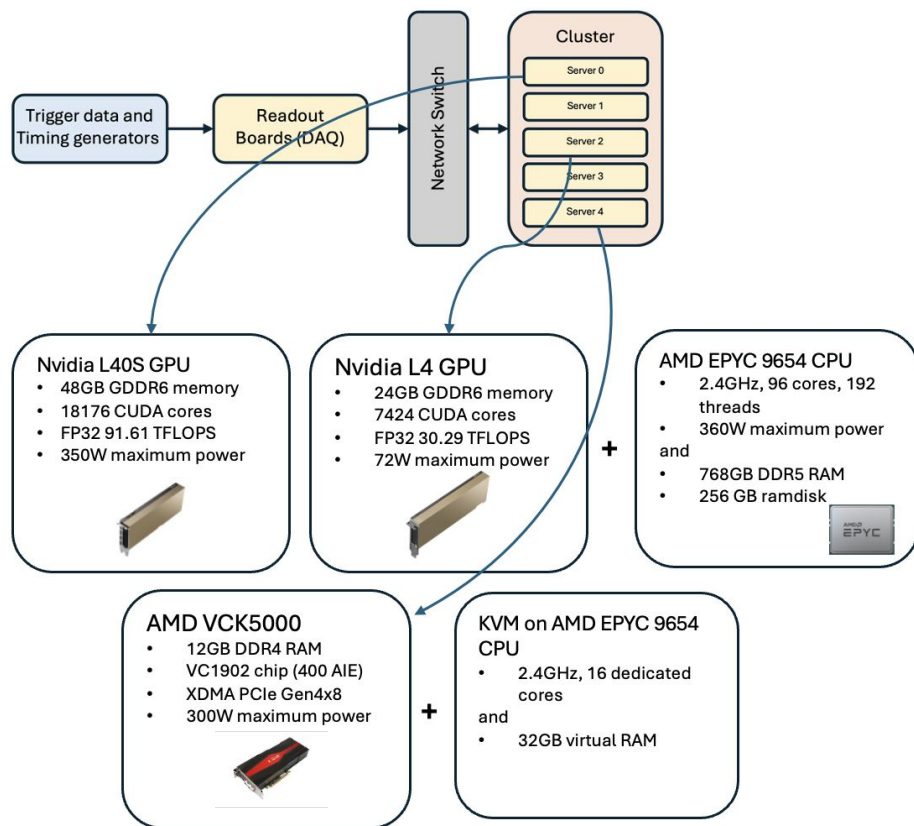
Feasibility study for low-mass ditau resonance

- Transformer model to identify hadronic taus, regress tau p_T and perform charge classification
- Develop event-level strategy for di-tau pair selection in fully hadronic states.
- Benchmark on low-mass $\phi \rightarrow \tau\tau$ resonances.



Increased sensitivity for ScoutPFTaus with respect to NNPuppiTaus baseline.

3.5 Enhancing L1T Scouting for HL-LHC: processing

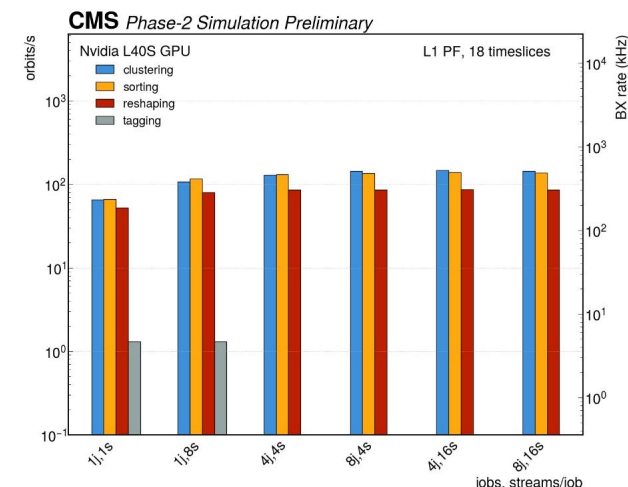
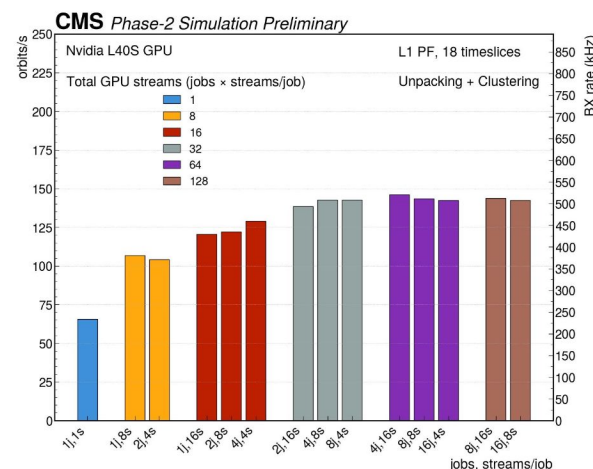


Material: [DP-2025-058](#), Talk at CHEP26.

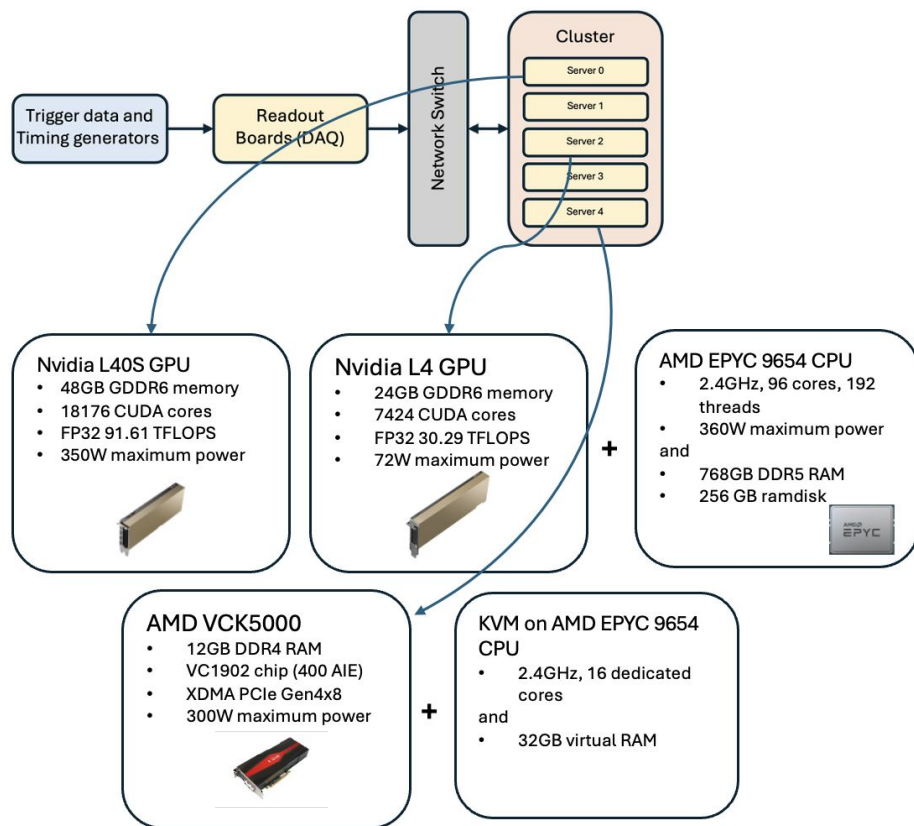
NEW RESULTS

R&D on Accelerators for the Phase-2 L1

- A key part of the Phase-2 L1 Scouting system is the **online processing farm** for data selection, aggregation and storage to disk for offline analysis.
- **Orbit-based processing** → event processing runs on a batch of 3564 events at a time, corresponding to a LHC orbit
- Developed a **full pipeline for the GPU deployment of clustering and ML inference** in CMS software framework for low- p_T tau reconstruction and identification.
 - Code is portable, i.e. is compiled and can run on both CPU and GPU backends
 - Benchmarks have been performed to assess the throughput performance



3.5 Enhancing L1T Scouting for HL-LHC: **processing**

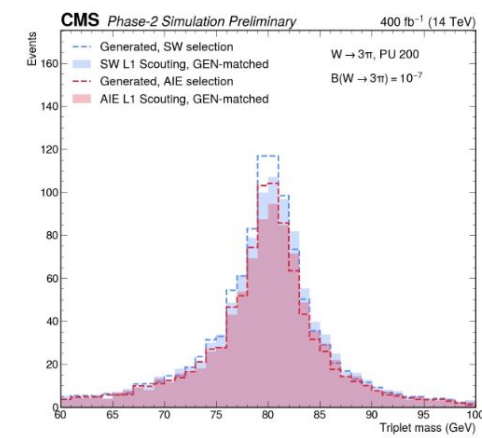
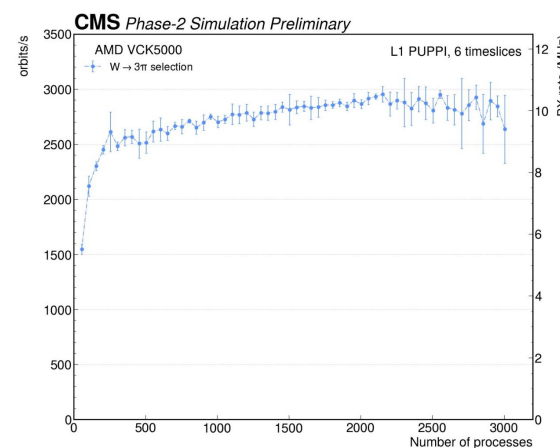


Material: [DP-2025-058](#), Talk at CHEP26

NEW RESULTS

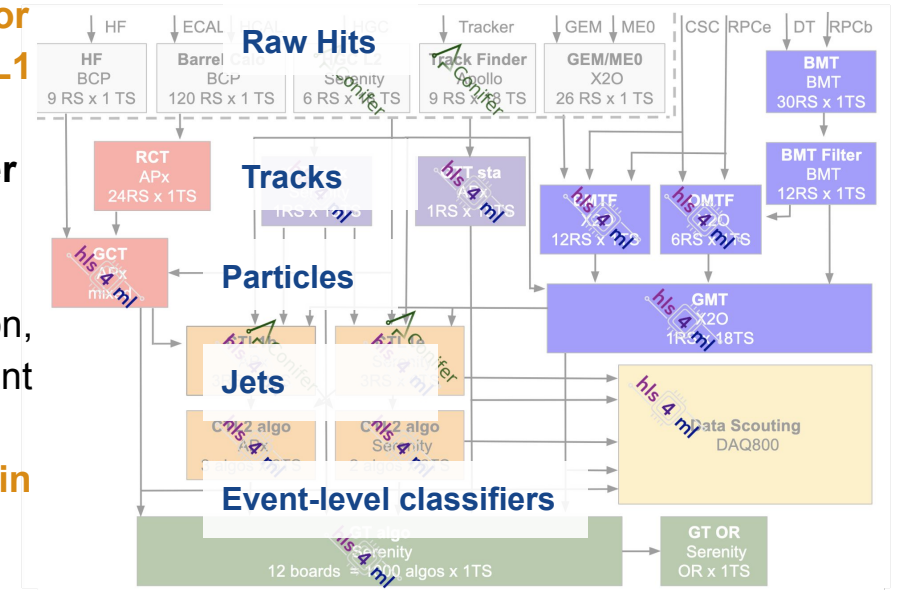
$W \rightarrow 3\pi$ Online Selection on AMD Versal AI Engines

- A key part of the Phase-2 L1 Scouting system is the **online processing farm** for data selection, aggregation and storage to disk for offline analysis.
- Rare $W \rightarrow 3\pi$ decay is identified using a **cut-based selection** that aims to reconstruct a triplet of charged particles from the L1 PF reconstruction and PUPPI pileup suppression.
- Benchmark a full AIE-based design on a AMD VCK5000 Versal Development Card.**
 - ~10 MHz of peak event throughput is obtained by running ~2000 concurrent processes on the host.
 - Efficiency of 11.1% comparable to CPU 12.6%.



3.6 Practical Real Time AI for L1T/Scouting: goals

- Boost the development and integration of advanced and fast **ML algorithms for improving the HL-LHC L1T reconstruction (used by standard triggers, and L1 Scouting) thanks to NGT**
 - Into FPGAs of **Global Trigger, Correlator and Global Track Trigger subsystems**
- Develop **setup for complex trainings of multiple algorithms**
 - e.g. simultaneously optimize algorithms for particle ID, object reconstruction, and event selection (implemented in different subsystems, only limited amount of information is propagated between subsystems)
- (Joint with 3.7) **Develop practices for training & deployment of ML algorithms in L1T (MLOps)**
 - Updating & redeploying algorithms for changes in detector conditions
 - tracking, archiving & retrieving ML models (e.g. for consistent emulation)
- Fundamental the interplay with **WP1.2, WP1.3 for hls4ml tools support & WP1.1 for the heterogeneous platform for training/monitoring/synthesis/deployment.**



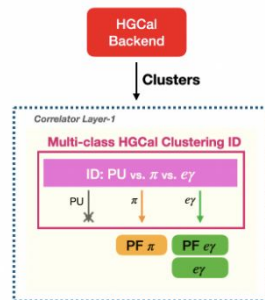
25 billion inference/second estimated from known projects in development

3.6 Practical Real Time AI for L1T/Scouting: algorithms

DONE

HGCAL Clusters:

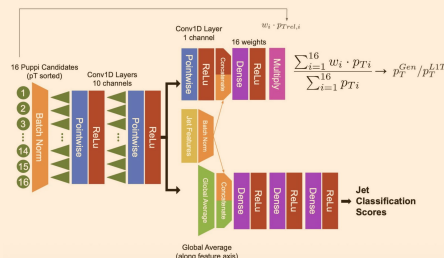
Multiclass discriminator to improve and harmonize IDs for e/γ, pions, pile-up separation for PF reconstruction: **increase efficiency at low p_T for e/γ clusters.**



Material: [DP-2024-098](#), [Talk at CHEP24](#)

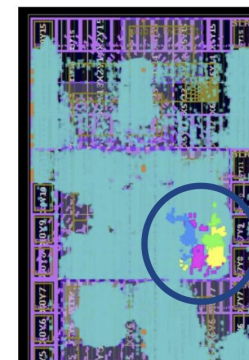
NEW RESULTS

Jet tagging: DeepSet based architecture trained for object tagging and p_T regression.



Material: [DP-2026-037](#), [DP-2025-032](#), [Talk at FastML25](#), [Talk at CHEP26](#).

NEW RESULTS



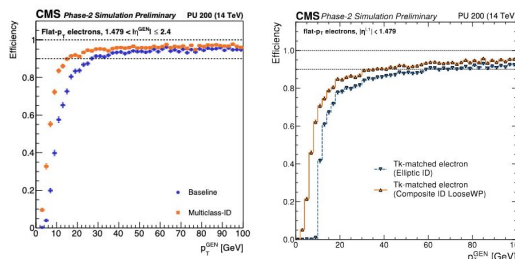
Track-to-Vertex association: NN based improve vertex finding and track association to reduce pileup impact.

Correlator Trigger firmware and emulator now implemented, and model has been run in hardware.

