

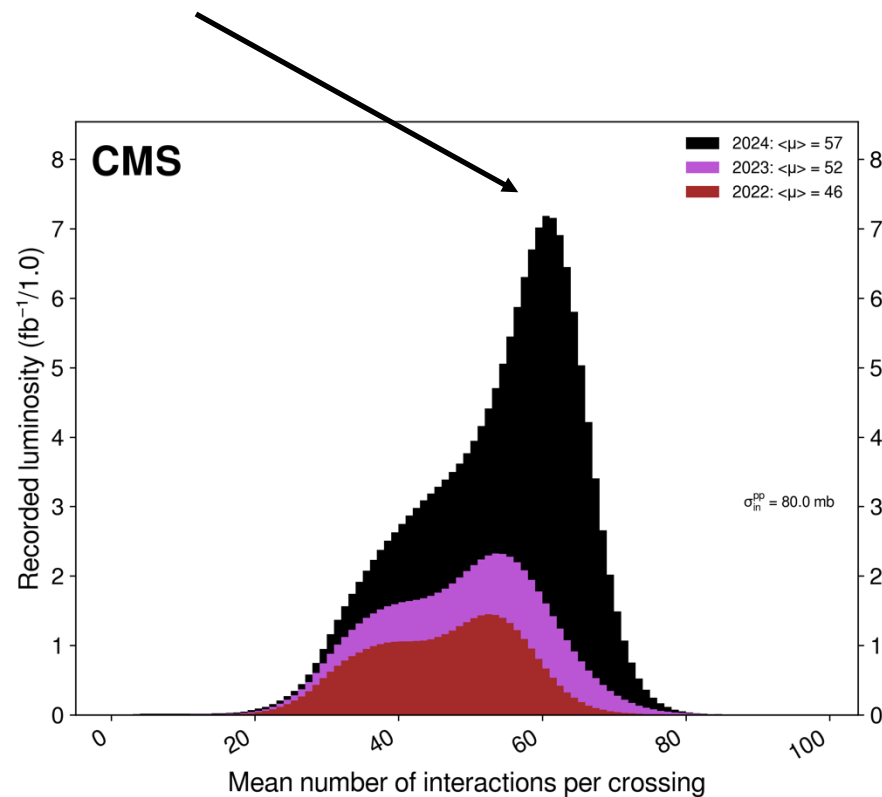
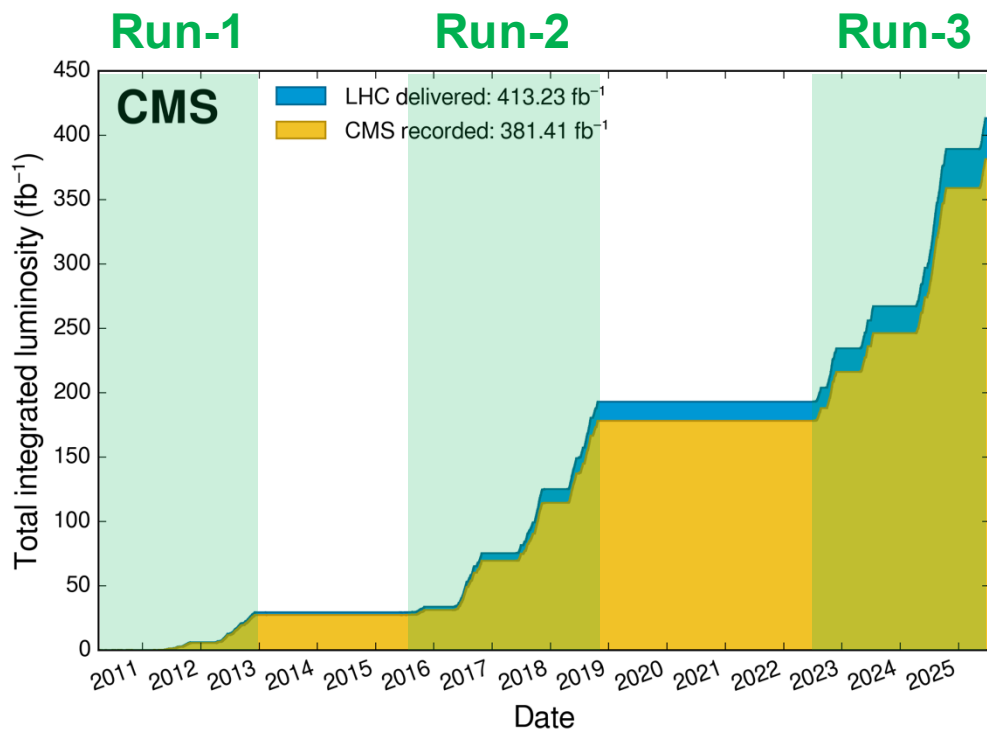
Overview of the trigger and DAQ systems in CMS

Daniele Trocino — *INFN Torino*
on behalf of the CMS Collaboration

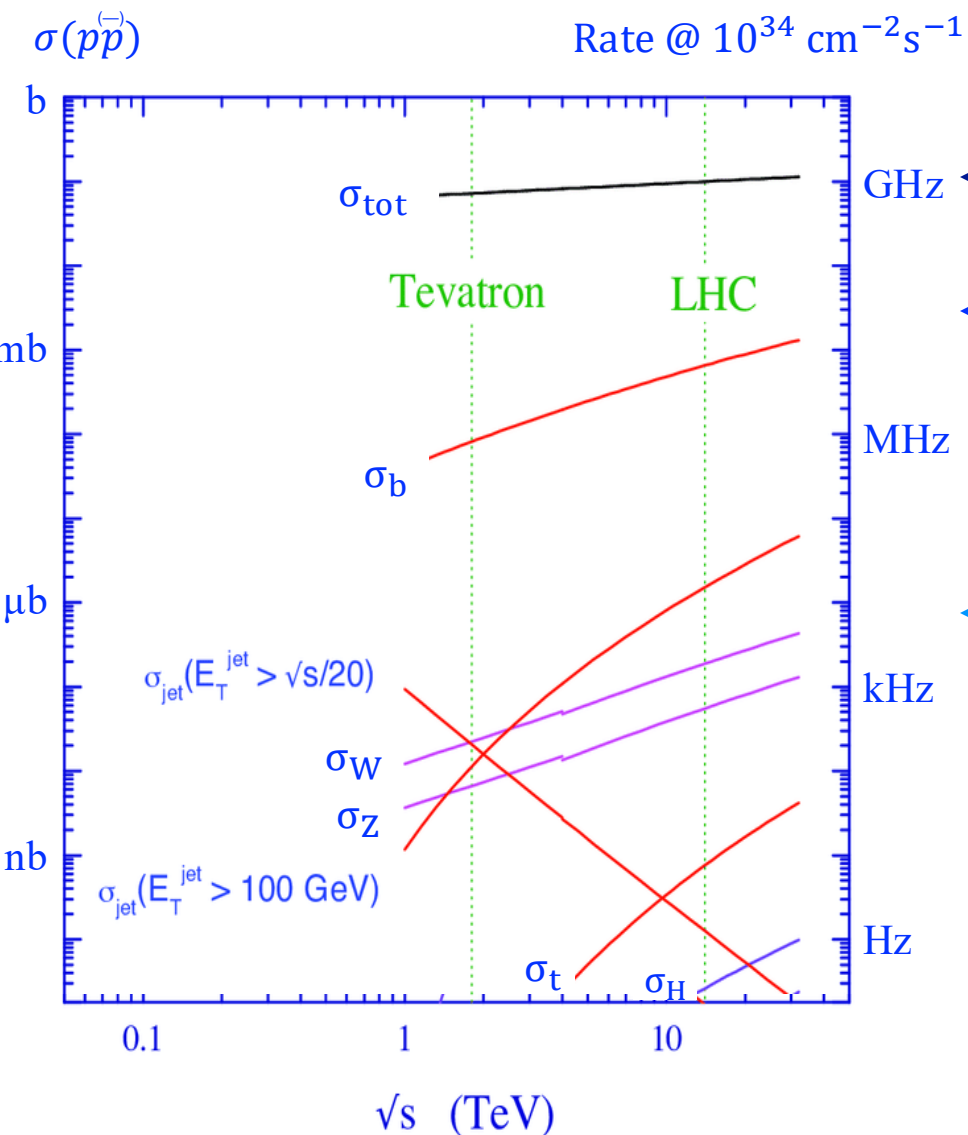
XIV International Conference on New Frontiers in Physics (ICNFP 2025)

22nd July 2025 — Kolymbari, Greece

- The LHC Run-3 (2022–2026) is in progress
 - 13.6 TeV collisions, ~ 30 MHz collision rate, peak luminosity $> 2 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$
 - In 2024–25, average pileup interactions $\langle \text{PU} \rangle \sim 63$ at peak luminosity

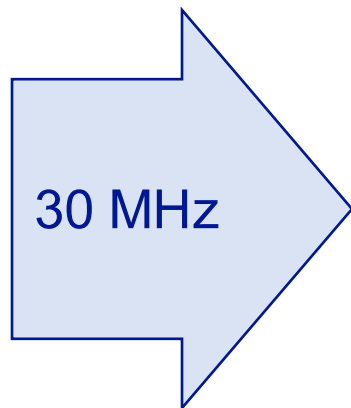
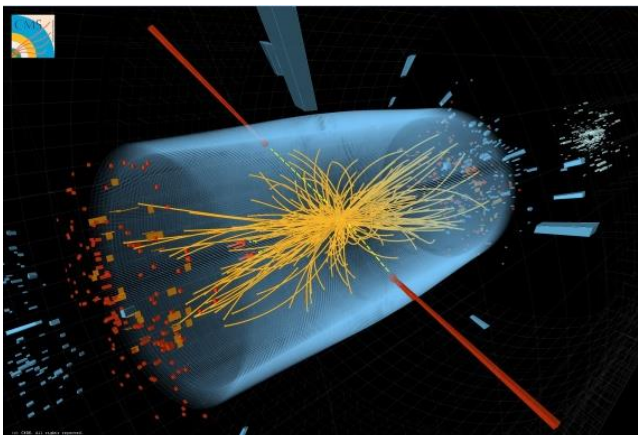


Why do we trigger?