Material: [DP-2021-035](#), [DP-2022-020](#)

DONE

Tracker-Tracks+Clusters BDT instead of geometrical track-matching, **increase efficiency and better p_T resolution for electrons**



Material: [DP-2024-098](#), [DP-2025-059](#), [Talk at CHEP24](#), [Poster at FastML25](#)

ONGOING

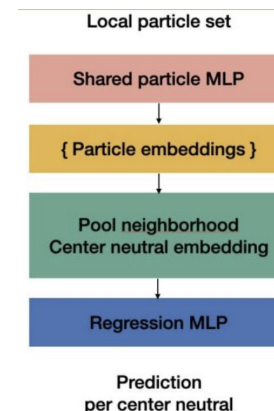
Topological Triggers: NN using all high-level objects in Global Trigger or all PF/Puppi Candidates in Correlator to enhance efficiency for VBF $H \rightarrow b\bar{b}$, VBF $H \rightarrow \text{invisible}$, $HH \rightarrow 4b$, $HH \rightarrow b\bar{b}\tau\tau$... reach signatures.

Next step: focus on inputs rather than model optimization to recover VBF jets.

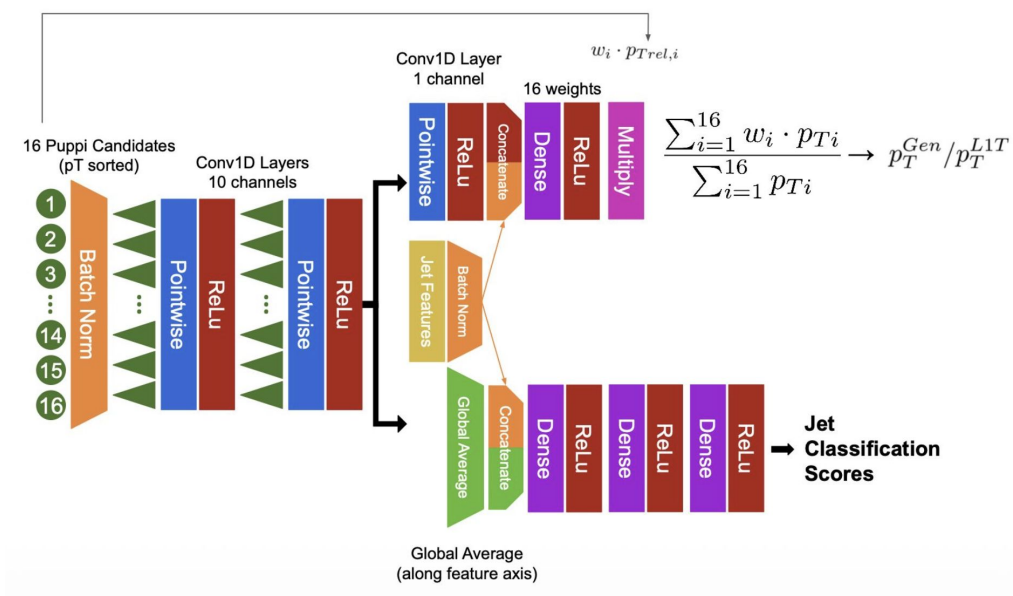
ONGOING

ML PUPPI:

DeepSet-based regression model to improve neutral PF candidate weighting using local particle information, targeting improved L1T jet and MET performance.



3.6 Practical Real Time AI for L1T/Scouting: algorithms

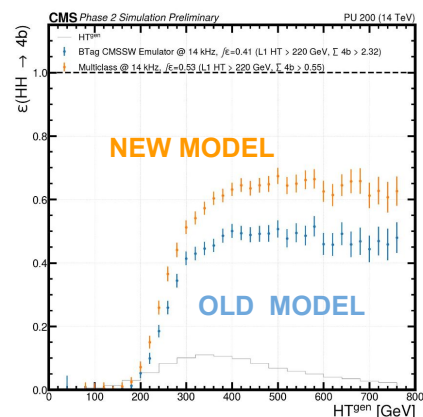


Material: [Regression \(DP-2026-037\)](#), [Tagging \(DP-2025-032\)](#), Talk at [FastML25](#), Talk at [CHEP26](#).

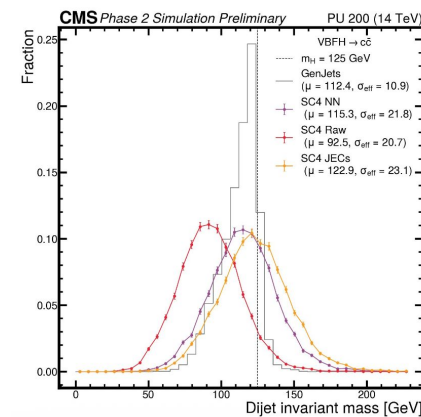
NEW RESULTS

Jet Tagging Strategy for Phase-2

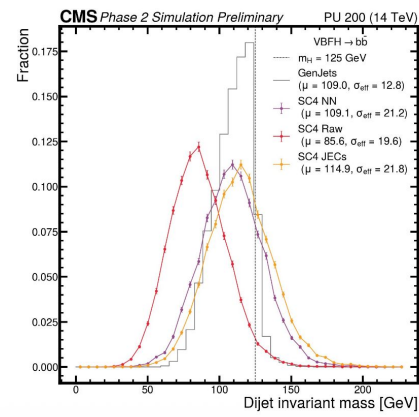
- PUPPI Jet Tagging:** introduce multiclass tagging trained on a variety of target classes: **b jets, c jets, τ_h (positively charged), τ_h (negatively charged), gluon jets, muon, electron, and light quark (uds) jets, plus new pileup category.**
 - The algorithm provides improvement on triggers for **HH $\rightarrow$ bb $\tau\tau$, HH $\rightarrow$ 4b, VBF H $\rightarrow$ $\tau\tau$**
 - Fully integrated into **CMSSW**, and its **FPGA firmware implementation** fits in the available resources and has been cross-checked to ensure bit-level accuracy.
- NEW p_T regression:** novel NN branch to derive the jet p_T correction factor to improve the baseline jet energy correction (JECs).



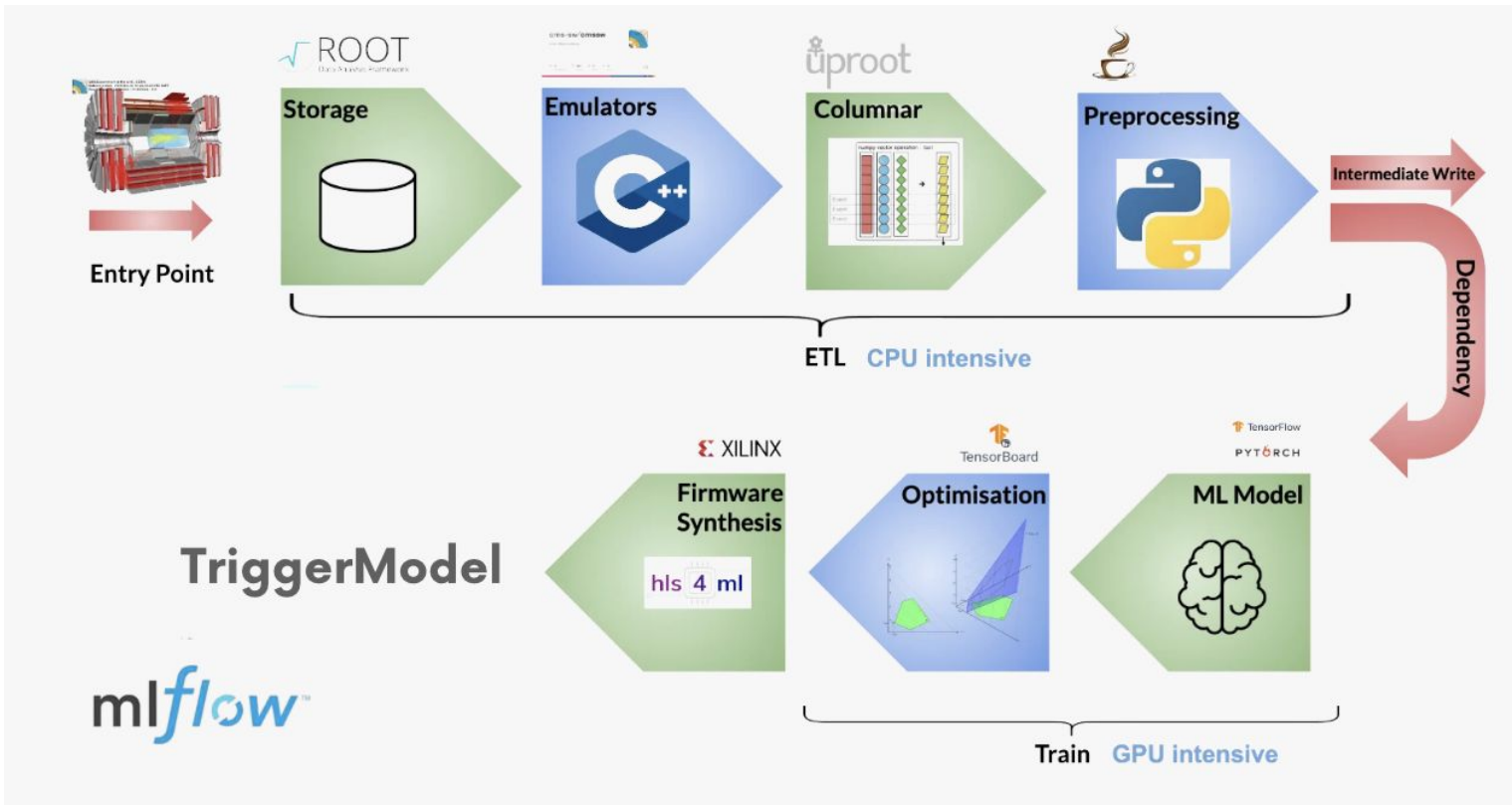
Increased HH $\rightarrow$ 4b efficiency.



Improved peak position relative to **genJets** and reduced peak width for **SC4 NN jets**.



3.6 Practical Real Time AI for L1T/Scouting: **MLOps**



How to maintain ML in L1T ?

Dedicated work on ML workflows for model development and deployment: use of Gitlab CI pipelines, MLFlow.

Release public package: [triggerflow](#) targeting ML developments on FPGA.

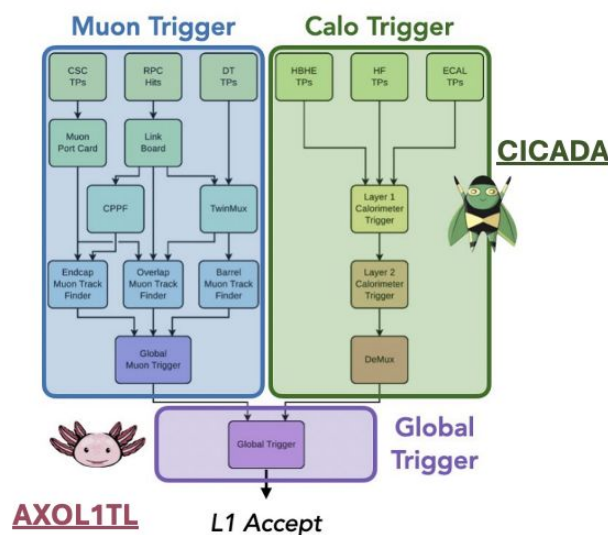
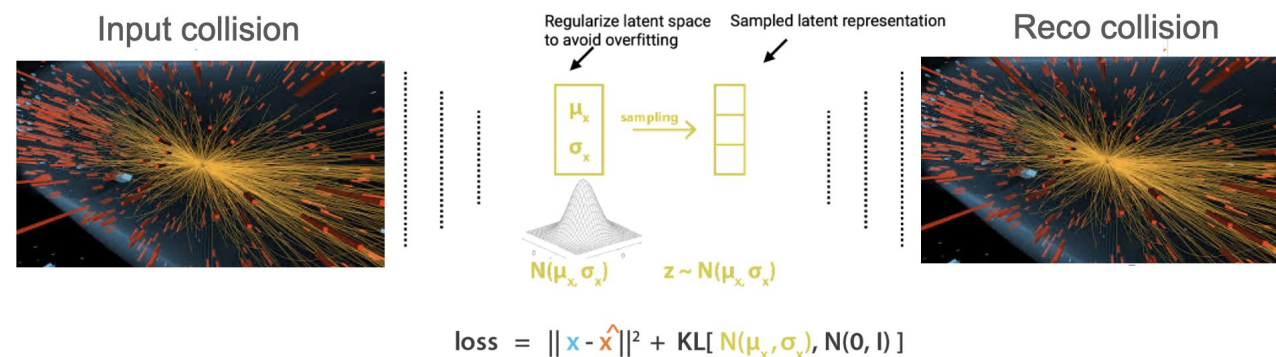
Working on training all the 3.6 ML algorithms together.

Exploit NGT WP1.1 infrastructure for resources.

Material: [FastML25 topical](#), [Talk at ACAT25](#).

3.7 L1T anomaly detection and data compression

- Redefine the way events are selected with anomaly detection in a **model-independent way** : trigger on “**anomalous**” events using ML unsupervised models (**sub- μ s autoencoders!**)



- In Spring 2024 CMS deployed a new trigger called AXOL1TL in L1 Global Trigger
- AXOL1TL was designed and integrated over last ~3 years by CMS Collaborators.
 - Collected over 100 fb^{-1} of data with a L1T pure rate of up to 250 Hz, passthrough at HLT Scouting.

NGT contributed to the development of AXOL1TL and it's driving the AD developments for the HL-LHC together with novel and advanced ML solutions for the L1T system, e.g. targeting data compression.

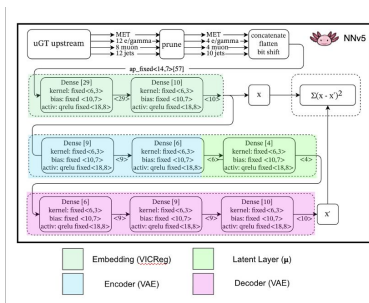
Material: [DP-2024-059](#), [DP-2025-061](#), [DP-2024-121](#), [Keynote talk at ICHEP24](#)

3.7 L1T anomaly detection and data compression

DONE

AXOL1TL V5

Deployed new model with contrastive block to enhance feature extraction and improve performance for 2025 data-taking.

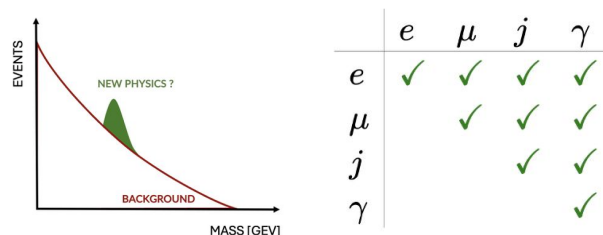


Material: DP-2025-061, Talk at FastML25, Poster at ACAT25

ONGOING

Physics Analysis

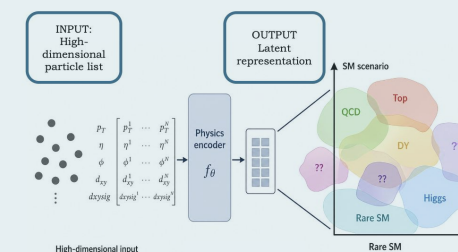
Model independent bump hunt search for new physics with AXOL1TL triggered data.



ONGOING

Foundation Model for Phase-2 L1T

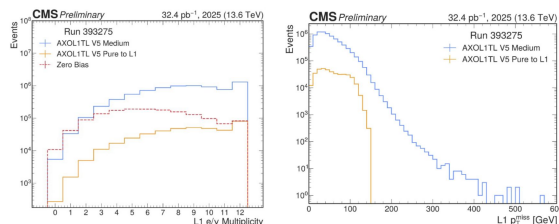
Developing an upgraded AD model, extendable to other tasks, incorporating new inputs and improved capabilities.



DONE

AXOL1TL trigger studies

What is anomalous? Extensive trigger characterization: AXOL1TL provides enhanced sensitivity to high multiplicity and soft objects!

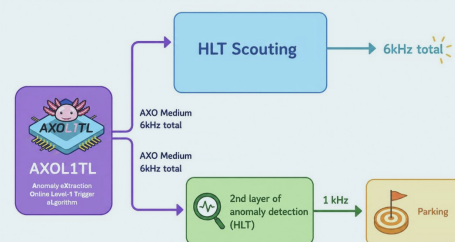


Material: [DP-2025-061](#), [Talk at FastML25](#)

NEW RESULTS

2nd Layer of Anomaly Detection

Enhancing the purity of anomalous events in the AD Parking stream.

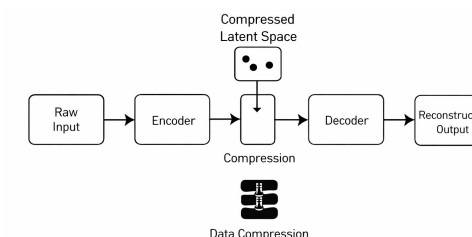


Material: [DP-2026-046](#), [Talk at CHEP26](#)

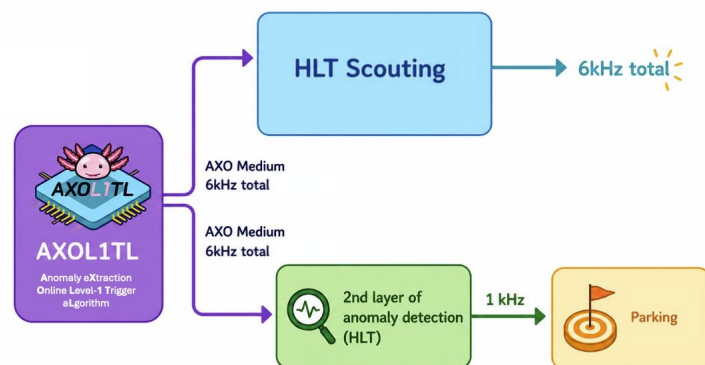
ONGOING

Data Compression for L1 Scouting

Use ML approaches, like tokenization, to obtain a compressed representation of data for downstream physics.

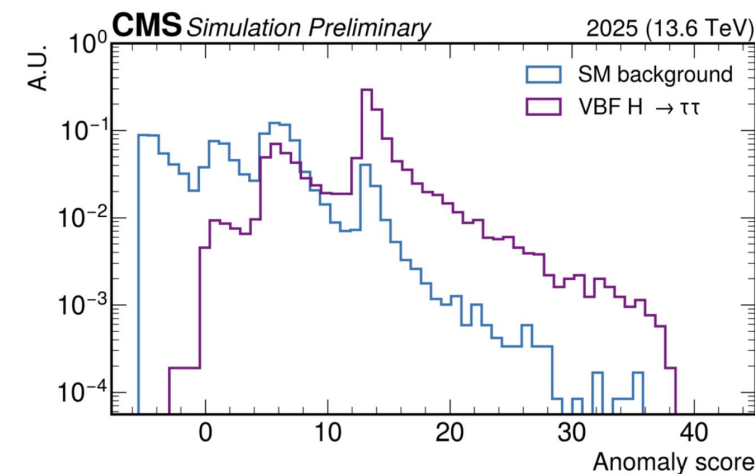
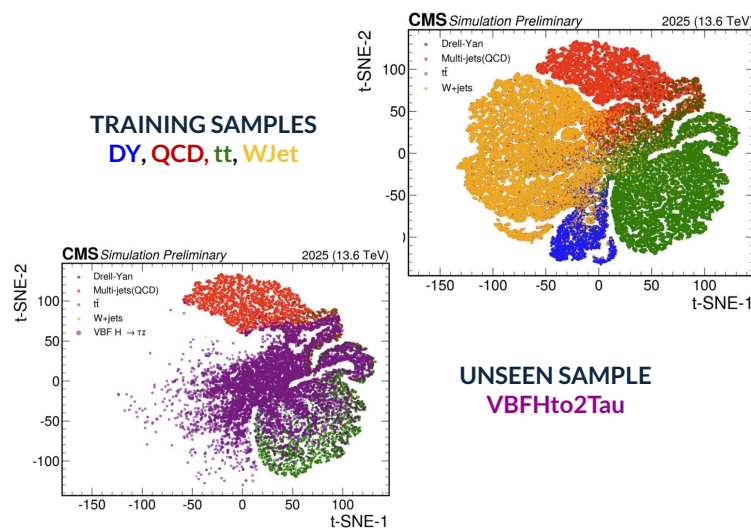
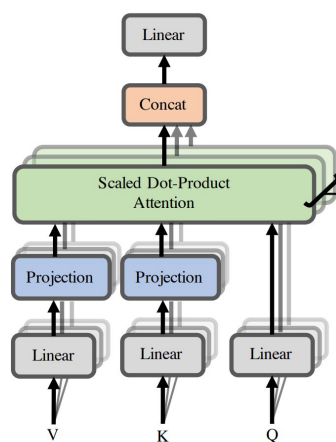


3.7 L1T anomaly detection and data compression



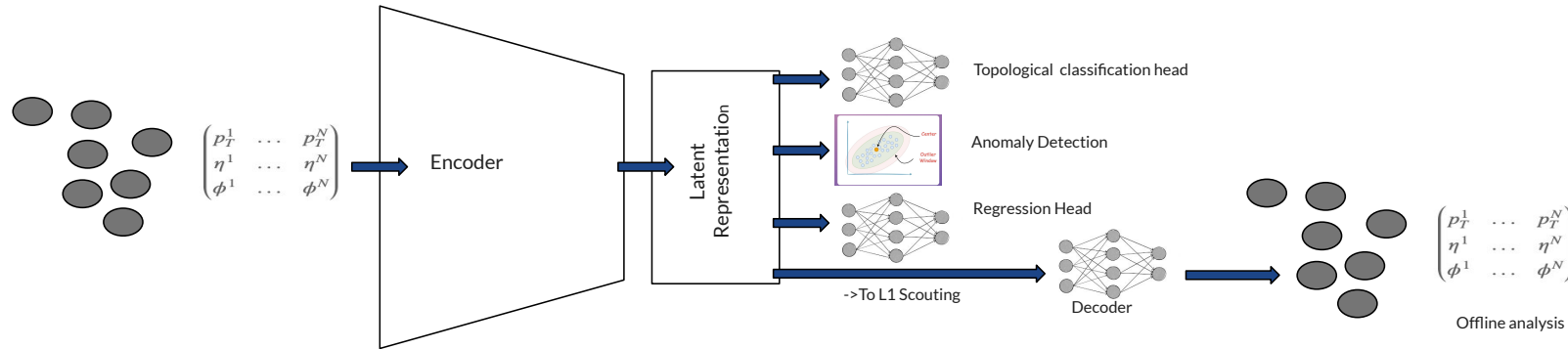
R&D on 2nd layer of anomaly detection

- Enhance the purity of **AXOL1TL** anomalous events in the AD parking stream.
- Linformer-based architecture to construct a **physics-informed embedding**.
 - HLT PF candidates kinematic features (p_T , η , ϕ , d_{xy} , particle ID).
 - More freedom at the HLT as model runs on CPU/GPU.
- HLT AD model serves as a **proxy for the Phase-2 L1T AD**.



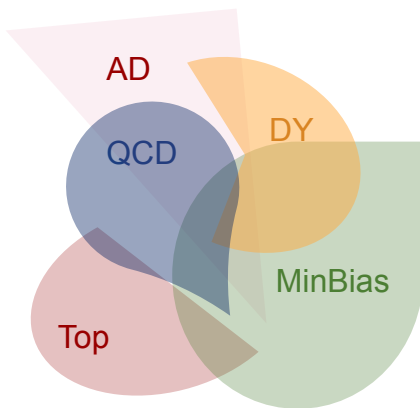
Material: [DP-2026-046](#), [Talk at CHEP26](#)

3.7 L1T anomaly detection and data compression



Foundation Model for the L1T Correlator Layer 2

- Transformer-based model development for the L1T Correlator Layer 2 to build a physics-informed latent space.
- **Can we build one common embedding from which we develop several lightweight NN/BDT heads for various ML/AD task?**



So far

- Train on PF PUPPI candidates kinematic features (p_T , η , ϕ , d_{xy} , particle ID).
- First version of the model with **ongoing studies on physics performances for AD**.
- **Firmware meets** the timing/throughput **constraints** of CMS L1T.

Challenges

- Transformer model HLS representations does **not match** the software and firmware (VHDL) one.
- **Synergy with NGT WP1 work: looking for help and inputs on to debug this!**

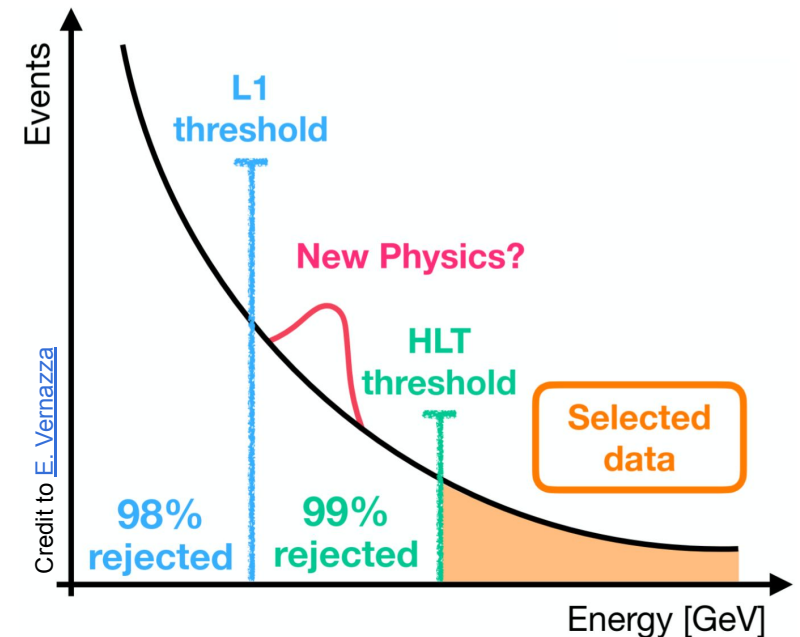
NGT HLT

Goal: Expand the CMS data scouting strategy by implementing the **NGT scouting**, i.e., fully reconstruct all L1-accepted events @ HLT, providing offline-like data at 750 kHz!

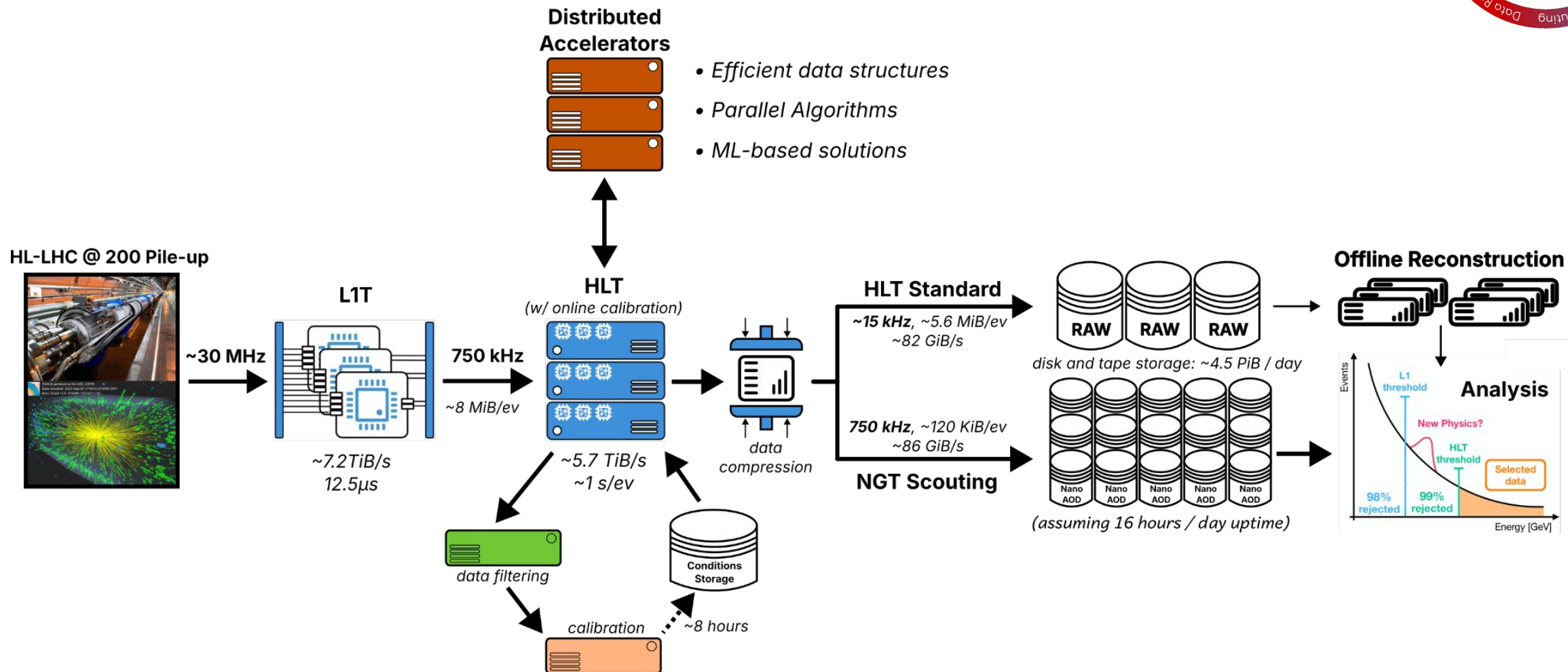
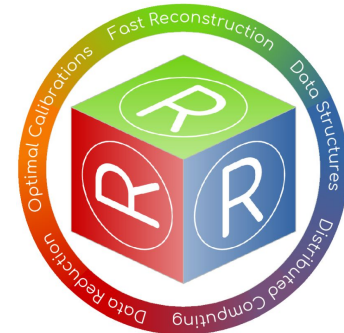
- Requires the design and implementation of efficient and heterogeneous algorithms
- High-quality real-time calibrations have to be derived and applied directly at HLT
- Save data in custom lightweight analysis-driven format to afford storage
- Provide offline-like objects for physics analysis

In parallel: Provide CMS with a faster, lighter, and physics-improved event reconstruction