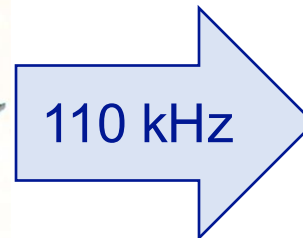


- Bandwidth limited by **data storage** and **analysis sustainability**
 - Physics processes of interest typically have rates **< kHz**
- \Rightarrow **physics-driven selection**
- In CMS this is achieved by means of a **two-level trigger system**

LHC

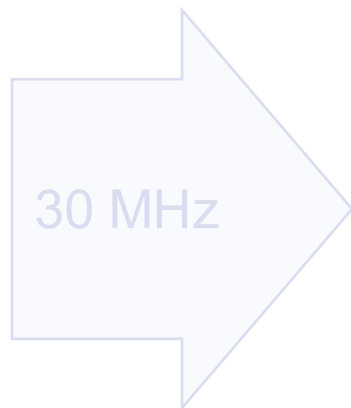
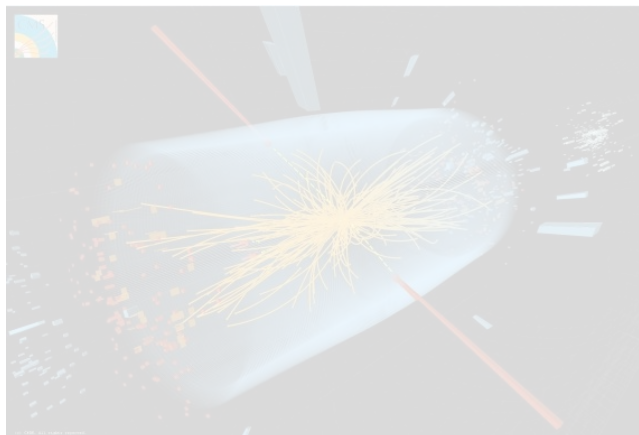


Level-1 trigger

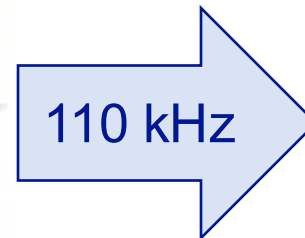
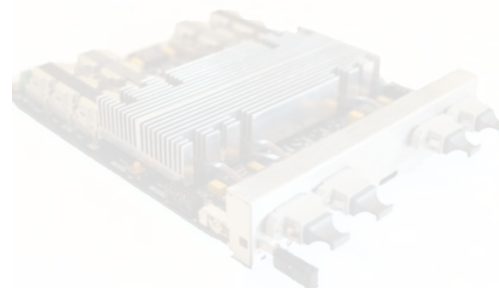


- ~ 30 MHz input rate, maximum latency $< 4 \mu\text{s}$
- Implemented on custom programmable processors (FPGAs), optical links for faster communication
- Partial detector information: only muon spectrometer and calorimeters, with limited granularity
- Maximum output rate allowed by the CMS readout and data acquisition: 110–115 kHz (1 event every ~ 300)

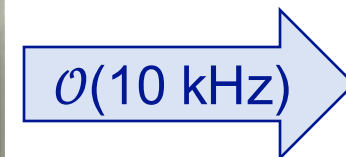
LHC



Level-1 trigger



High-Level trigger

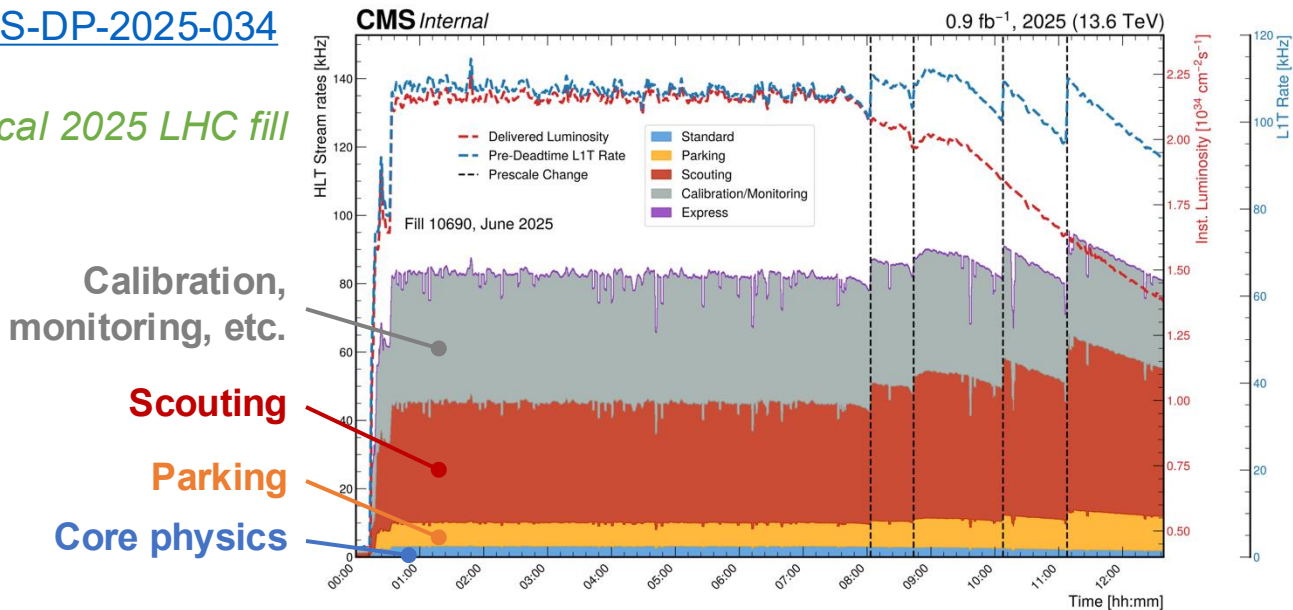


- ~110 kHz input from L1, average processing time ~420 ms
- Full software implementation, running on a CPU + GPU farm
- Information from **all subdetectors** (incl. the inner silicon tracker) with full granularity
- **Offline-like** reconstruction and analysis **software**, optimized for HLT time constraints → ~100× faster
- Average output rate for physics ~ 7 kHz (+ 30 kHz from “HLT scouting”)

NEW!
GPU in Run-3

CMS-DP-2025-034

Typical 2025 LHC fill



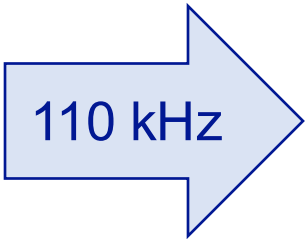
Calibration,
monitoring, etc.

Scouting

Parking

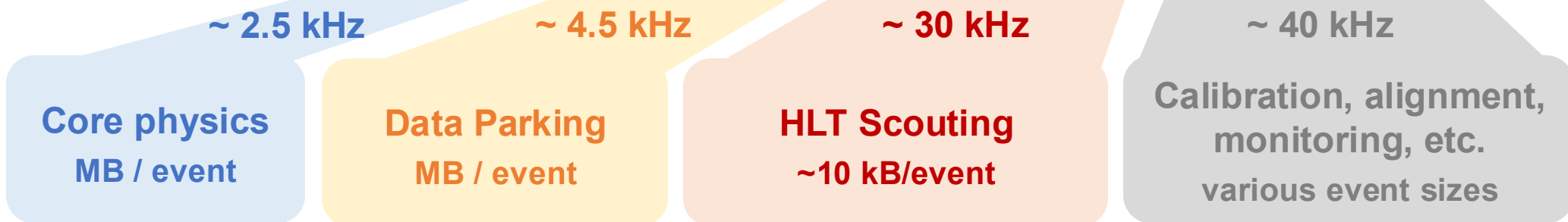
Core physics

High-Level trigger



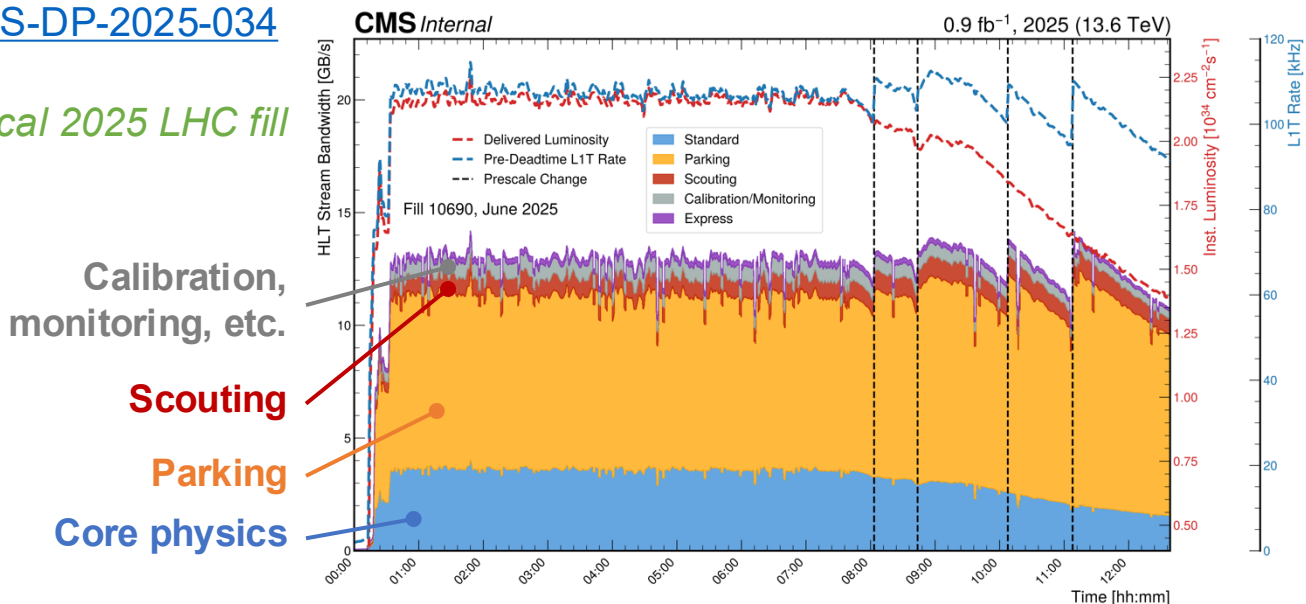
The **HLT output** is split into non-exclusive **streams**,
each with different **event content**
and **event size**

Ballpark numbers
at $\langle \text{PU} \rangle = 60$

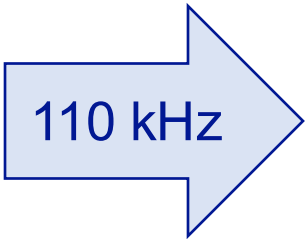


CMS-DP-2025-034

Typical 2025 LHC fill



High-Level trigger



The **HLT output** is split into non-exclusive **streams**, each with different **event content** and **event size**