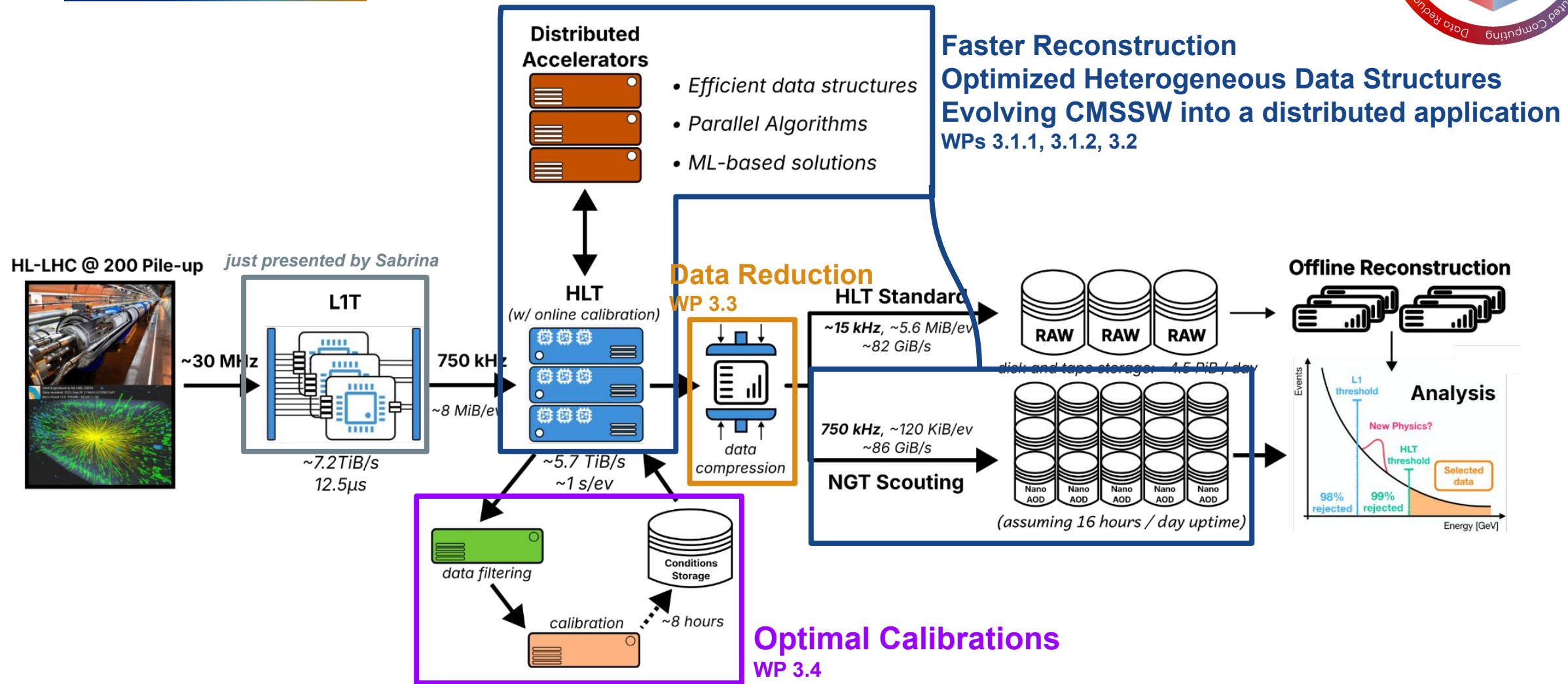
- NGT HLT work and ideas can be integrated and further developed into the standard CMS HLT strategy
- Benefits can percolate to the offline reconstruction
- Expertise will be created and benefit CMS at large



A bird's-eye view of NGT HLT



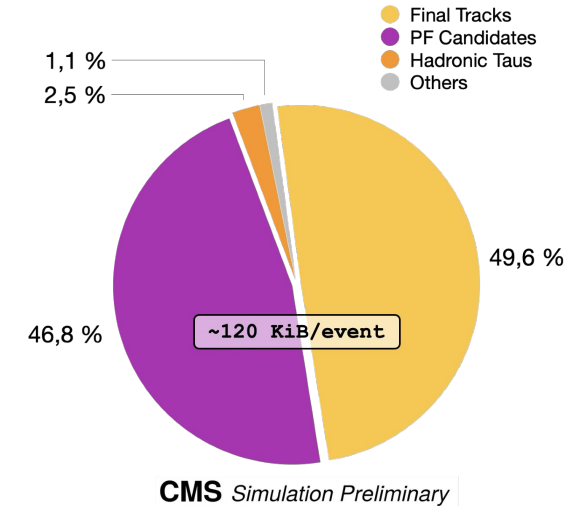
A bird's-eye view of NGT HLT



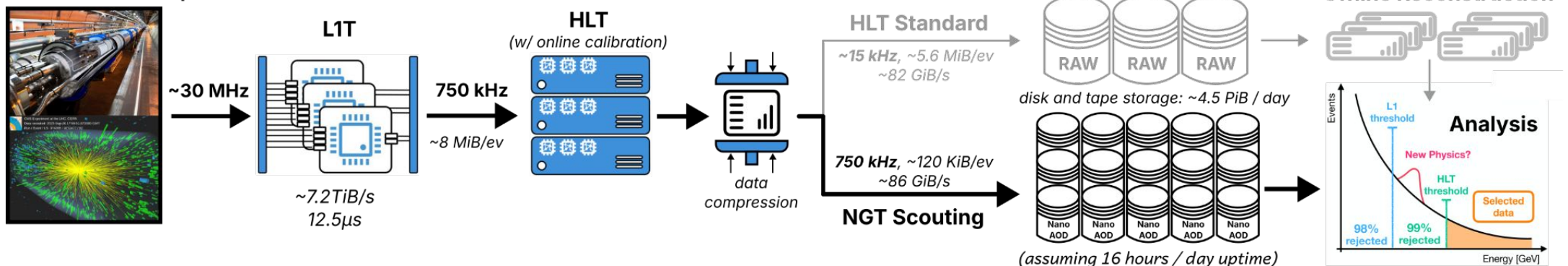
NGT Scouting stream and output format

- **Setup a dedicated NGT Scouting path @ 750kHz**
 - Runs the full reconstruction on the L1-accept input rate
 - More challenging than the baseline HLT Scouting @ 225 kHz (~30% L1T rate)
 - Testing ground for all kinds of experimental reconstruction strategies, aiming at reducing the timing while keeping or improving the physics
- **Dedicated Phase-2 NGT HLT output format, including a compressed version of all relevant objects**
 - Also useful for a-posteriori validation and as input for ML training
 - O(120) KiB/event without selections applied, but with duplicated info

Event Content

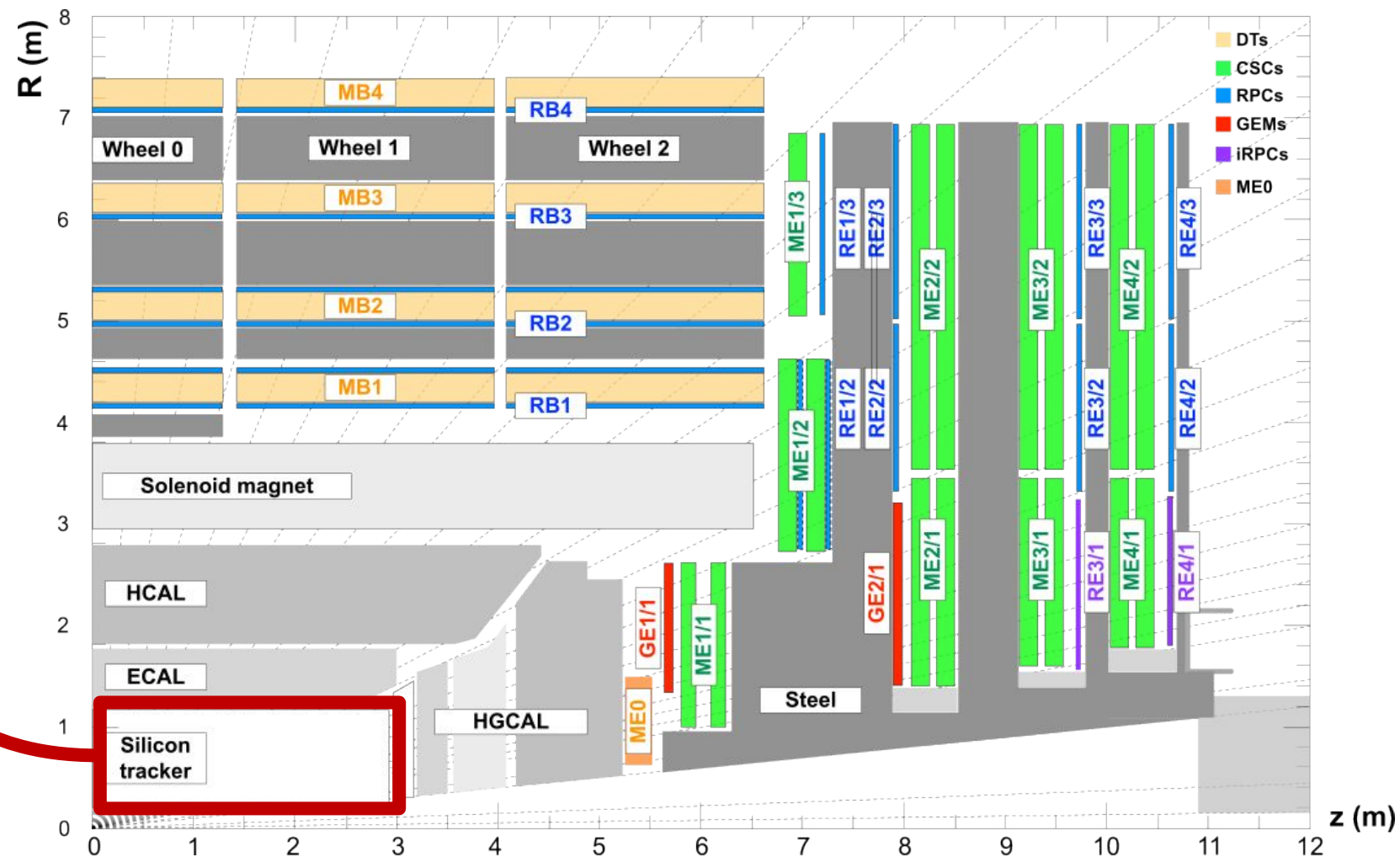


HL-LHC @ 200 Pile-up



Faster algorithms... where?

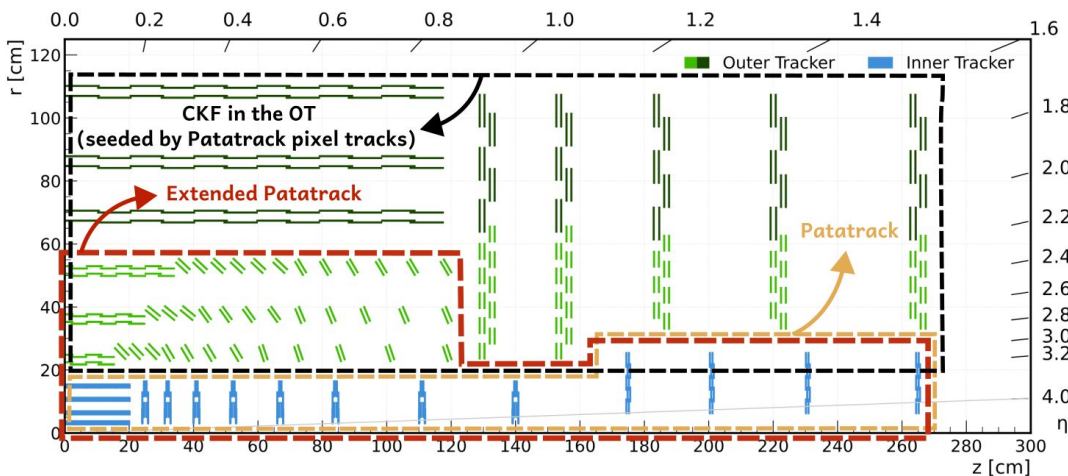
Tracker



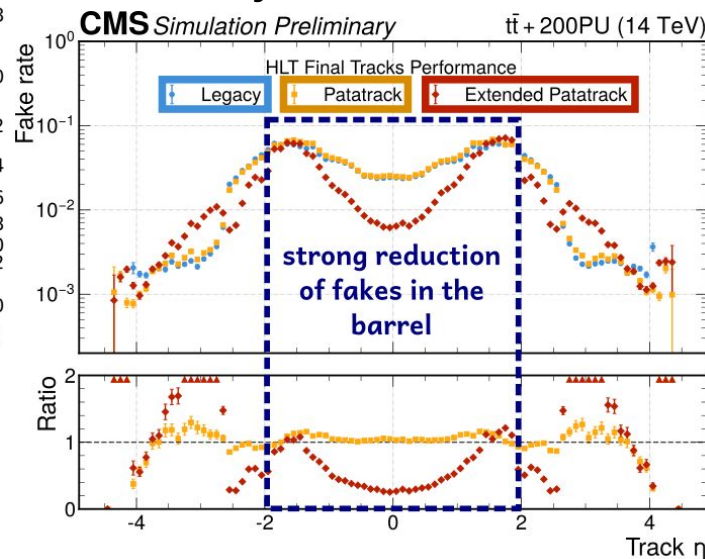
Pixel Tracking: Extension and Acceleration

- ~34% speed-up with new NGT “**Extended Patatrack**” pixel tracks
 - large fake rate reduction
 - 5% inefficiency “only”, despite not using the endcap outer tracker and the silicon strip layers in the barrel

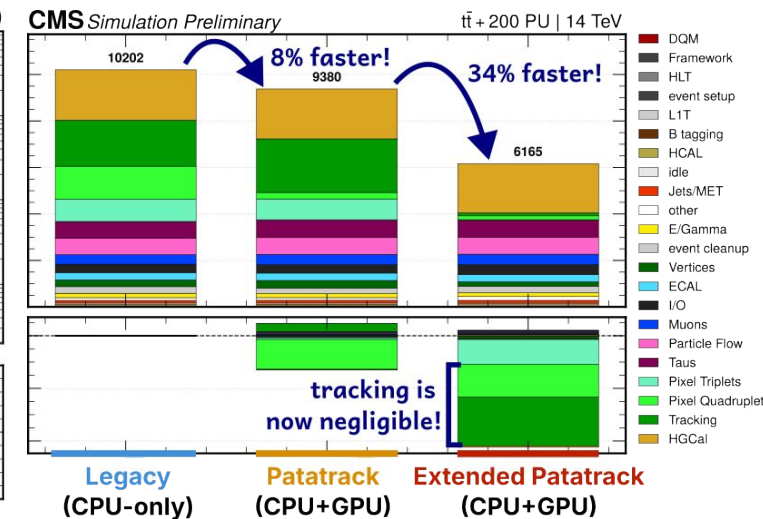
The regions of influence of tracking algorithms in the CMS Phase-2 Tracker



Physics Performance



Timing Measurements



CERN-CMS-DP-2026-011



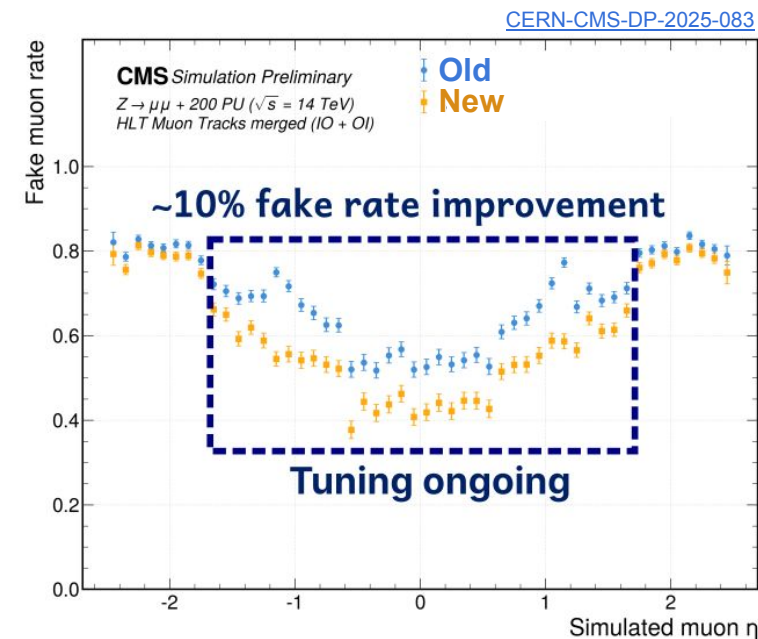
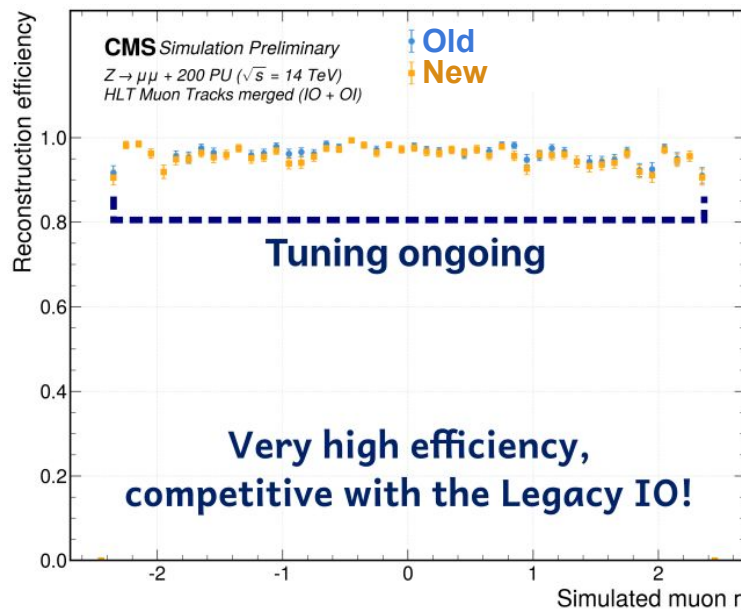
Muon Tracking: Rethink and Optimize

Old:

- The HLT muon reconstruction uses muon-only seeds and L1T seeds
- Muons seeded by L1T are reconstructed with two steps for all the tracks:
 - Inside-Out (IO, from the inner to the outer tracker)
 - Outside-In (OI, vice-versa)

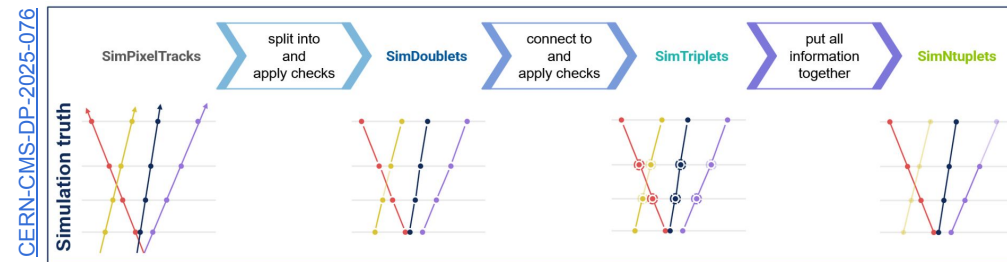
New:

- The use of tracking already @ L1 enabled the simplification of the HLT seeding strategy
- The L1 tracker-seeded IO + OI chain was overhauled and simplified, **using Extended Patatrack**, and achieving **76% speed-up!**



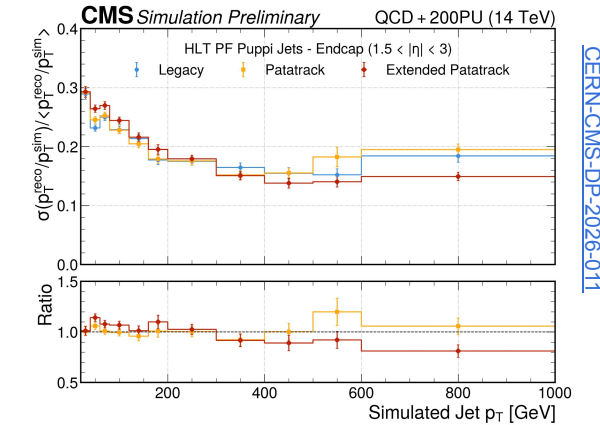
Physics Validation

- Limited CMS tools to perform validation, @ HLT and offline
 - Crucial to compare algorithms, conditions and software releases
 - Carefully monitor Phase-2 developments
- During the last year, NGT has developed validation tools for:
 - Jet/MET and Taus
 - **Primary vertex:** outcome of the [19th Patatrack hackaton](#)
 - **PF ECAL clusters:** part of an effort to deploy a central PF validation tool @ HLT
 - ΔR matching breaks at high PU: use hit-by-hit comparison instead
 - **SimPixelTracks:** new simulated objects enable the performance evaluation of the Patatrack pixel tracking algorithm
- To improve the definition and usage of MC Truth Information: [NGT MCTI Workshop](#)
 - MC information is crucial for reconstruction validation, performance studies, machine-learning training and algorithm developments for Phase-2
 - A shared, consistent and complete definition of truth info is being developed across multiple WPs

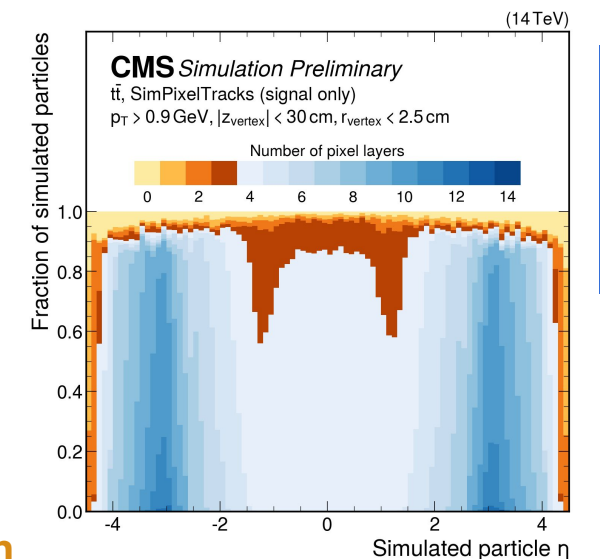


Simulation chain mimics the pixel track reconstruction

Jet Validation: Endcap Resolution



SimPixelTrack Performance



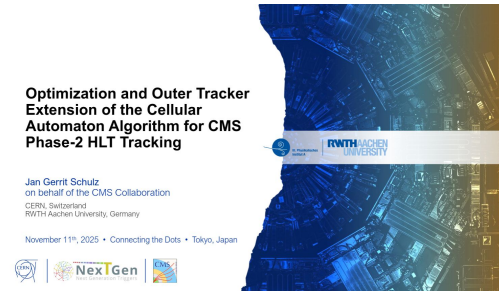
Deliverables and next steps

- Accelerate the HLT tracking and muon reconstruction without major physics compromises, successfully exploiting heterogeneous architectures and efficient data structures ✓
- Provide multiple developments for a new validation suite ✓
- Ongoing and planned studies
 - Improve the muon tracking iteration with new selection strategies
 - Optimize the tracking and muon reconstruction for displaced tracks
 - Explore new computing solutions for the reco. of secondary vertices
 - Exploit ML techniques for improving the removal of fake tracks
 - Explore new objects for improved tracking
 - Further reduce and compress the scouting stream output size
 - Extend the validation software considering MET, PF HCAL clusters and Taus

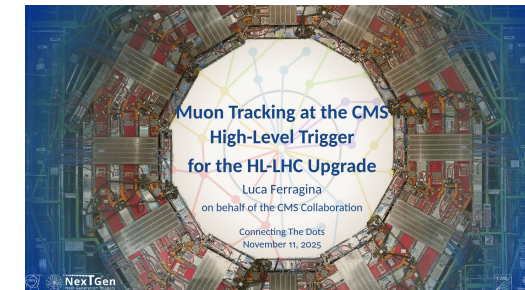
Additional documentation

- NGT MCTI Workshop [Talk](#)
- [CERN-CMS-DP-2025-076](#)
- [CERN-CMS-DP-2025-077](#)
- [CERN-CMS-DP-2025-083](#)
- [CERN-CMS-DP-2026-011](#)

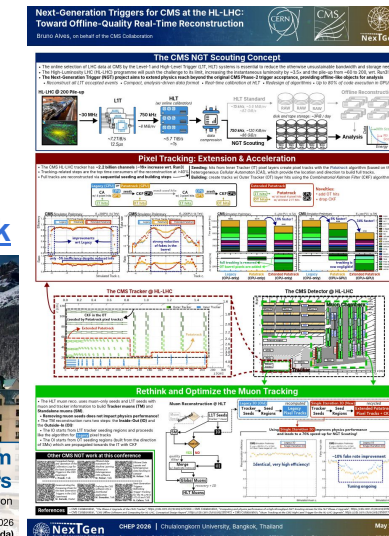
CTD'25, J. Schulz, [Talk](#)



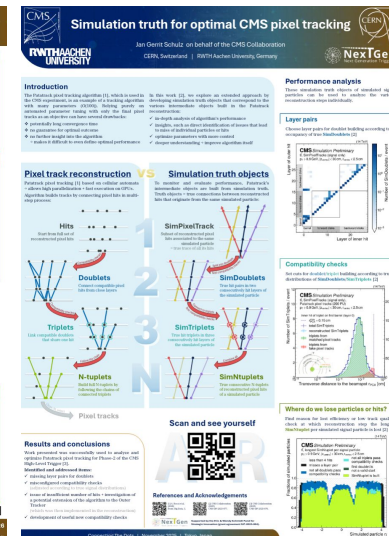
CTD'25, L. Ferragina, [Talk](#)



CHEP'26, B. Alves, [Poster](#) CDT'25, J. Schulz, [Poster](#)



LLWI'26, E. Vernazza, [Talk](#)



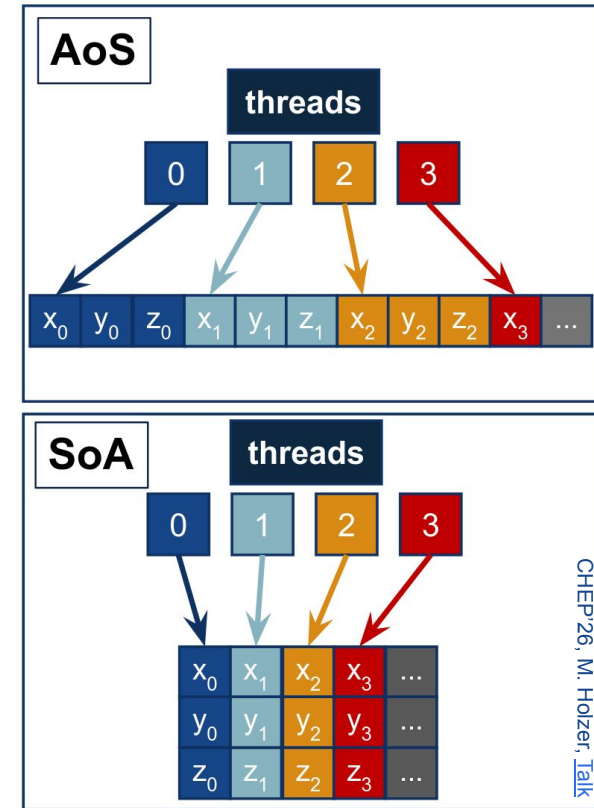
Efficient Data Structures

Requirements:

- Flexible implementation of Structure of Arrays (SoAs)
 - Access the data with **user-friendly and intuitive “AoS-like” operators**
- **Single memory buffer** to minimize data allocations
- **Same memory layout on all architectures** to simplify data transfers
- Separation of Concerns:
 - **Layout:** Organizes the buffer into columns and handles persistent storage
 - **View:** Data access (in AoS-like fashion for user friendliness)
 - **PortableCollection:** Templated on the layout, manages the memory buffer

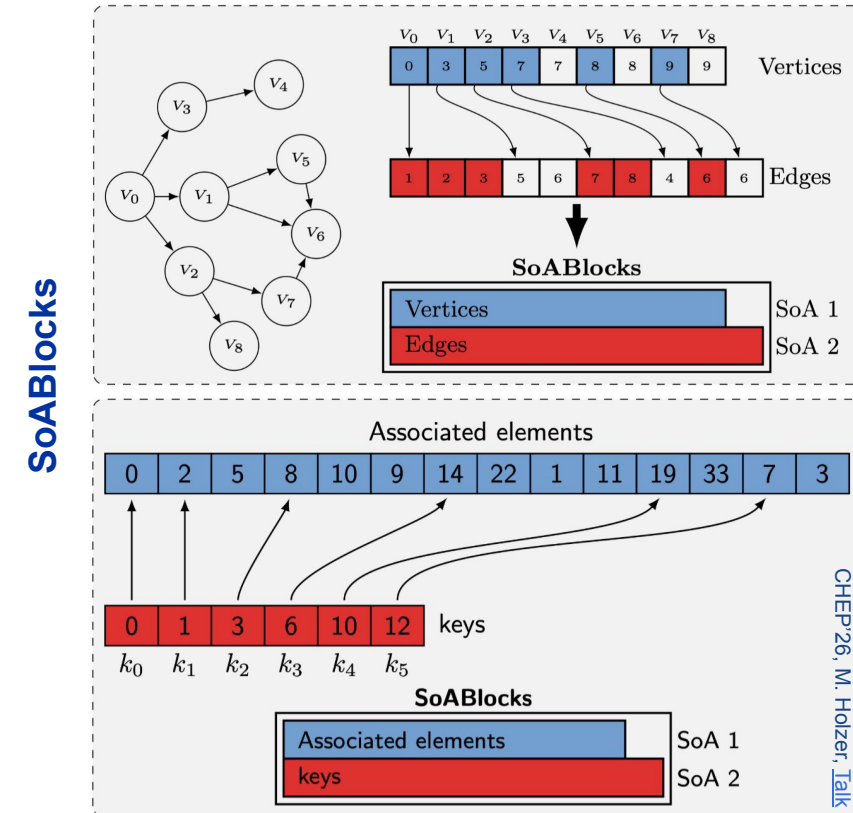
Challenges:

- Flexible composition of columns from different SoAs
- User-friendly macro- and template-based implementation (fully supported in C++20)
- Minimize data copying in heterogeneous Machine Learning workflows (in collaboration with WP 1.7)
- Efficient serialization: strong collaboration with the ROOT team to develop RNTupleSoAs