Ballpark numbers at $\langle \text{PU} \rangle = 60$



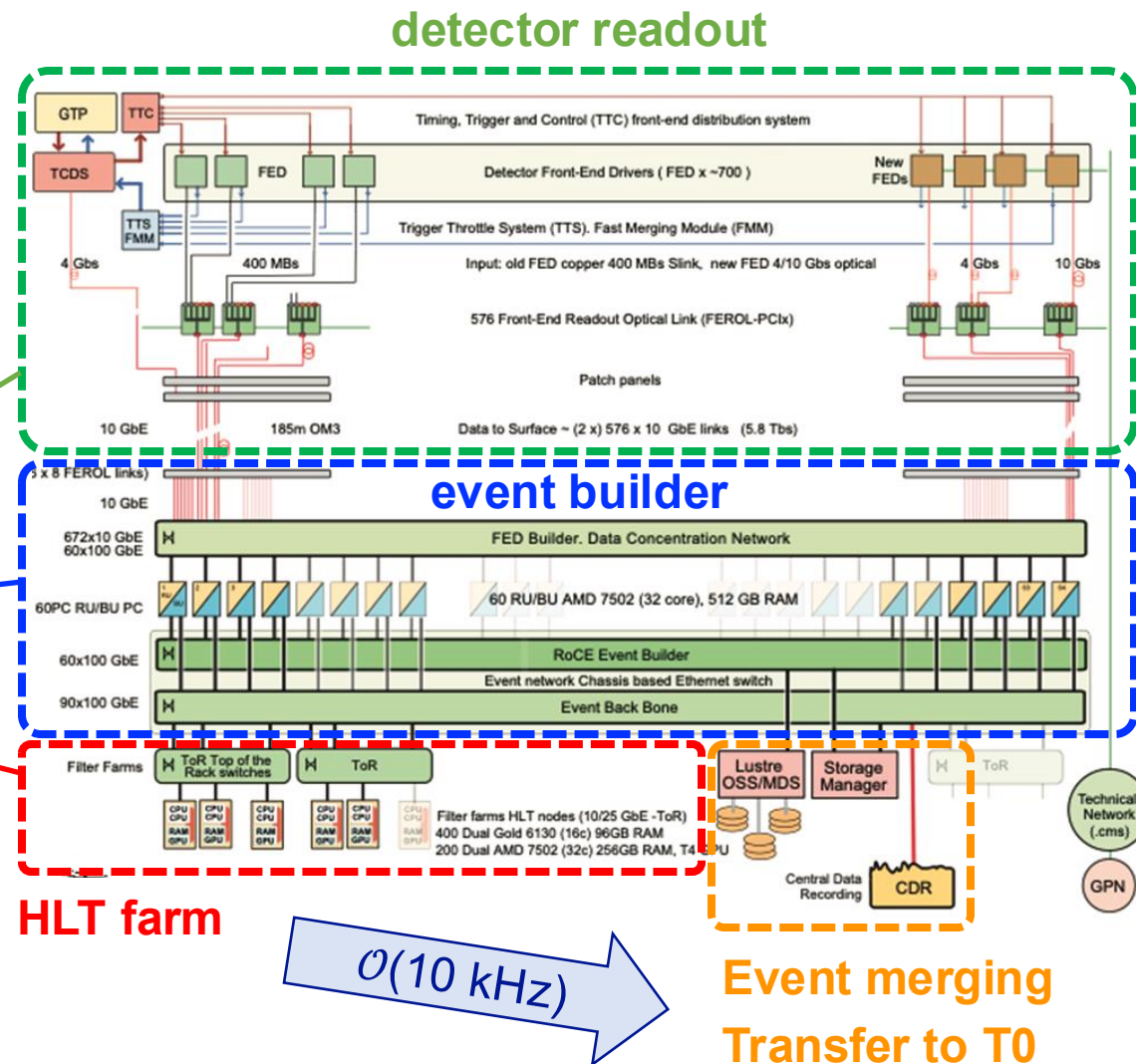
JINST 19 (2024) P05064



110 kHz

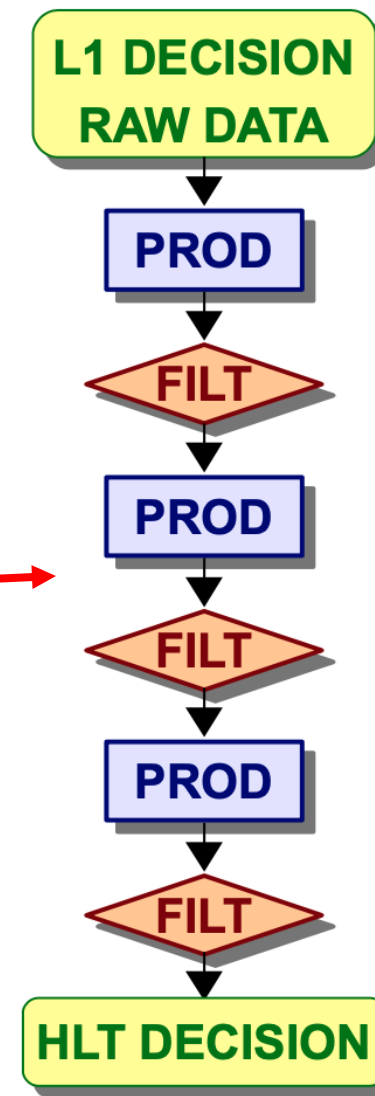
DAQ handles data transfer and event building at the L1-accept rate of ~ 110 kHz

- Read out and transfer data from subdetector front-ends
- Aggregate subdetector fragments to build full events
- Send events to the filter farm that runs the HLT
- At the HLT output rate, events are merged and transferred to the CERN Tier-0 for permanent storage
 - Data throughput for proton-proton in 2025 ~ 16 GB/s (but can handle up to ~ 30 GB/s)



Trade-off of **efficiency**, **rate**, and **processing time**

- **High efficiency** is ensured by sophisticated **offline-like reconstruction** algorithms
- The **rate** can be kept under control by modeling identification criteria after the **offline analysis** selections → *reduce the rate while keeping high efficiency*
- Constraints on the **processing time** are driven by the size and computing power of the **filter farm**
 - **Modular approach (trigger “path”)**
 - Sequence of **reconstruction** and **filtering** modules of increasing complexity
 - Reject events as soon as possible, skip more time-consuming reconstruction
 - **Regional reconstruction**
 - Wherever possible, read out the detector only around L1 or HLT object



- Run-3 HLT farm equipped with hybrid CPU + GPU machines
 - Comprised of ~220 nodes → 30'000 CPU cores + 450 GPUs (NVIDIA T4/L4)



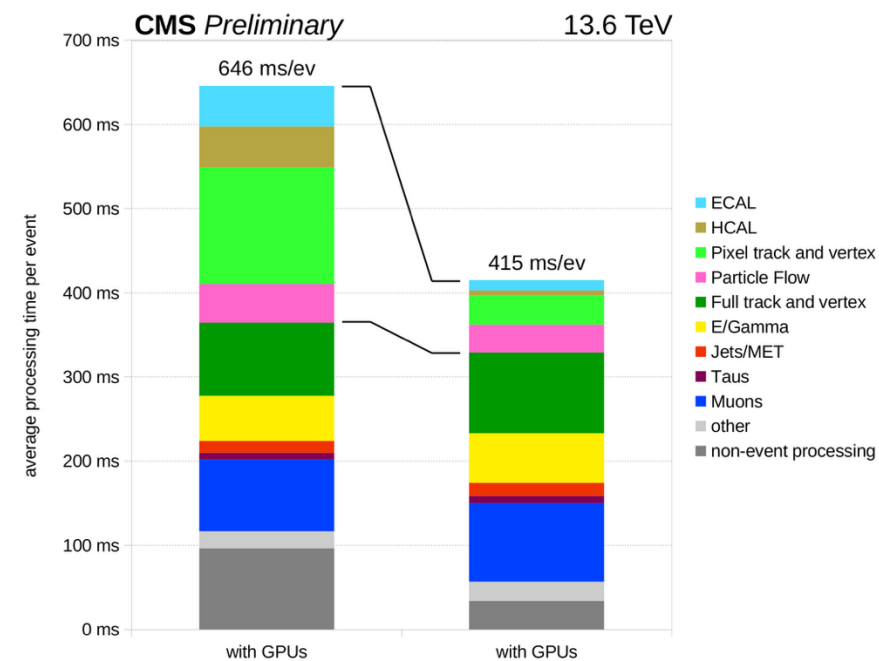
- 35% of reconstruction offloaded to GPUs (Pixel, ECAL, HCAL)

[CMS-DP-2024-082](#)

- 55% faster reconstruction!
- Algorithms written with **Alpaka** (portability library)

- Additional processing power with GPUs allows for

- higher input rates from L1
- more sophisticated reconstruction methods

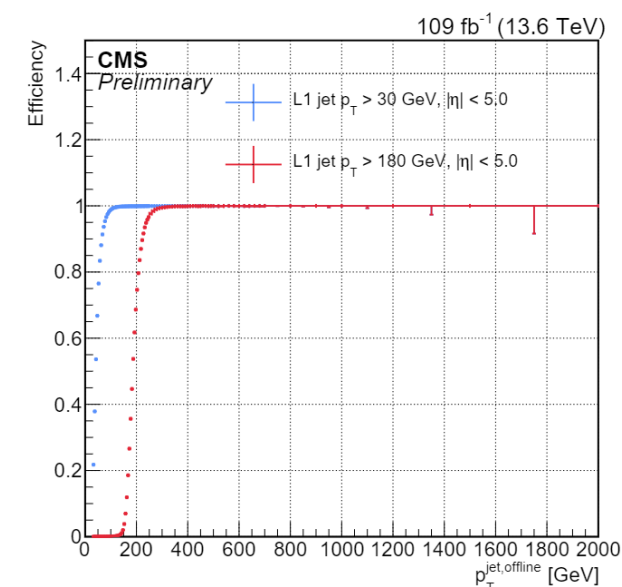
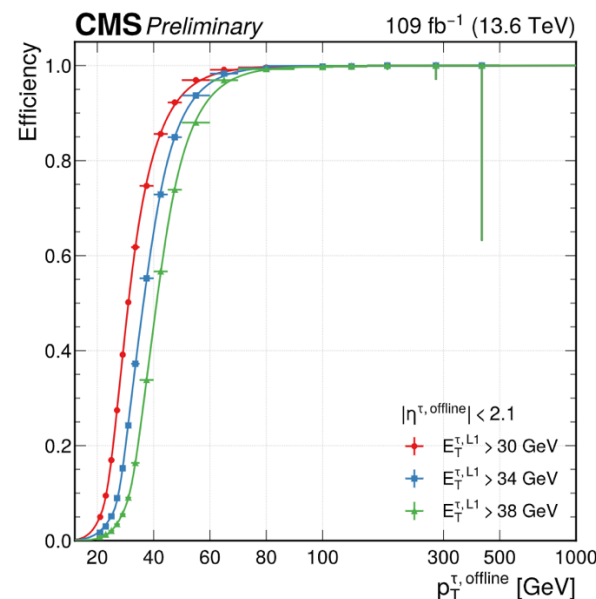
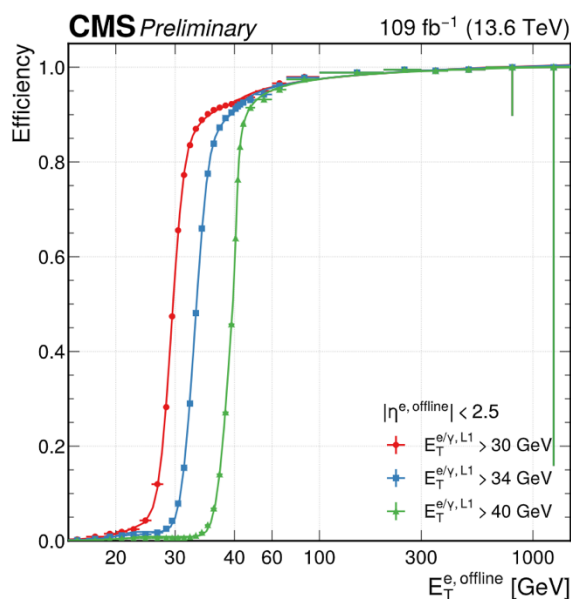
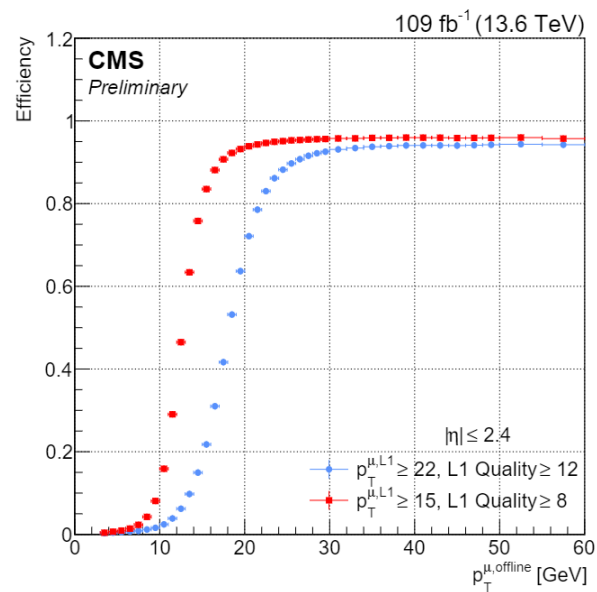


muons

electrons / photons

hadronic τ

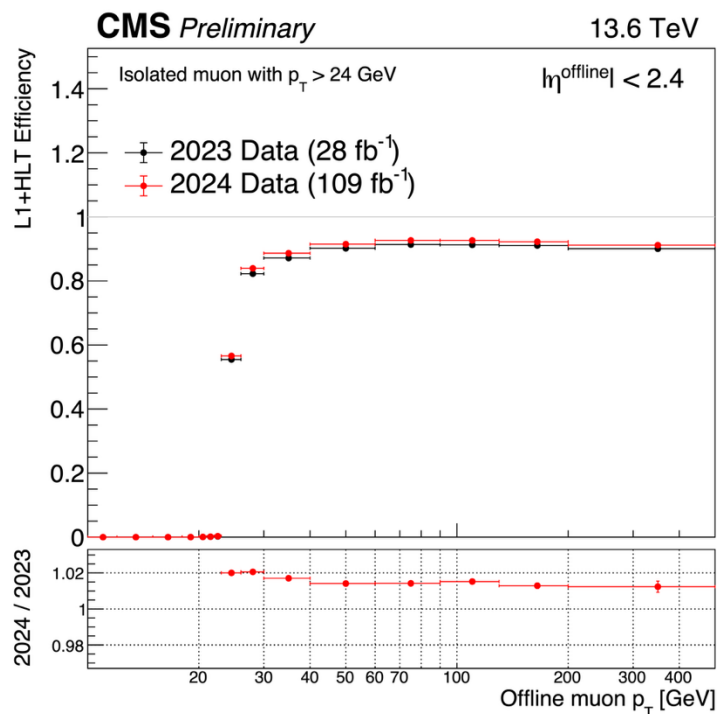
jets



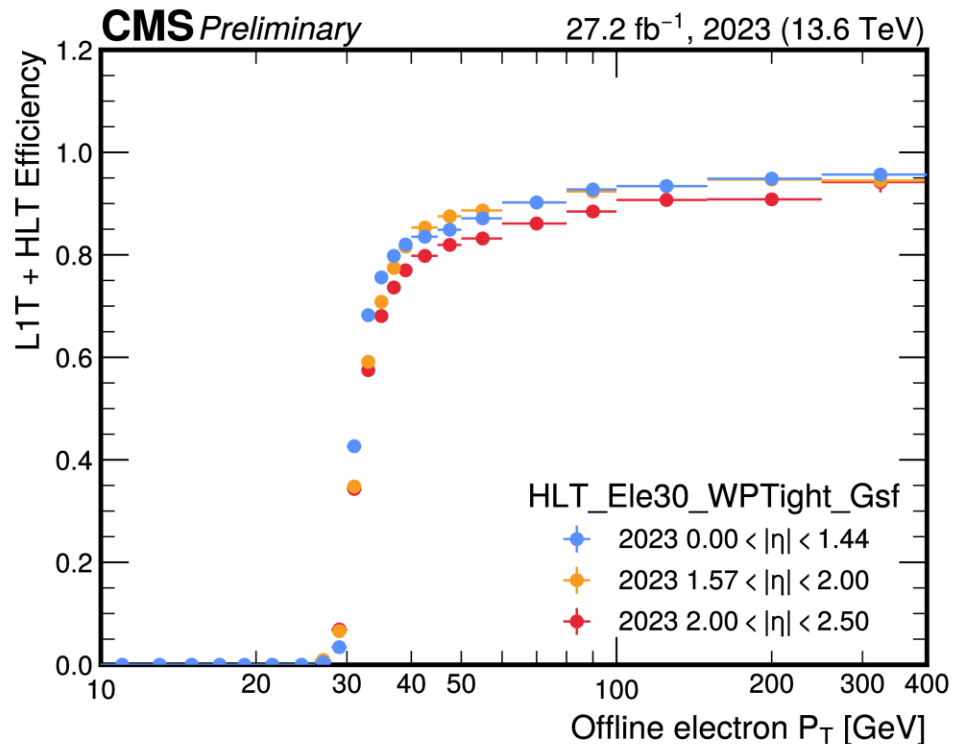
- Sharp efficiency turn-on at the p_T thresholds
- Efficiencies as high as **95–100%** up to the **TeV** scale

[CMS-DP-2024-122](#)

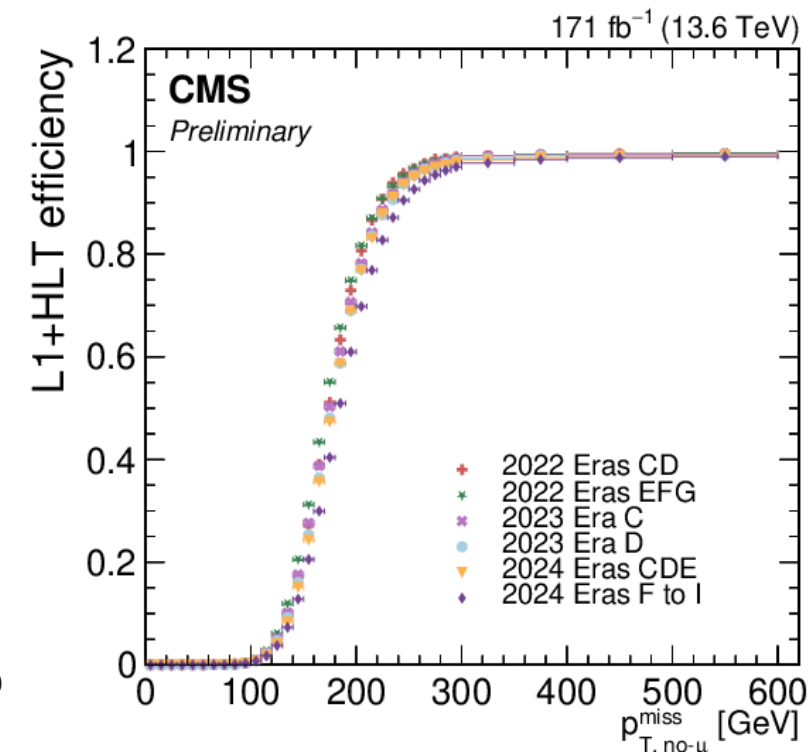
muons



electrons



Missing p_T



- Combined L1 + HLT efficiencies above 90–95% at plateau

[CMS-DP-2025-014](#)

[CMS-DP-2024-041](#)

[CMS-DP-2025-007](#)

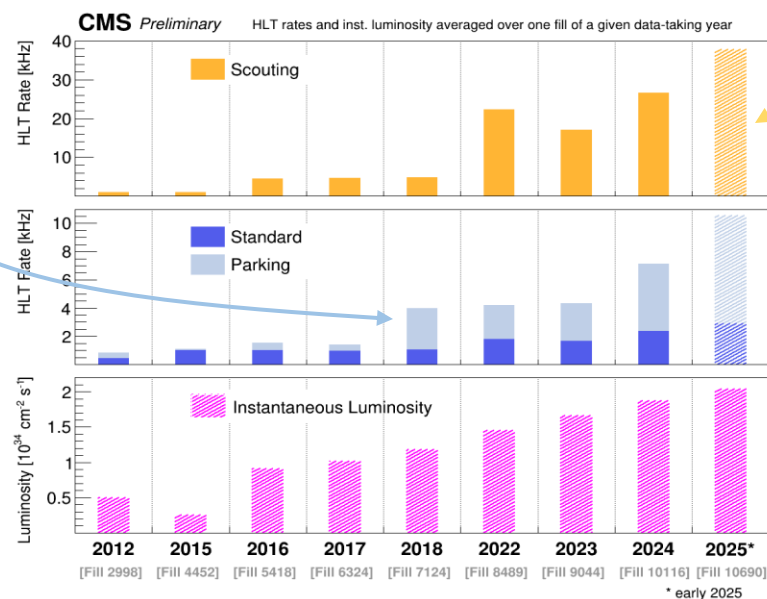
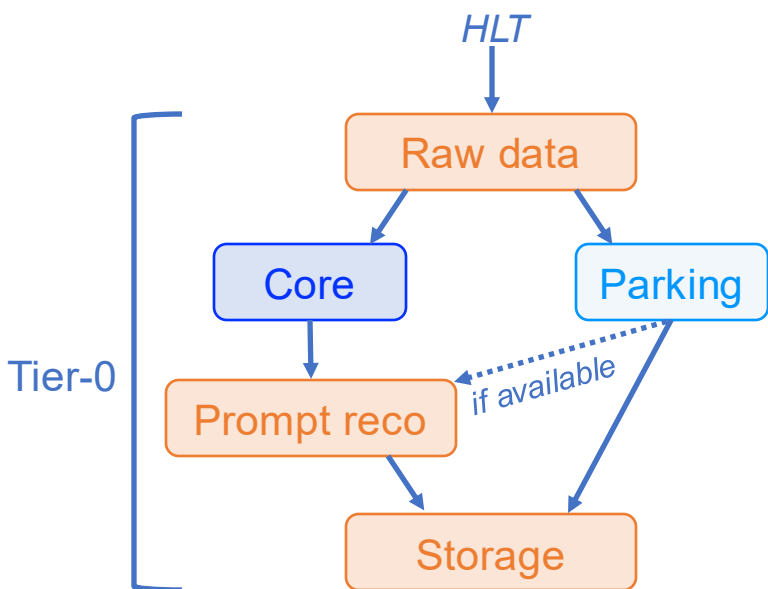
- Two strategies to circumvent standard rate limitations — since 2012!

Data Parking

- High-rate stream with full event content (~1 MB/event)
- Offline reconstruction delayed until computing resources become available
 - Always promptly reconstructed so far in Run-3

HLT Scouting

- Very-high-rate stream with reduced event content
 - No offline reconstruction, only trigger-level objects
 - ~10 kB/event → manageable bandwidth



In 2025 (*)