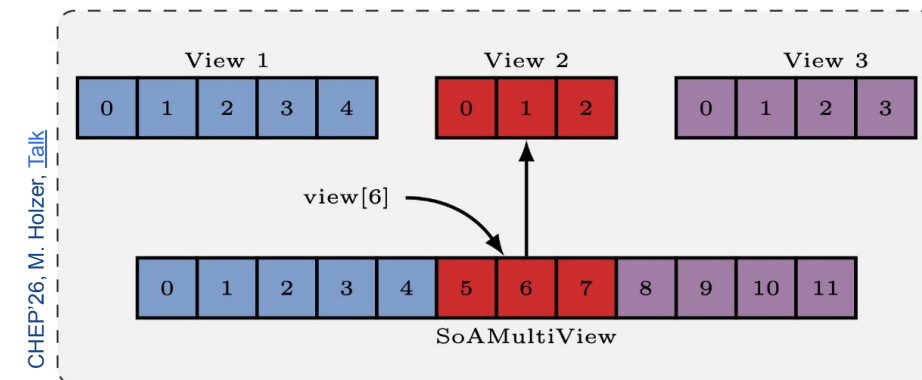


Flexible Data Structures

- **SoABlocks:** combine multiple SoA layouts
 - used to represent complex map- or graph-like structures
 - layouts can be completely different!
- **SoAMultiView:** enables the iteration over multiple SoAs with different sources but identical layouts
 - **no performance penalty**
 - under development
- **Serialization with ROOT**
 - Recent focus on the persistent storage of SoAs
 - Collaboration with the RNTuple team to **avoid data conversions / copies between SoAs and ROOT in the future RNTuple integration**
 - discussions on ROOT's [automatic scheme evolution](#)



SoAMultiView



Efficient Memory

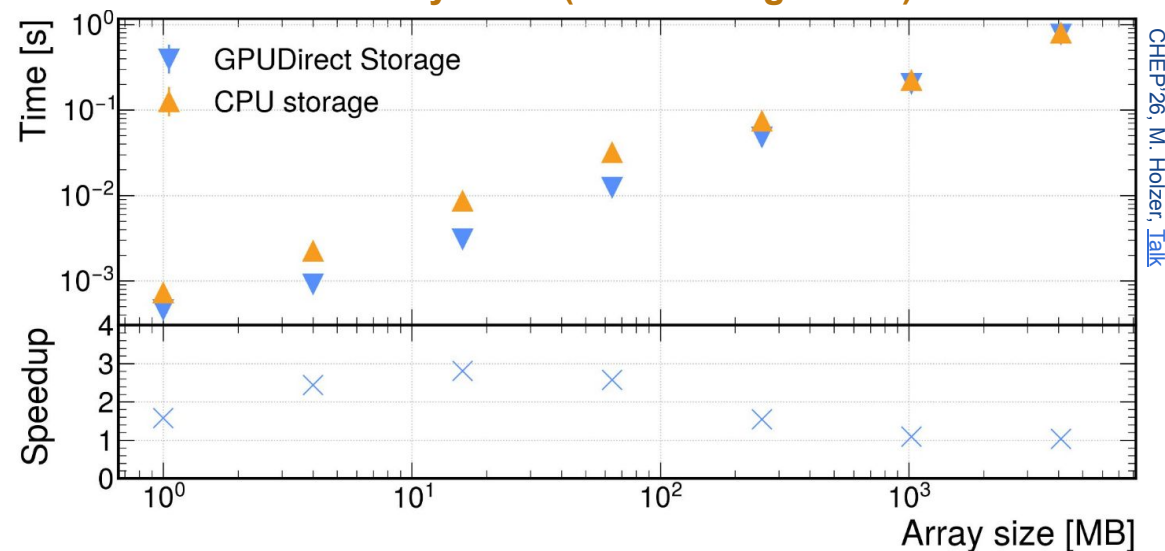
- **GPUDirect Storage**

- bypass the CPU: copy data from storage (NVMe) to GPU memory
- avoids extra copies in the CPU buffers
- tested using synthetic benchmarks

- **For the benefit of all: ML integration**

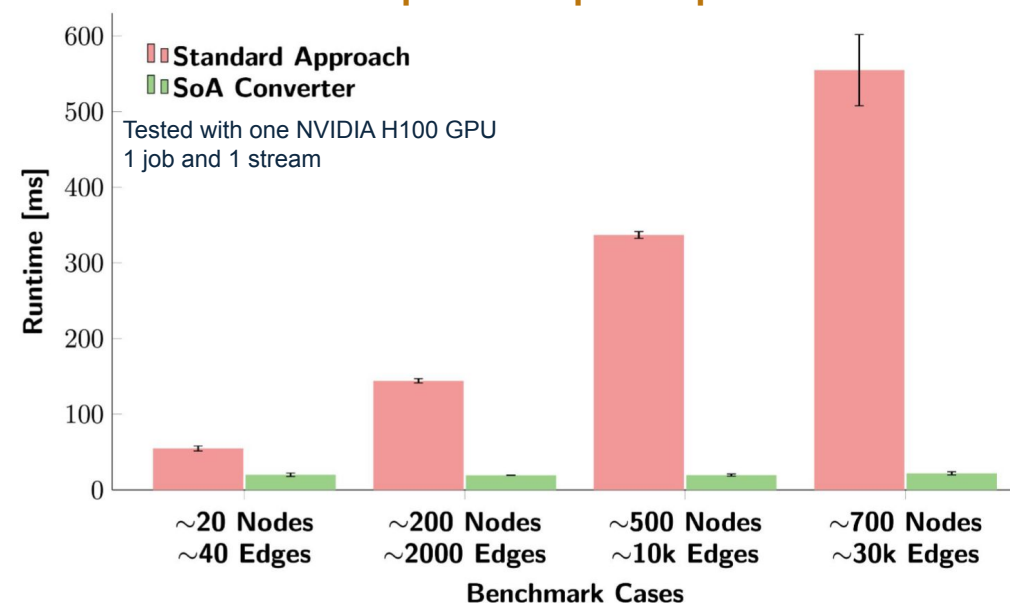
- In collaboration with WP1.7: **CMS SoAs can now directly interact with PyTorch's memory layer**
- Efficient memory reuse
- Successfully tested under production-like conditions!

Very significant benefit for intermediate array sizes (note the log scale!)



CHEP'26, M. Holzer, Talk

Up to 25x speed-up!



ACAT'25, C. Zeh, Poster

Efficient Data Movement for Machine Learning Inference in Heterogeneous CMS Software
C. Zeh^{1,5}, L. Michalski^{2,5}, L. Beltrame^{3,4}, D. Valsecchi⁴, F. Pantaleo⁵, E. Cano⁵
On behalf of the CMS Collaboration
¹Technical University Vienna ²Technical University of Denmark and Technology ³Physics of Higgs ⁴ETH Zurich ⁵European Organization for Nuclear Research (CERN)

Abstract

Machine learning (ML) on GPUs is often slowed by redundant data movement between CPUs, GPUs, and local memory. We present a new efficient data access structure of CMS data directly into PyTorch tensors. This approach integrates a tensor, computes data for different data types, and uses memory to buffer tensor blocks for direct GPU access.

Our approach eliminates unnecessary memory copies and accelerates model execution. By integrating with the data layout, it also supports heterogeneous environments. The result is improved GPU efficiency and simplified integration of ML models into high-performance workflows. It reduces data movement time and improves the use of the GPU data, this is tested on 1000+ in the benchmarks.

Motivation

- Adoption of efficient memory structures like Tensors and PyTorch Collections.
- Lack of a streamlined interface for ML model inference in CMS-based workflows.
- CMS requires multiple copies and linear conversions, leading to inefficiency.
- TensorFlow has limited GPU support for SoA usage.
- Users rely on ad-hoc memory management, such as buffer conversion.
- Optimizing to directly interact with memory using `torch.tensor` (e.g., `torch.tensor`) is not readily available in TensorFlow (TF).

Heterogeneous Inference

- Integrating PyTorch operations with a control over buffering, device streams, and resource management to create safe and scalable ML inference.
- Tensor (e.g., PyTorch) internal tiling is disabled.
- Input tiling of an iteration is based on CMS's stream path, ensuring no unmanaged dataflow.
- Stream device execution: ML operations run on CMS's managed CUDA/HIP streams using device-based resource control.
- Kernel Design: Efficient for both large and small batch sizes and high-throughput processing steps.
- Host Device: Ready-to-use conversion across multiple GPUs, with CPU-side memory.
- Device-side memory management: Host and device memory are managed by the host device, triggered by the first scheduled event.

Model Compilation for Low-Latency Inference

To enable deep learning in high-energy physics workflows, integrate both ahead-of-time (AOT) and just-in-time (JIT) compilation strategies.

just-in-time

- Compile models dynamically at runtime.
- Good for model development and testing.
- Supports on-the-fly execution without build-time preparation.

ahead-of-time

- Use TensorBuilder and Tensor to compile models into optimized libraries for specific hardware (e.g., NVIDIA GPUs).
- Produce a shared library, loaded at runtime using standard C++ interfaces.
- AOT should be done ahead of production to reduce runtime overhead, reproducibility, and maintain performance.

Copy-Efficient Memory Buffer Manipulation

Optimizing data movement is essential for efficient ML on heterogeneous hardware, such as Higgs and CMS. We introduce a flexible interface that converts structured physics data into PyTorch tensors without copying.

- Zero-Copy Transfer:** Direct conversion to PyTorch tensors, no data movement.
- Inline Data:** Complete time-based copying using C++ templates.
- Low-Overhead:** High performance without low-level memory management.

Benchmarking Results

Implementation Performance on CMS Backend

The benchmarking was conducted on a GPU for reconstruction, integrating segmented particle-level and code-level benchmarks.

Deliverables and next steps

- Flexible SoA datastructure deeply integrated within the CMS software ✓
 - used in production for data taking in Run-3
- Complex use cases covered by custom SoA extensions ✓
- Direct integration in Machine Learning workflows ✓
 - no data copies required
- Future: Try GPUDirect Storage in realistic scenarios
 - A Summer student will soon be working on it!
- Continue collaboration with the RNTuple team

CHEP'26, F. Pantaleo, [Talk](#)

Efficient data movement for Machine Learning inference in heterogeneous CMS software

Christine Zeh^{1,2}, Lukasz Michalski³, Emanuele Coradin⁴, Leonardo Beltrame⁵, Eric Cano¹, Felice Pantaleo¹, Davide Valsecchi⁶, Markus Holzer¹
on behalf of the CMS Collaboration

¹CERN, ²Imperial College London, ³AMD, ⁴Università di Padova, ⁵Politecnico di Milano, ⁶ETH Zurich

28th Conference on Computing in High Energy and Nuclear Physics

Efficient Data Layouts and heterogeneous data handling in CMSSW

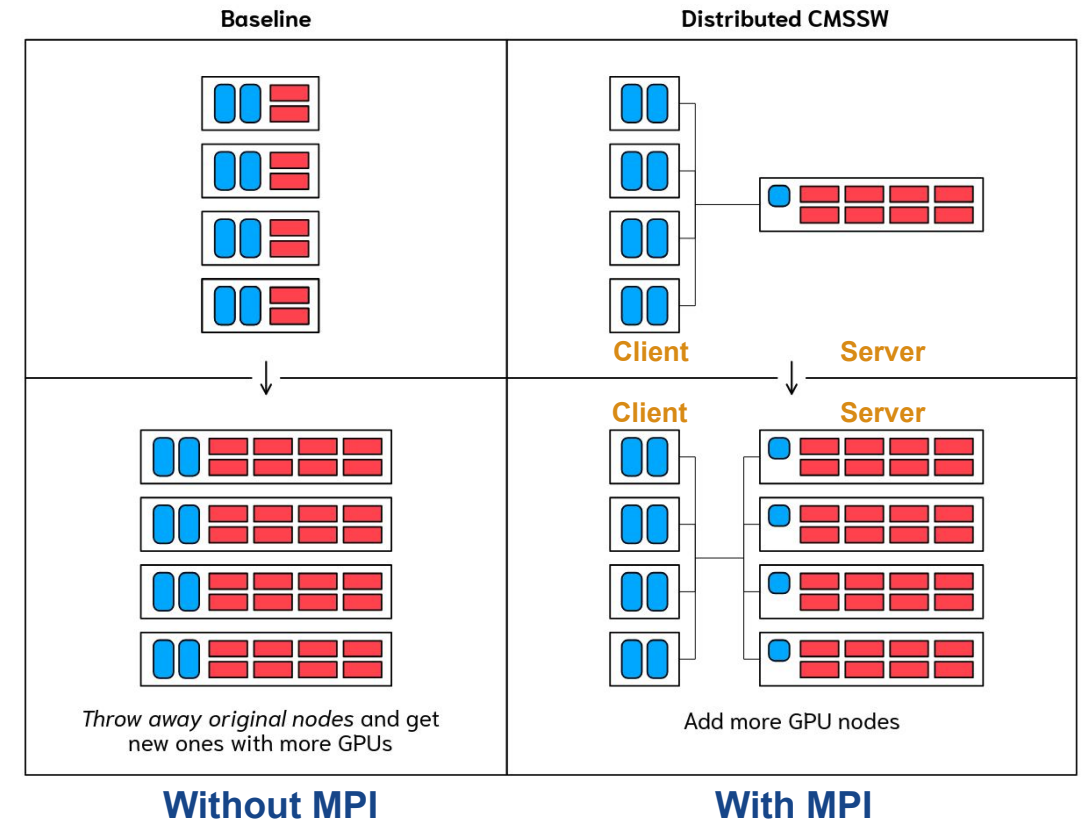
Markus Holzer¹, Simone Balducci^{1,2,3}, Leonardo Beltrame⁴, Felice Pantaleo¹, Andrea Bocci¹
on behalf of the CMS Collaboration

¹CERN, ²University of Bologna, ³INFN-CNAF, ⁴Politecnico di Milano

28th Conference on Computing in High Energy and Nuclear Physics

Evolving CMSSW into a distributed application

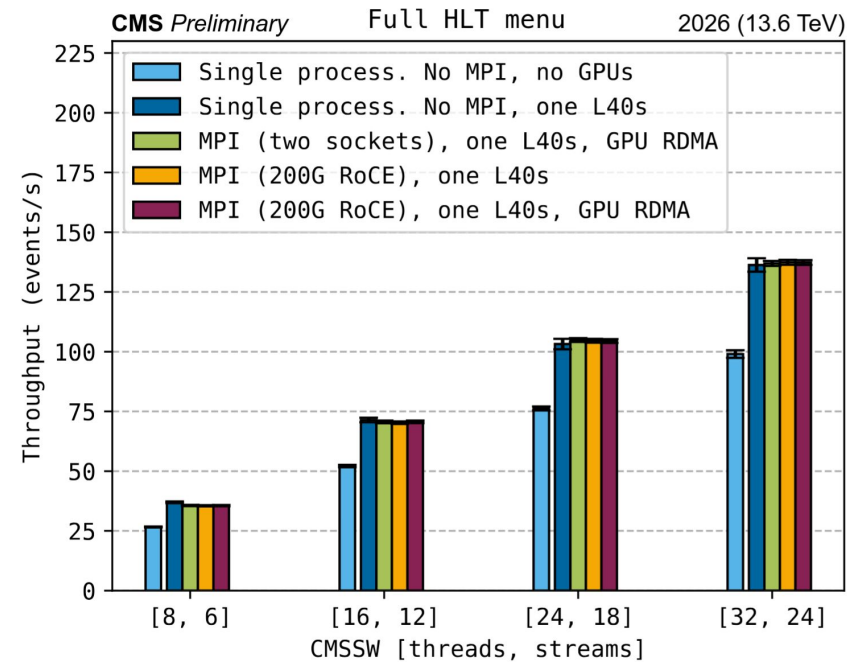
- The CMSSW framework is constantly evolving
 - Data processing is being increasingly ported to GPUs
 - Legacy code is being overhauled or removed
 - The available hardware is also evolving
- The answer: distributed application with MPI
 - Offload part of the processing to remote accelerators
 - Allow independent scaling of CPU and GPU resources
- Implemented new “trivial serialization” mechanism:
 - Describe memory layouts as “address, size” pairs
 - Directly send and receive data to/from GPU memory