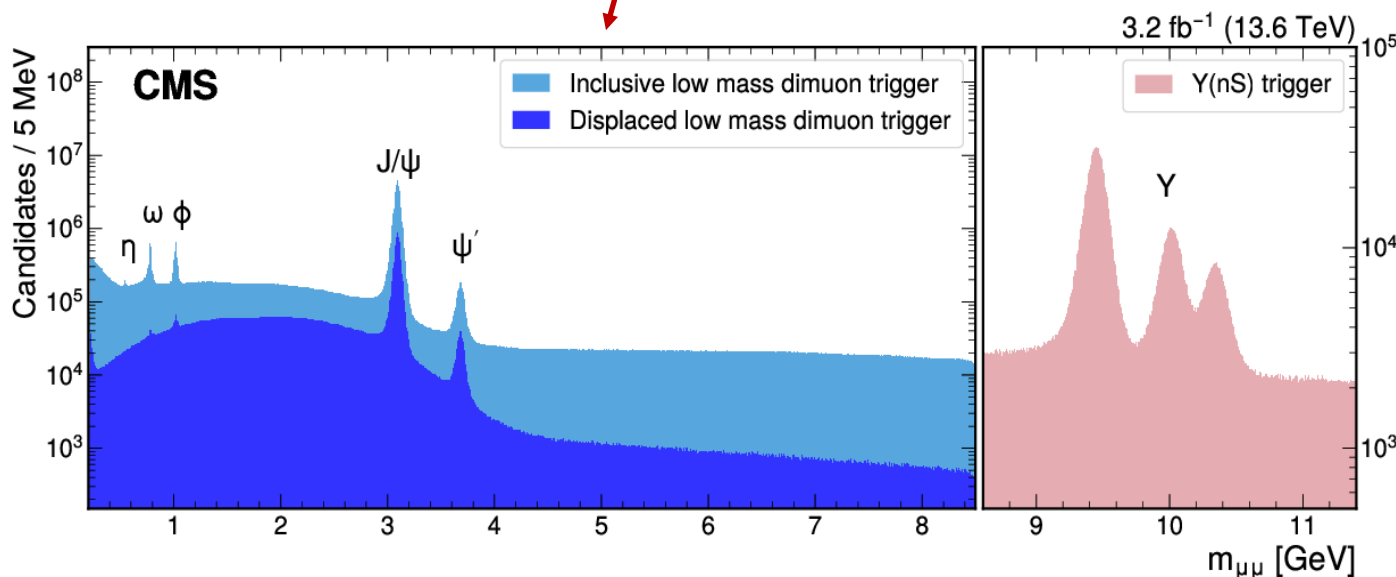
→ 30 kHz — 0.3 GB/s

→ 7 kHz — 8 GB/s

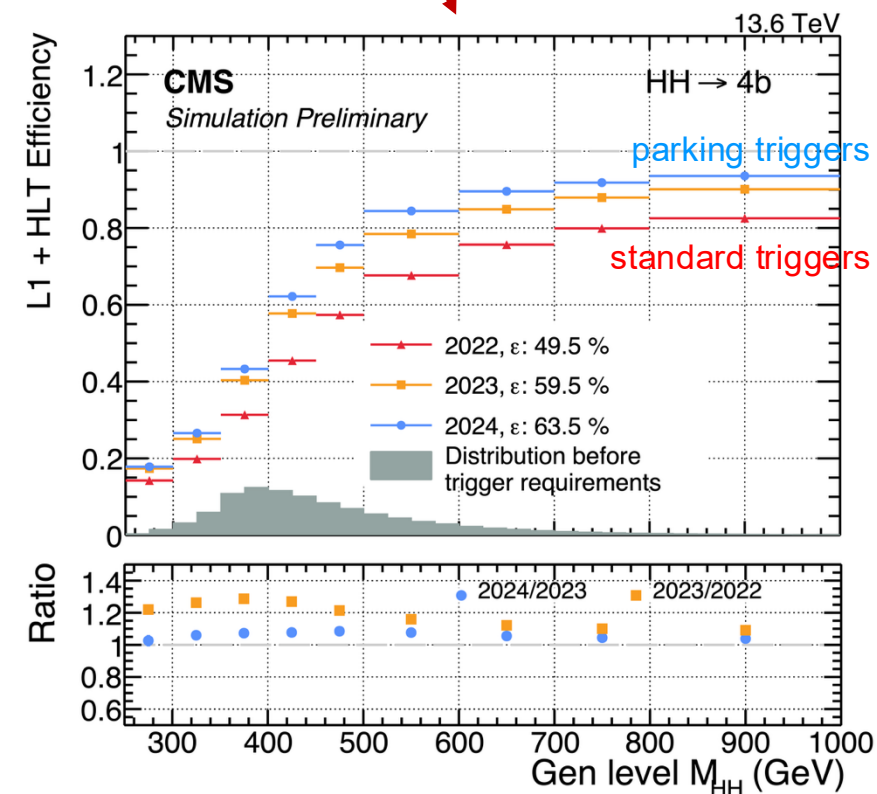
(*) *Ballpark numbers at peak luminosity*

[CMS-DP-2025-034](#)

- Stream content evolves following physics needs
 - Looser versions of core triggers or triggers for new final states
 - Currently triggers for B-physics, long-lived particles, vector boson fusion, and di-Higgs signature



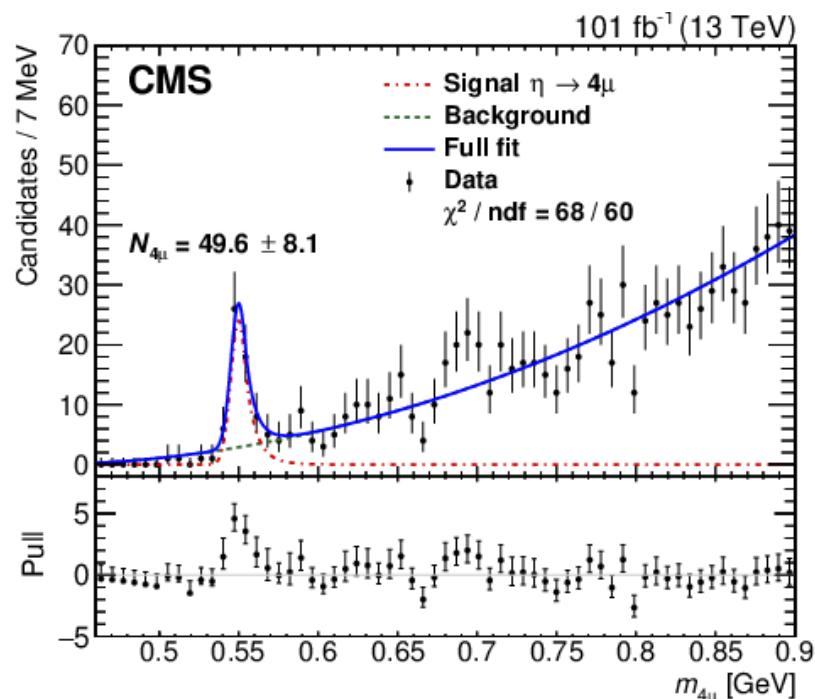
Parking di-muon triggers



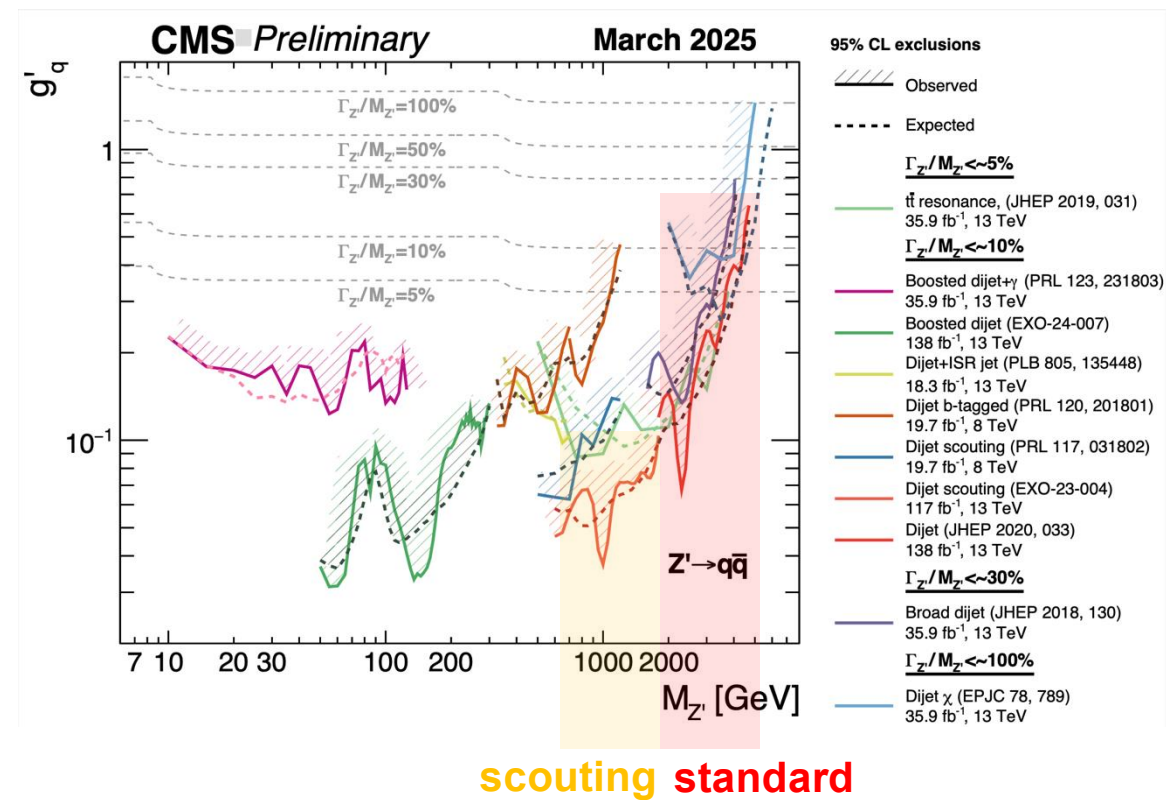
- Crucial for very low-mass searches or other “difficult” corners of phase space
 - B-physics, long-lived particles, di-jet resonances

$Z' \rightarrow 2 \text{ jets}$

First observation of $\eta \rightarrow 4\mu$ decay

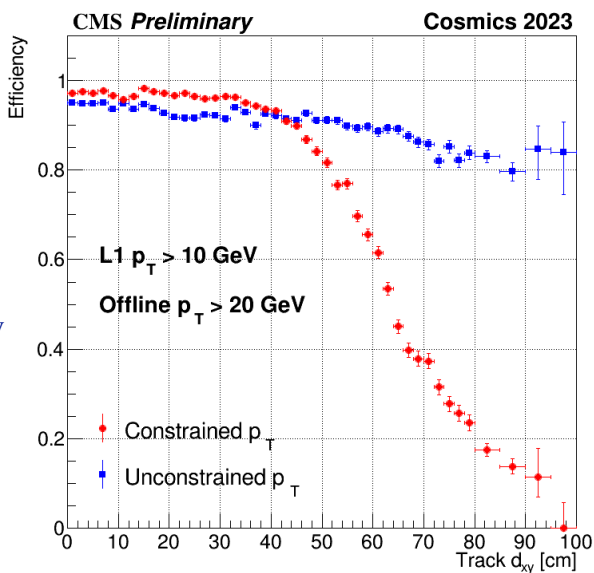
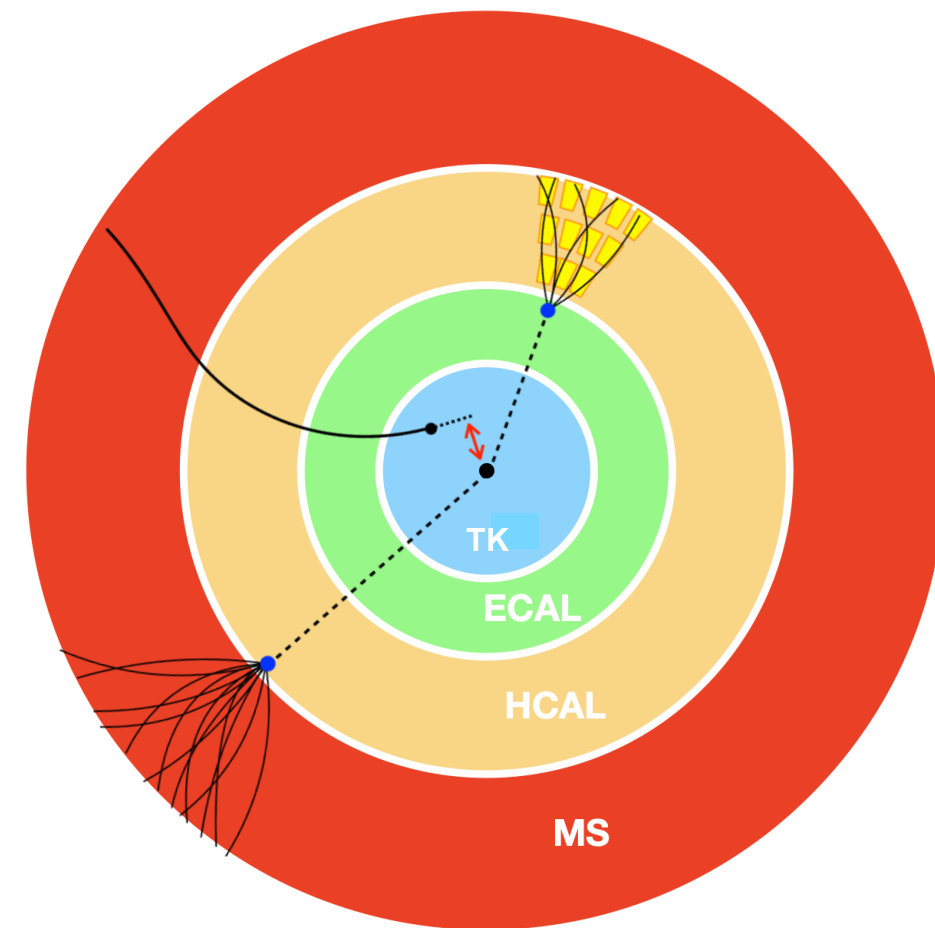