Throughput results

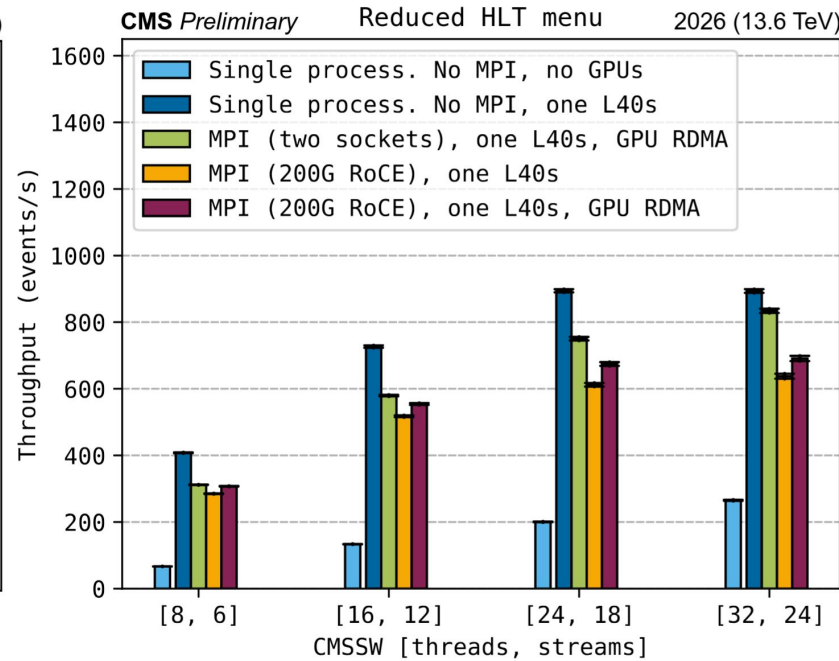
- Trade-off between MPI network overheads and the speed-up provided by remote accelerators

No penalty from using MPI



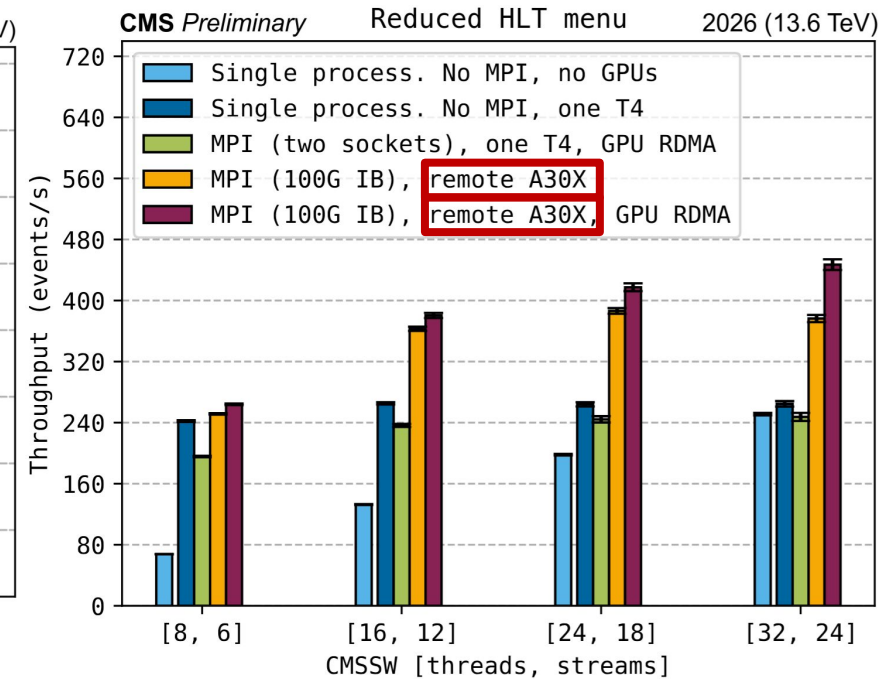
CPU-bound HLT menu

Significant penalty from using MPI



GPU-bound (reduced) HLT menu server GPU == client GPU

The overhead is largely compensated from using more powerful accelerators

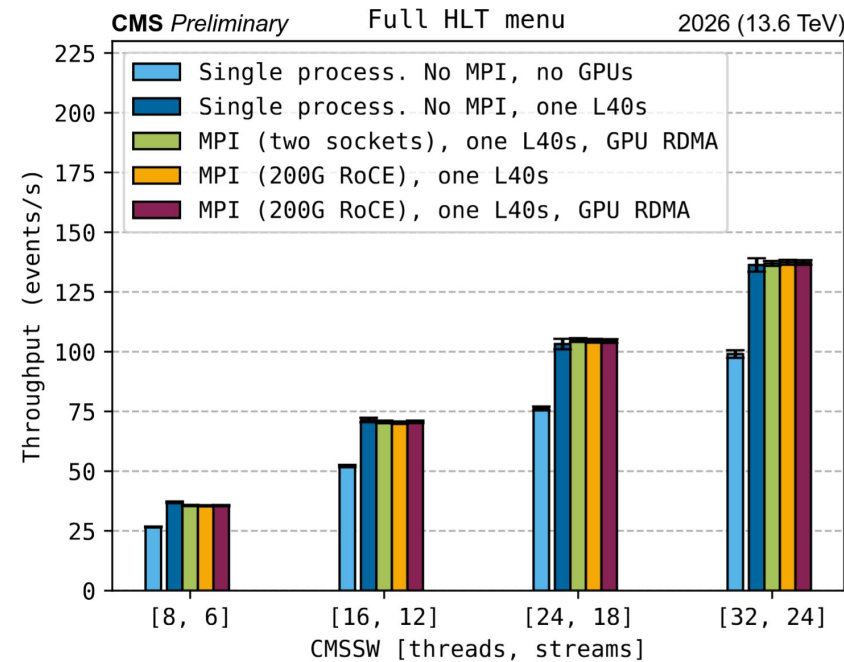


GPU-bound (reduced) HLT menu server GPU >> client GPU

Throughput results

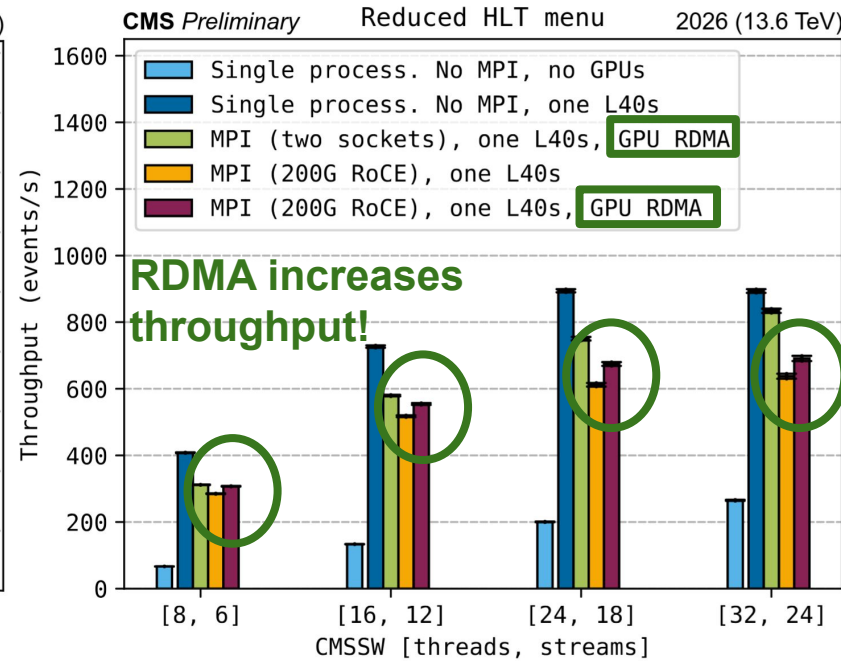
- Trade-off between MPI network overheads and the speed-up provided by remote accelerators
- **Study the impact of RDMA:** by-passing the server's CPU when transferring data from the server to the client

No penalty from using MPI



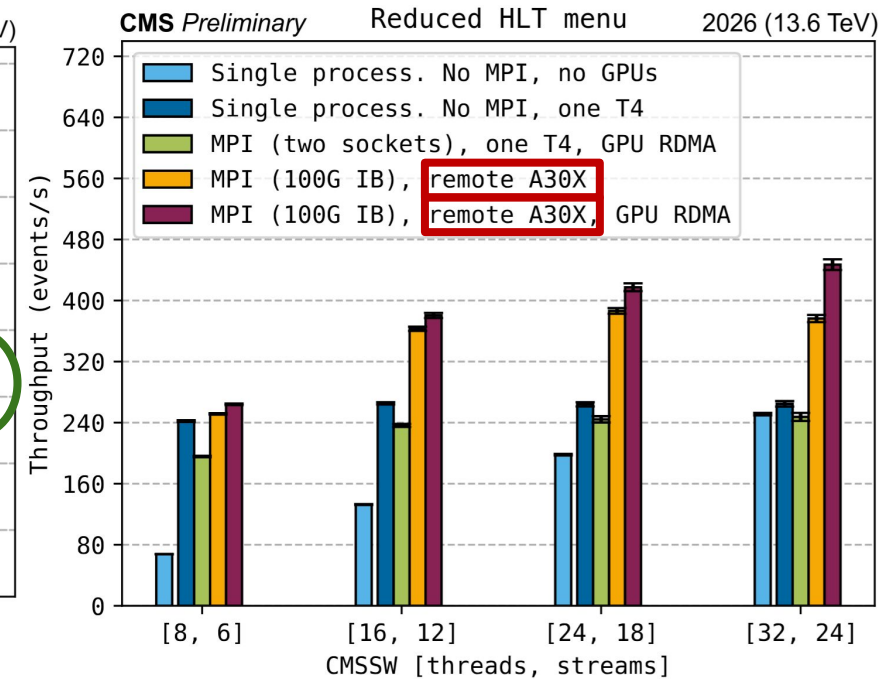
CPU-bound HLT menu

Significant penalty from using MPI



GPU-bound (reduced) HLT menu server GPU == client GPU

The overhead is largely compensated from using more powerful accelerators



GPU-bound (reduced) HLT menu server GPU >> client GPU

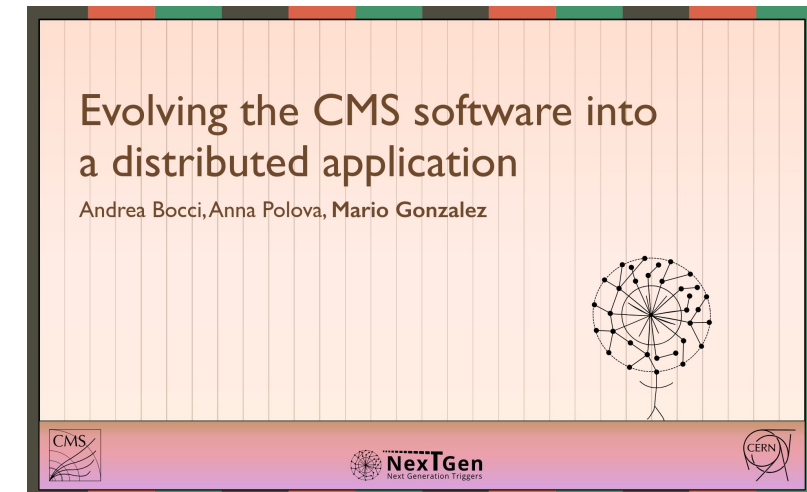
Deliverables and Next Steps

CMS Week, A. Polova, [Talk](#)

- The simultaneous offloading of parts of the HLT to remote servers with different accelerators has been demonstrated ✓
 - Opening the possibility of offloading specific HLT stages to the remote accelerators where they run most efficiently
 - Targeting multiple vendors at once becomes a possibility
- Ongoing and planned studies:
 - Reduce MPI's overhead by compressing data before sending
 - Further optimize MPI configurations, or using lower-level libraries like UCX instead
 - Compare different network protocols and interconnects (Infiniband and RoCE)
- Documentation
 - [Task 3.2 2025 Report](#)

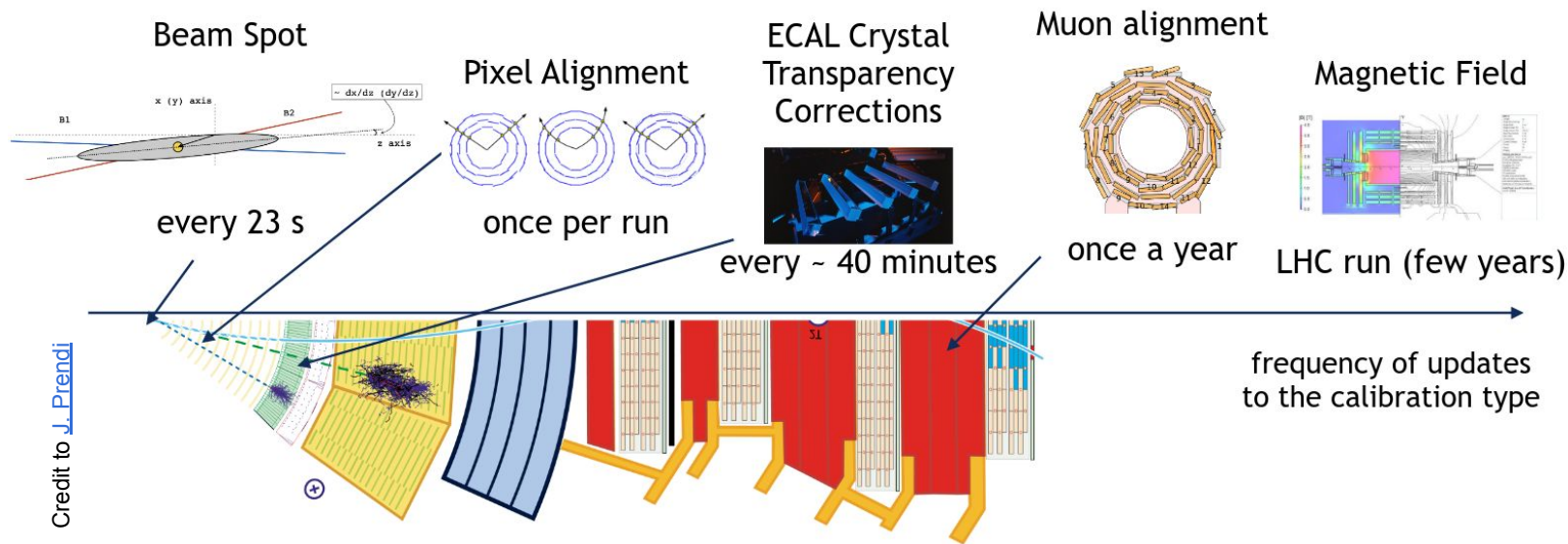


CHEP'26, M. Gonzalez, [Talk](#)