[Phys. Rev. Lett. 131 \(2023\) 091903](https://arxiv.org/abs/2303.10903)



<https://twiki.cern.ch/twiki/bin/view/CMSPublic/SummaryPlotsEXO13TeV>

- Since Run-2, CMS has established a strong program of **LLP searches**
 - Big efforts both at L1 and HLT to develop dedicated triggers
- Some of the new **L1 strategies**
 - **Displaced muons** with no constraints to the beamline
 - Hadronic **showers** in the **muon chambers**
 - **HCAL timing and depth** (with upgraded electronics)



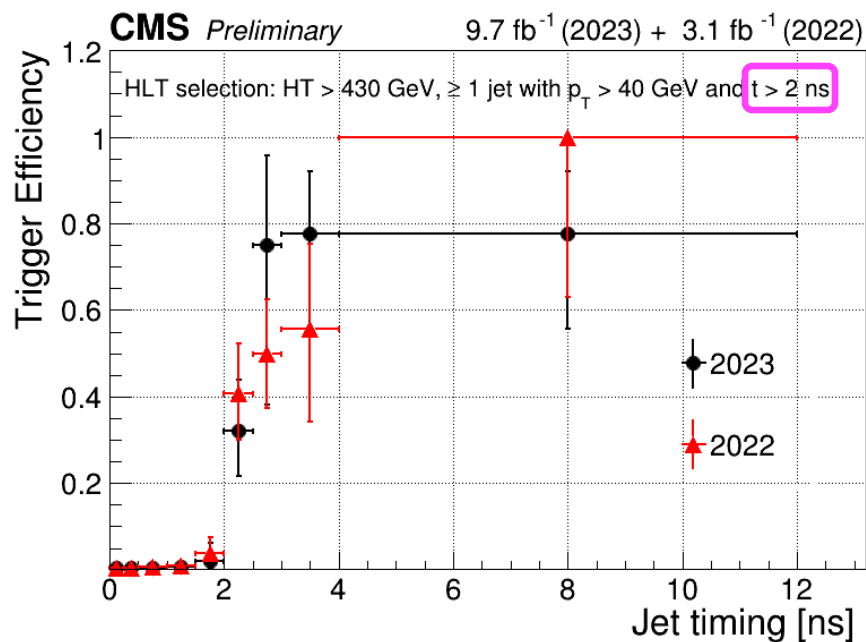
displaced L1 muon

standard L1 muon

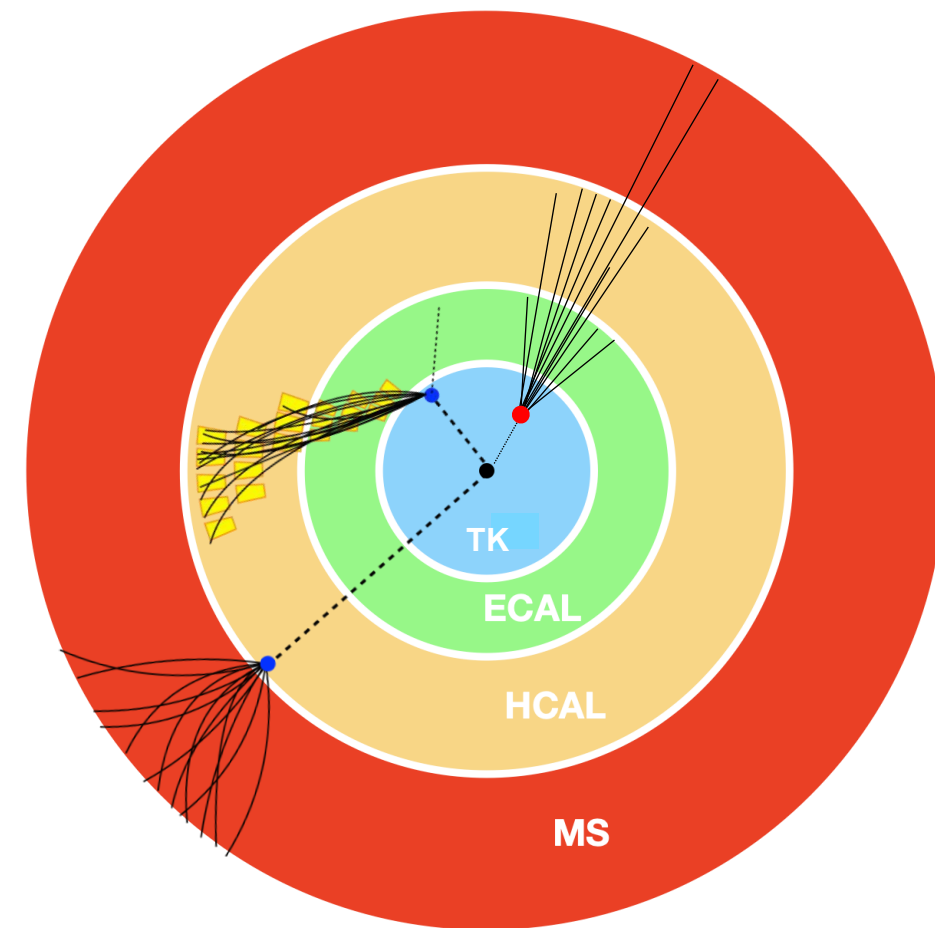
Muon efficiency vs d_{xy}

[CMS-DP-2023-056](#)

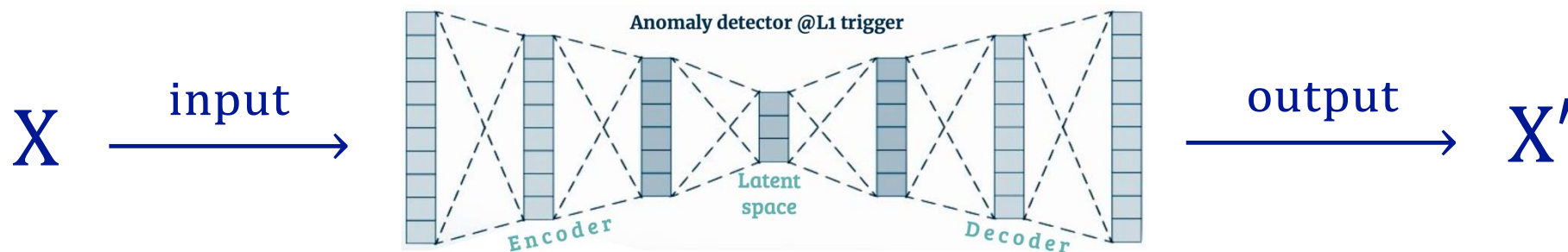
- Some of the **HLT** strategies
 - **Displaced jets** using non-prompt tracker tracks
 - **Delayed jets** using ECAL time measurements
 - **Muon showers** from hit clustering



*Delayed jet efficiency
vs
jet timing*

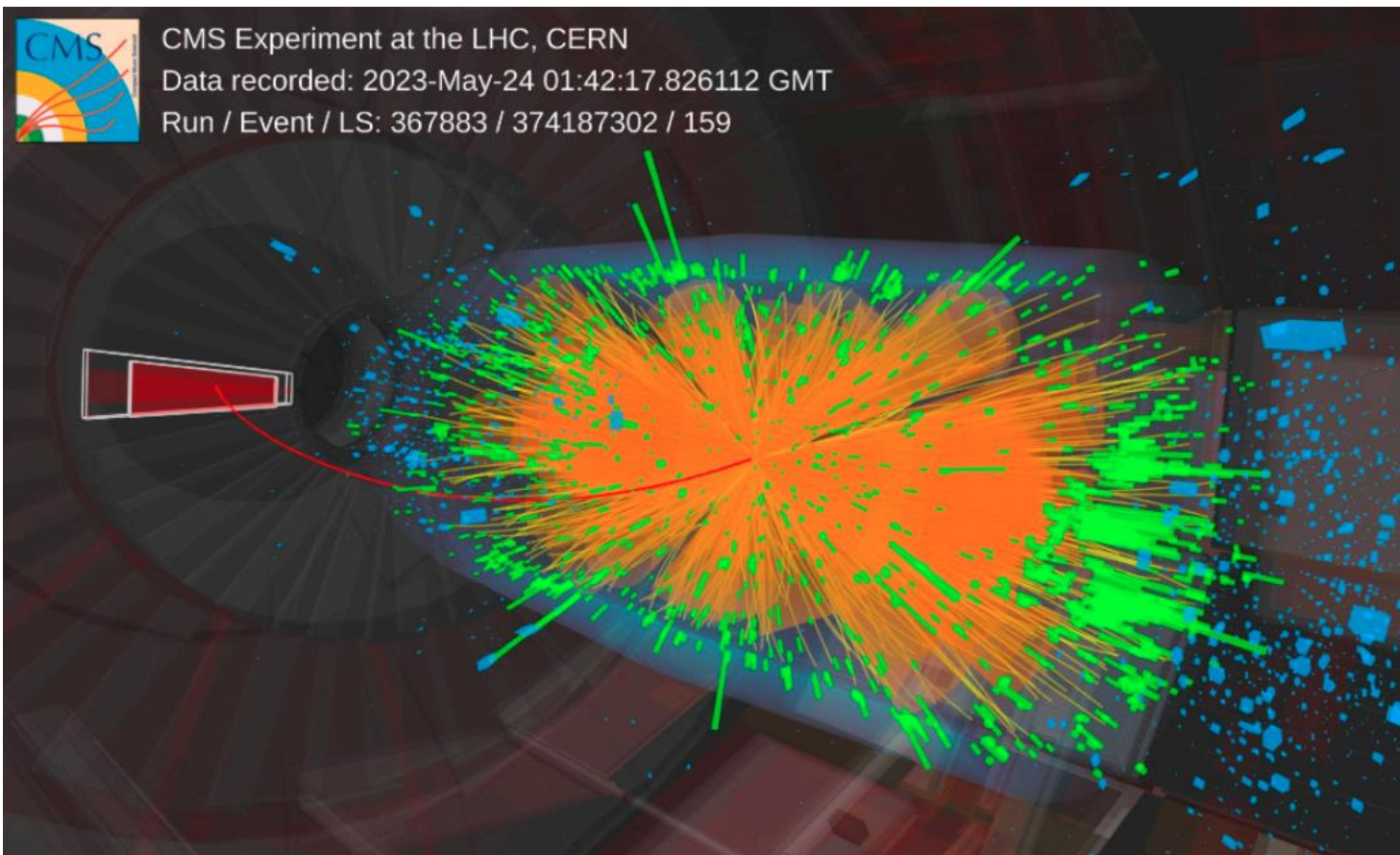


- Use **ML** to look for anything “unusual” in the collisions
 - Train an **autoencoder** on typical proton-proton collisions (“zero-bias” events)
 - Trigger on **outliers** using the **loss function** as metric: $\mathcal{L} = \|X - X'\|$



- Two complementary approaches at **L1**
 - **Anomaly eXtraction Online Level-1 Trigger aLgorithm**
 - Input: (p_T, η, ϕ) of muons, e/γ , jets, and MET
 - **Calorimeter Image Convolutional Anomaly Detection Algorithm**
 - Input: η - ϕ map of calorimeter deposits in image format

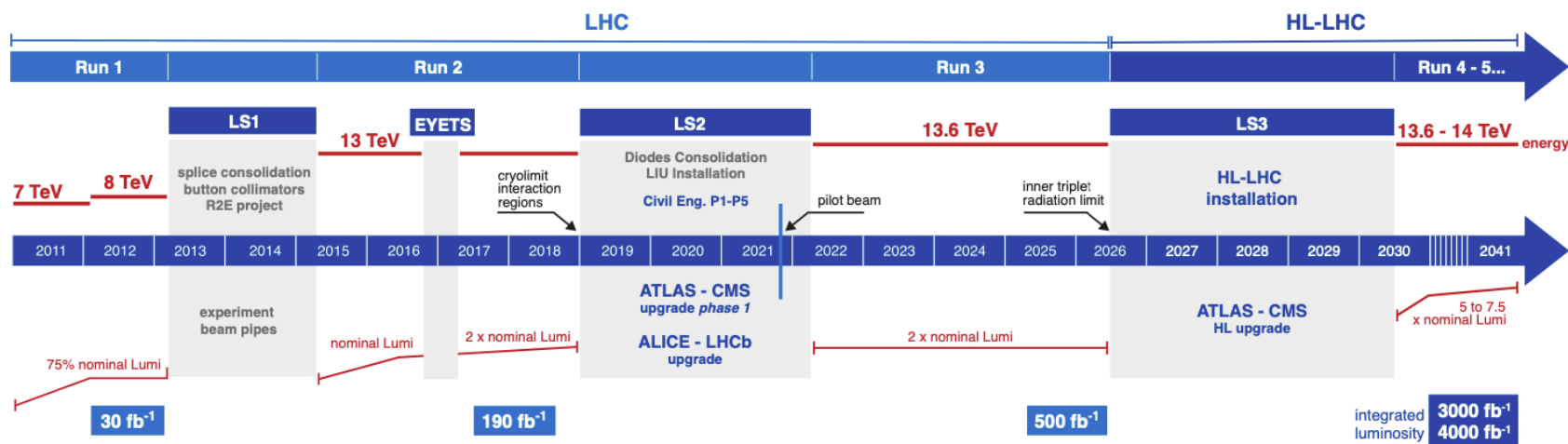




- CMS event (May 2023) selected exclusively by **AXOL1TL**
- Features:
 - 12 jets, 11 with $p_T > 20$ GeV
 - 1 muon with $p_T = 3$ GeV
 - 75 primary vertices (with $\langle \text{PU} \rangle \sim 64$)

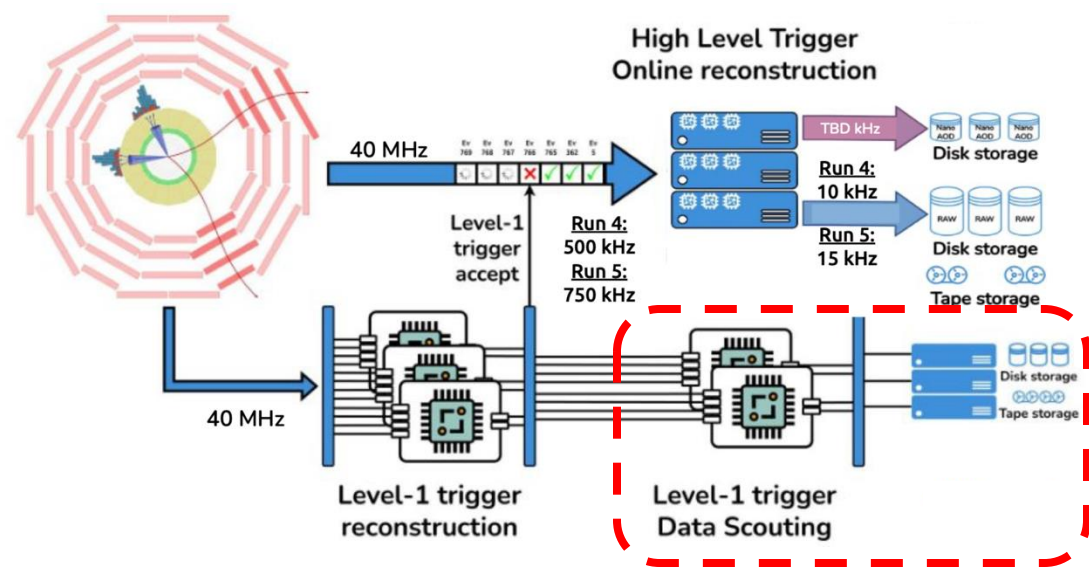
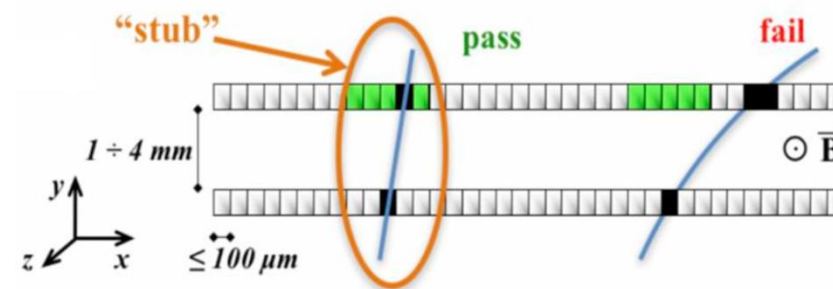
[CMS-DP-2023-079](#)

- The **High-Luminosity LHC** will present significantly harsher conditions

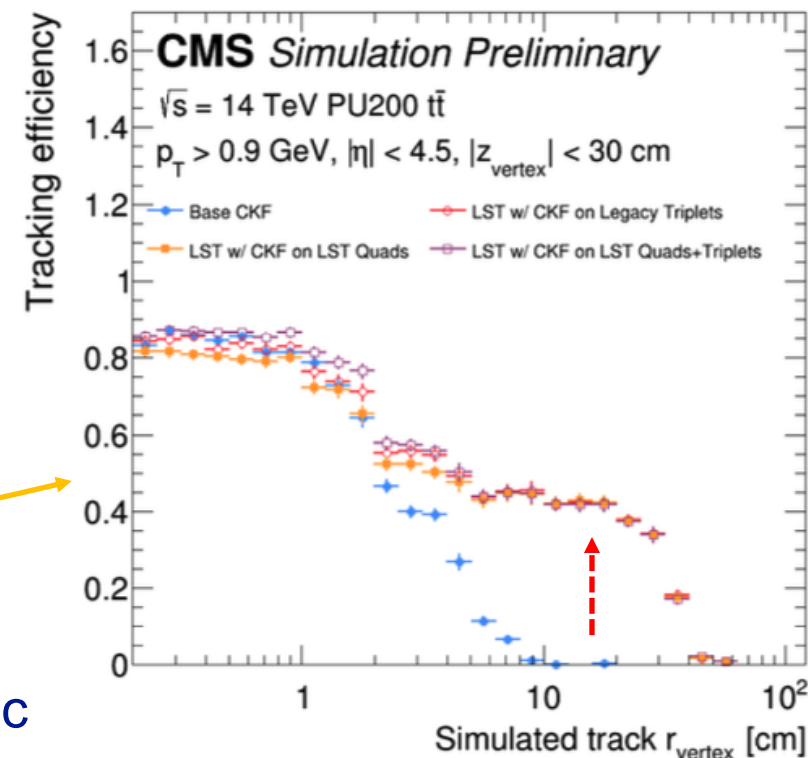


	HL-LHC (Phase-2)	
	Run-4	Run-5
Peak \langle PU \rangle	140	200
L1 accept rate	500 kHz	750 kHz
HLT accept rate (core)	5 kHz	7.5 kHz
DAQ throughput	24 GB/s	51 GB/s

- Upgrade **trigger electronics** for all subdetectors with state-of-the-art FPGAs
- **Outer tracker (OT)** will be included in the L1T
 - L1T tracking for prompt particles with $p_T > 2 \text{ GeV}$
 - Hit-pair patterns allow for a **10× rate reduction**
- **L1 Data Scouting @ 40 MHz**
 - Save all collision events before the L1 decision → **triggerless data!**
 - Store **L1 objects** only → limited resolution
 - Opportunity to peek at regions of phase space otherwise inaccessible!
 - A **demonstrator** has been running since 2024

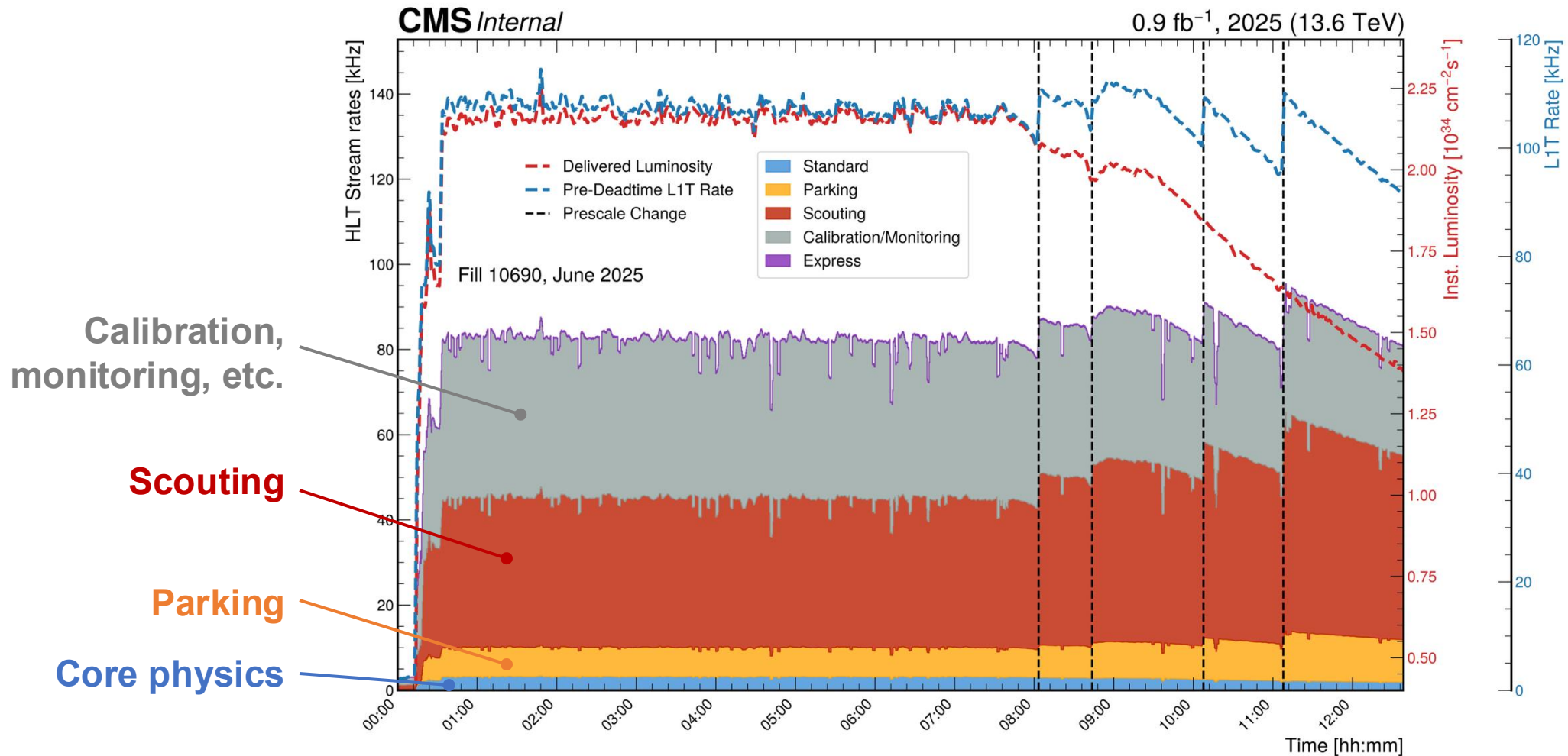


- HLT output rate target: 7.5 kHz
- Use of heterogeneous computing
- Tracking particularly challenging at $\langle \text{PU} \rangle \sim 140\text{--}200$
 - Fully running on GPUs (parallelization)
 - Use the OT double-sided sensors to reduce the combinatorics (“Line-Segment Tracking” or LST)
 - Extend acceptance of displaced tracks
- Major upgrades of the DAQ system to handle the larger data traffic
 - E.g., increase the number of event builder nodes from ~ 60 to ~ 200