The Need for Optimal Calibrations @ HLT

- Historically, offline calibrations are done more frequently and are more precise than at HLT
- NGT scouting requires fast, optimized and accurate calibrations directly at HLT
 - Buffer the full HLT input data for several hours, requiring $O(100 \text{ PB})$ storage with $\sim 7.2 \text{ TiB/s}$ @ 750 kHz
 - Calibration constants are derived in-situ during data taking, exploiting data buffering
 - Only **one chance** to reconstruct HLT data: **calibrations need to be optimal!**

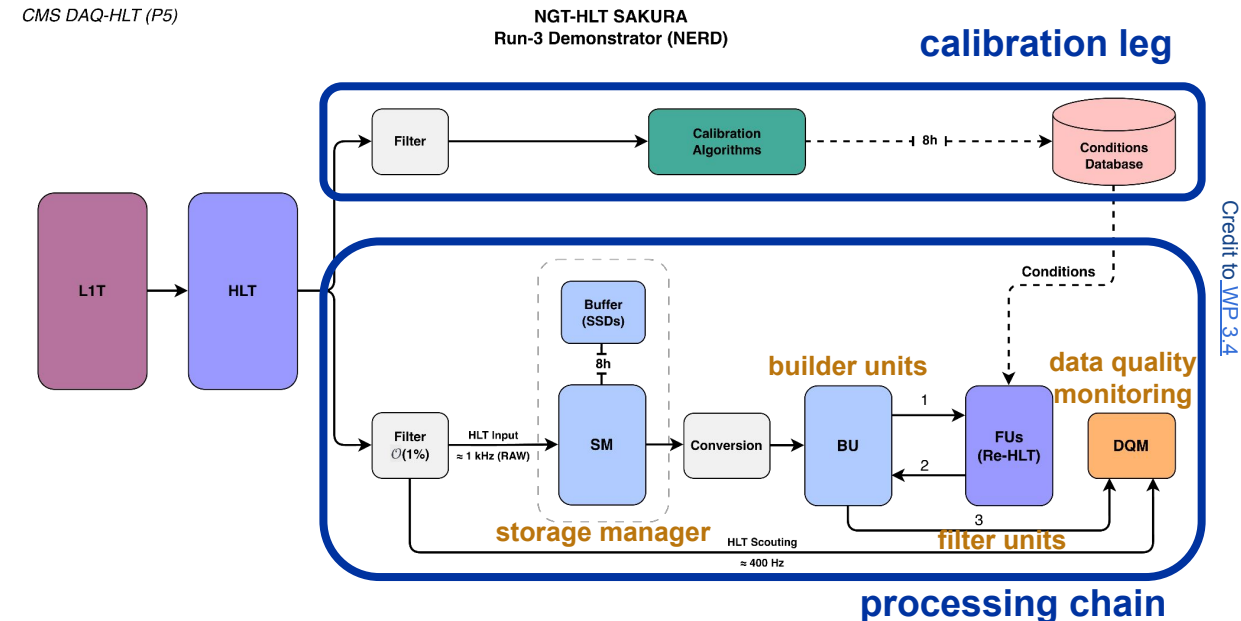


The NGT Demonstrator

Goal: Demonstrate the feasibility of the Phase-2 NGT HLT calibration, using the current Run-3 system as a benchmark

Workflow:

- Route $\mathcal{O}(1\%)$ of the HLT input data (~ 1 kHz) to a dedicated DAQ storage manager node at P5
- Buffer the data for 8 hours on SSDs
- Derive or pick-up improved calibrations within 8 hours
- Re-run the full HLT with the new calibrations
- Evaluate and compare physics performance between old and new calibrations



A significant effort was devoted to setting up the required hardware and software infrastructure!

The NGT Calibration Loop

Three tasks must be orchestrated:

- Event reconstruction with simplified tracking
- Deriving or picking-up calibrations
- Upload the calibration results to the conditions database

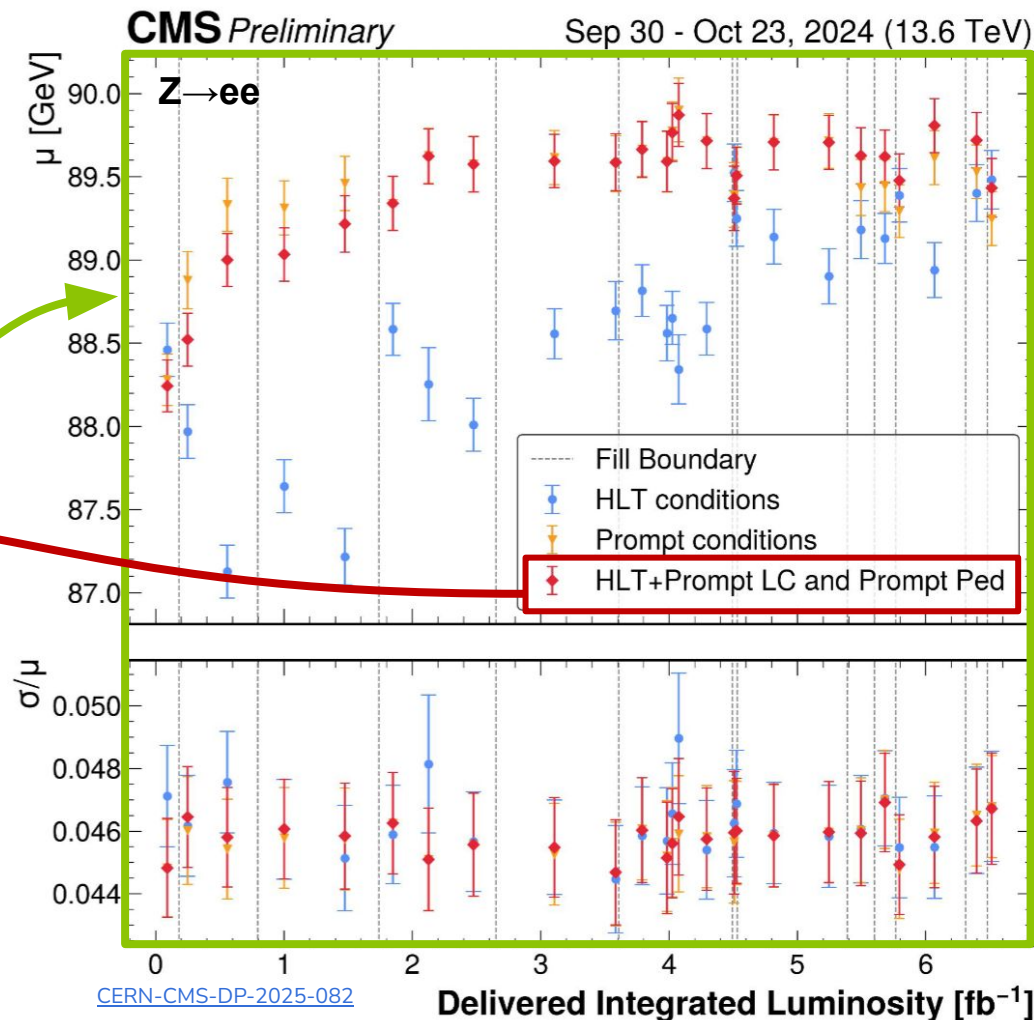
Proof of concept using offline calibrations @ HLT

- ECAL noise pedestals
- ECAL crystals' laser corrections

After successful validation, the calibration was extended:

- Bad components of the silicon strip tracker
- Pixel tracker alignment
- Strip tracker unpowered modules
- Improved beamspot (high-purity)

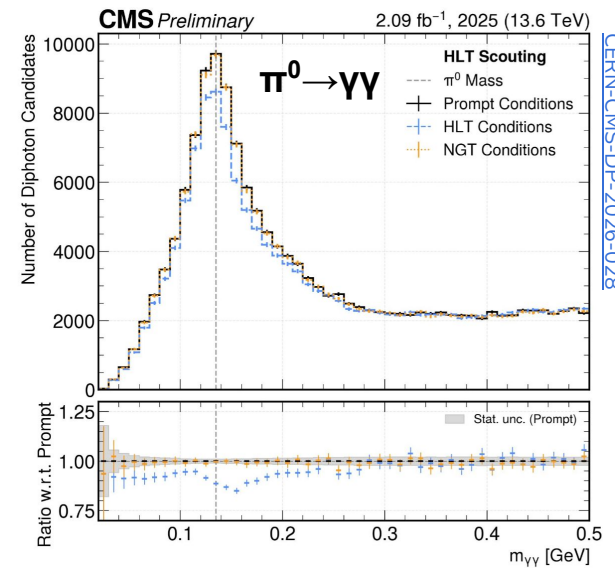
Two offline calibrations are sufficient to recover match prompt performance @ HLT!



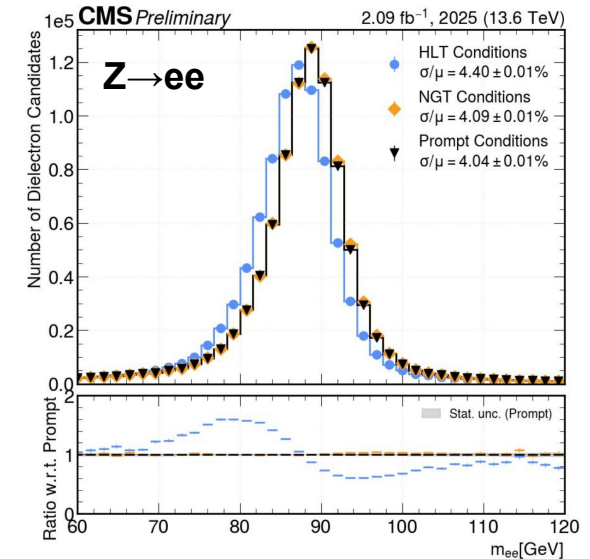
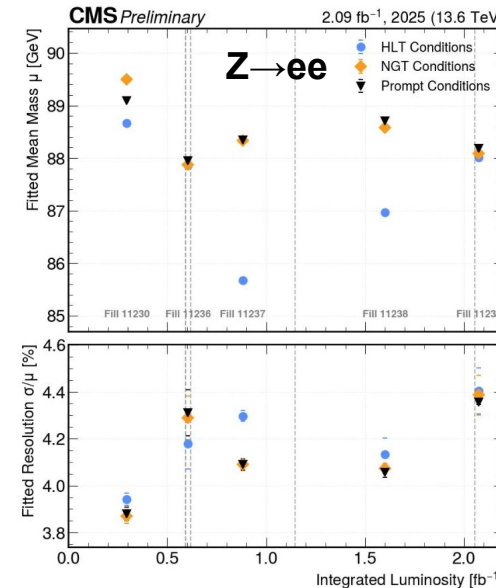
Monitoring and Validation

- Successful integration of the NGT demonstrator infrastructure into CMS monitoring systems
 - Data Quality Monitoring GUI
 - File-based Filter Farm monitoring tool
 - Conditions Database
- Physics validation highlights the **similarity between NGT online calibrations and standard offline calibration**, using:
 - $\pi^0 \rightarrow \gamma\gamma$
 - $Z \rightarrow ee$
 - Reconstructed hits

Compatibility is stable over time for the mass scale and resolution!

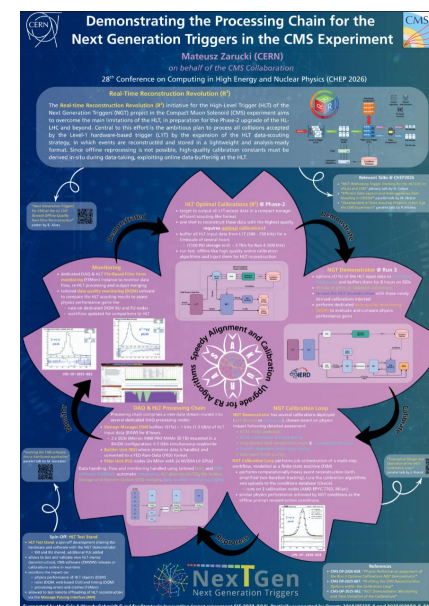
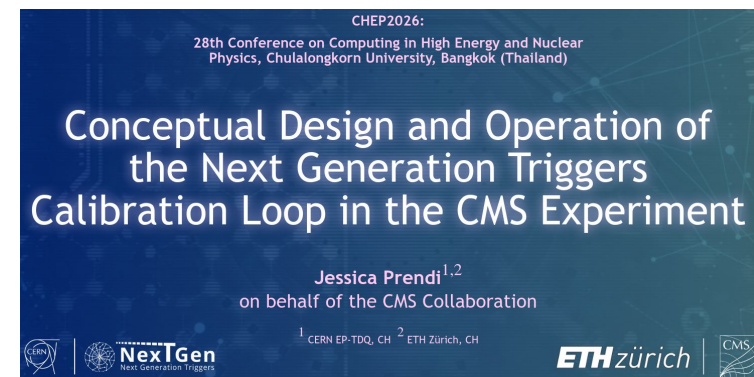


Mass spectra are remarkably similar between prompt (offline) and NGT calibrations !



Deliverables and Next Steps

- The NGT demonstrator was integrated with the rest of the CMS infrastructure, and was continuously run during part of the Run-3 data-taking period ✓
- Physics validation results are extremely promising ✓
- Phase-2 integration and validation represent our priority:
 - Implement a test integration system for LS3 (replay and simulated Phase-2 data) with the features of the NGT demonstrator
 - Detailed assessment of Phase-2 calibrations, including physics impact
 - Study different buffer solutions (SSDs vs HDDs, size, time, cost, ...)
 - Investigate orchestration solutions for the Phase-2 calibration loop (including “rolling” calibrations *à la* beam-spot)
 - Investigate other mechanisms for delivering calibrations to the HLT (e.g. HDF5 files)
 - Explore heterogeneous reconstruction and ML solutions for calibration
- Plan to discuss with LHCb on their real-time buffering and calibration system experience



CHEP'26,
M. Zarucki,
[Poster](#)

Award-winning
poster: 3 prizes
and counting!

Documentation

- [CERN-CMS-DP-2025-082](#)
- [CERN-CMS-DP-2025-087](#)
- [CERN-CMS-DP-2026-028](#)

Rethinking the HL-LHC CMS Real-Time data processing with NGT

But **what if New Physics doesn't look as expected**, and it's instead **buried under the bulk** of the **background events we are throwing away** due to the trigger selections?



Redesign the data collection and scouting strategy to **reduce the need to reject events in the Level-1 and High-Level triggers** aiming at complementing the current workflows.



At Level-1: develop an advanced Scouting workflow that, profiting from the increased granularity and flexible architecture of the Phase-2 CMS experiment, **can allow to perform trigger-level physics analyses at 40 MHz.**



At HLT: replace the trigger filtering task with an event processing task **similar to what happens with offline** events stored on disk and develop the next HLT scouting system: **NGT Scouting.**



Leverage traditional physics-based algorithms and **advanced AI solutions** to more efficiently select signal processes: **both in hardware and software triggers!**

To extend the CMS discovery and precision measurement reach!