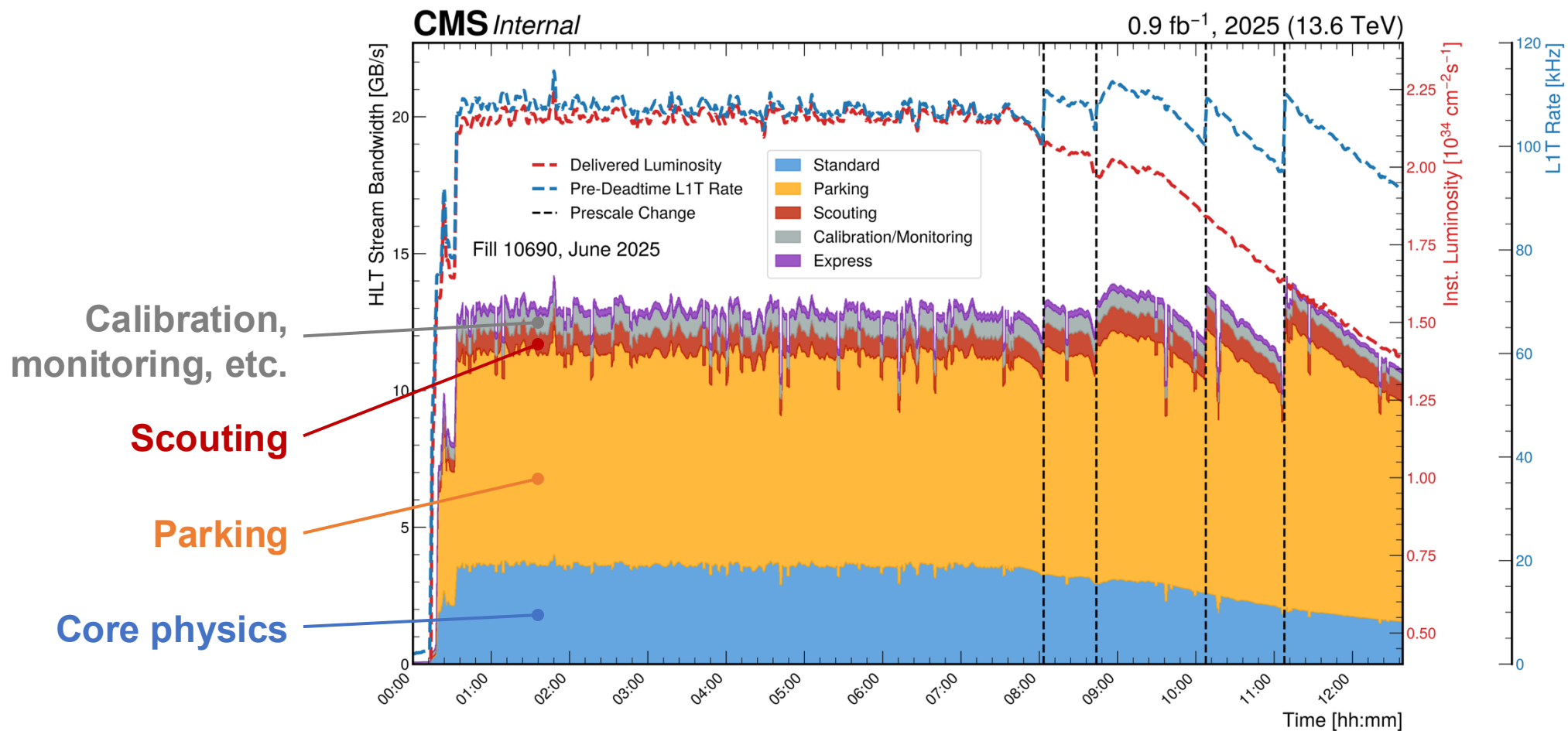


- CMS Trigger + DAQ systems successfully operating in Run-3 with upgraded L1 and HLT capabilities
 - Significant progress in L1 algorithms, HLT software, and GPU deployment
- Phase-2 upgrades will enable CMS to thrive even under the harsh HL-LHC conditions
 - L1 tracking, heterogeneous computing, and potentiated DAQ architecture are pillars of the future system
 - Ongoing development and validation effort to prepare CMS for the HL-LHC start in 2030!

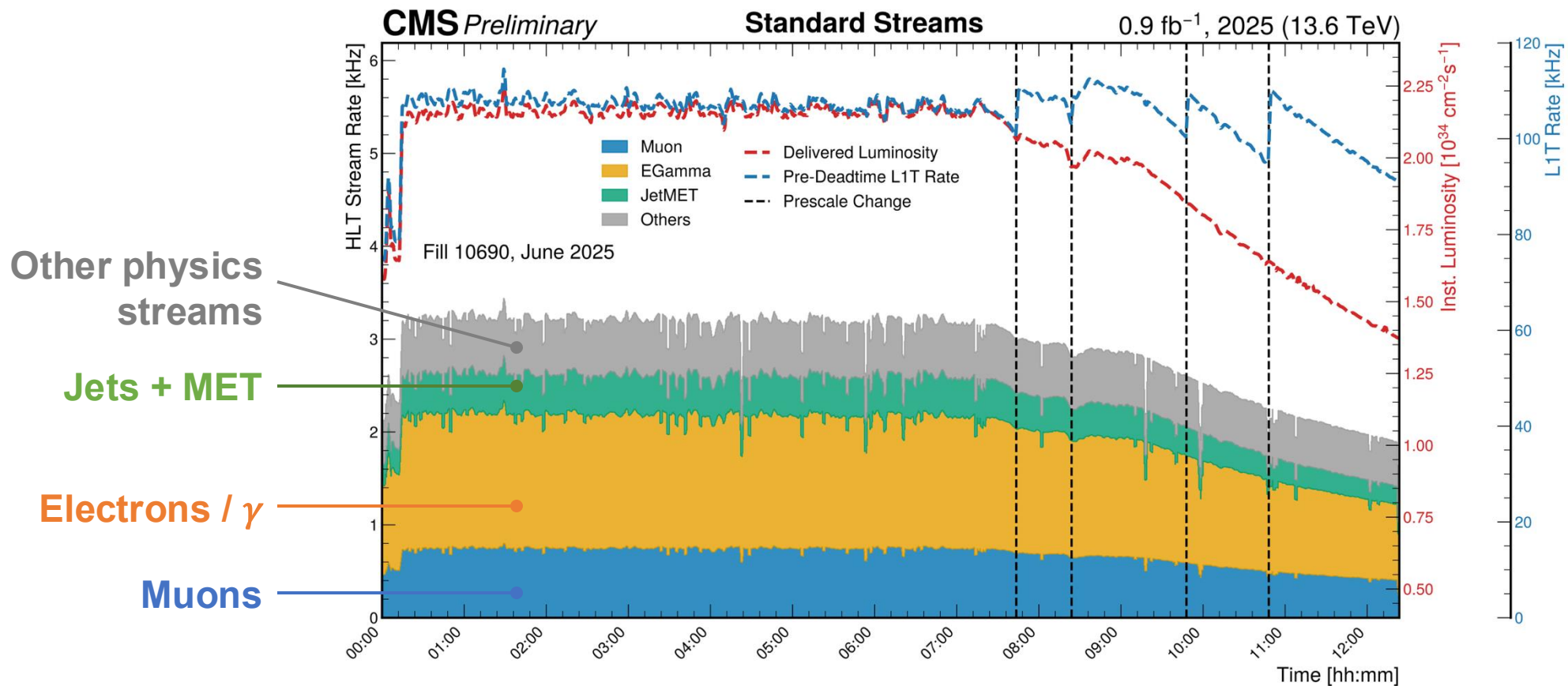
Backup



Typical 2025 LHC fill: stream bandwidth



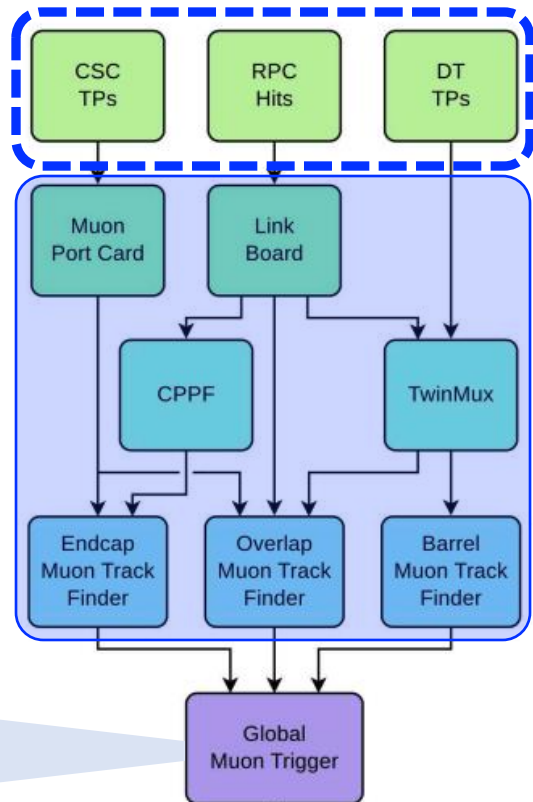
Typical 2025 LHC fill: physics streams



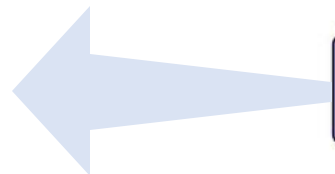
JINST 19 (2024) P05064

Muon chambers

Combine signals from chambers
 Reconstruct muon tracks
 Assign p_T , η , ϕ , muon quality

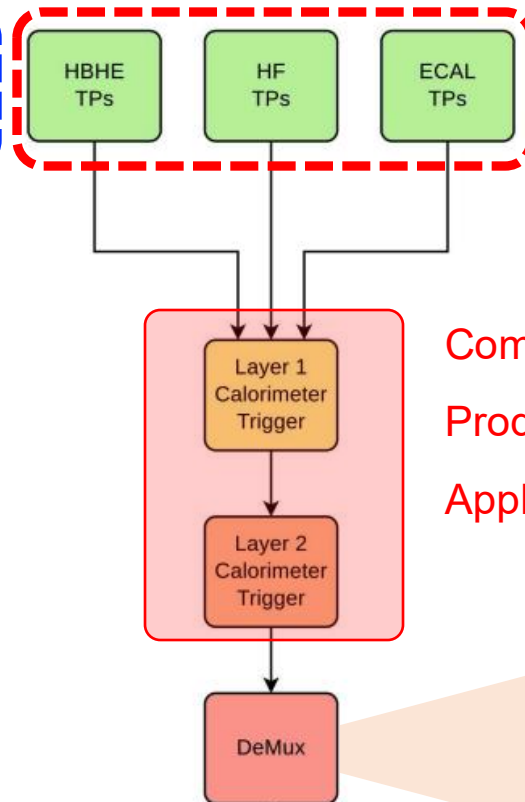


“prompt” muons
 (constrained to beamline)
“displaced” muons
 (no constraints to beamline)

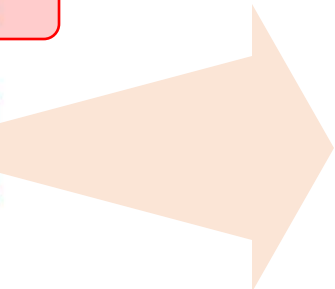


Calorimeters

Combine deposits from calorimeters
 Produce “calo-object” candidates
 Apply calibrations and corrections



electrons / photons
jets
 τ
MET, energy sums



Trigger decision

