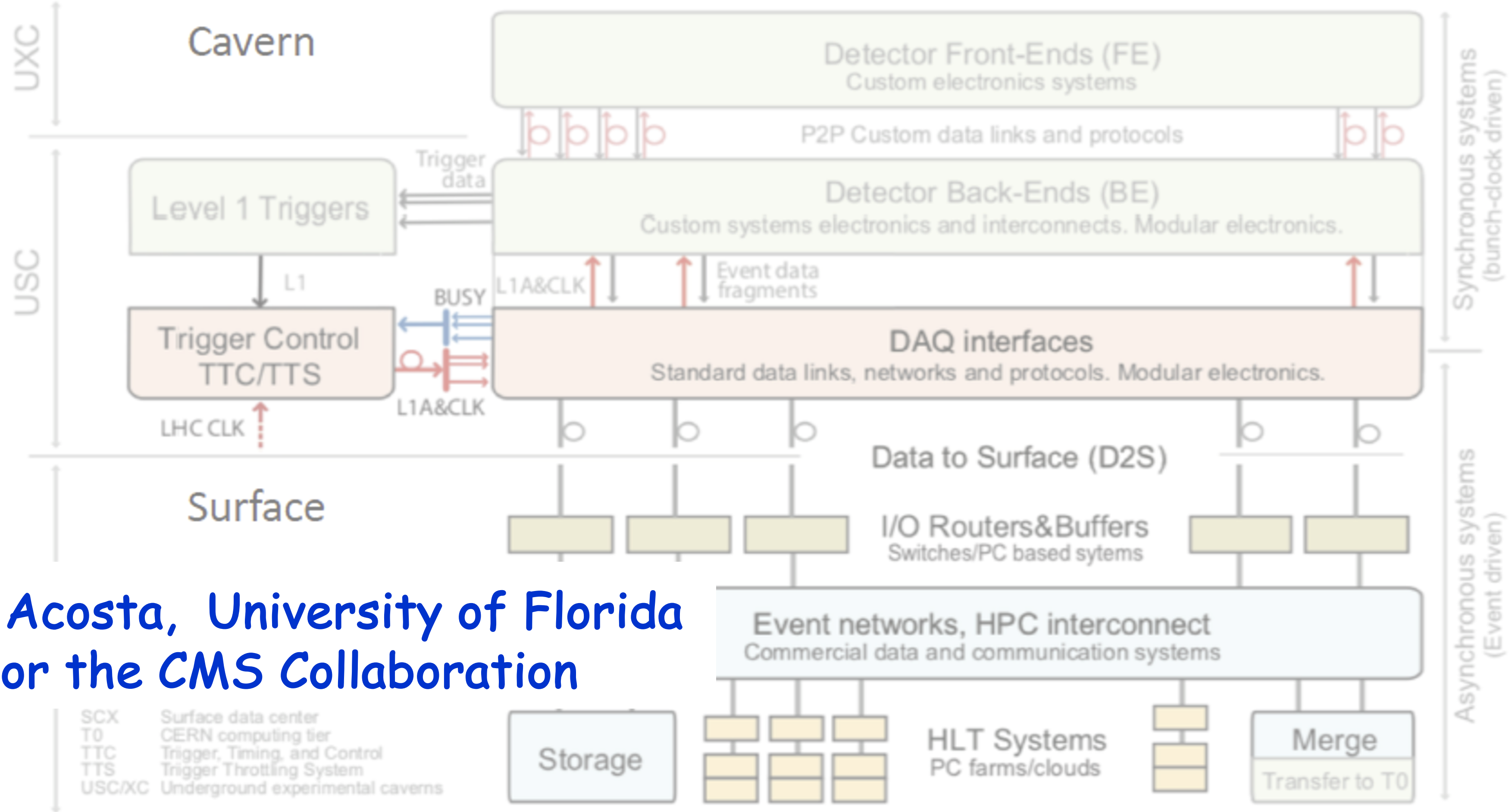
The CMS Trigger System



Darin Acosta, University of Florida
For the CMS Collaboration

General Outline

- Quick overview of a trigger
- CMS Trigger architecture
- Level-1 Trigger and performance
- High Level Trigger and performance
- Trigger menus
- Data parking
- Data scouting
- HLT processing time
- Summary

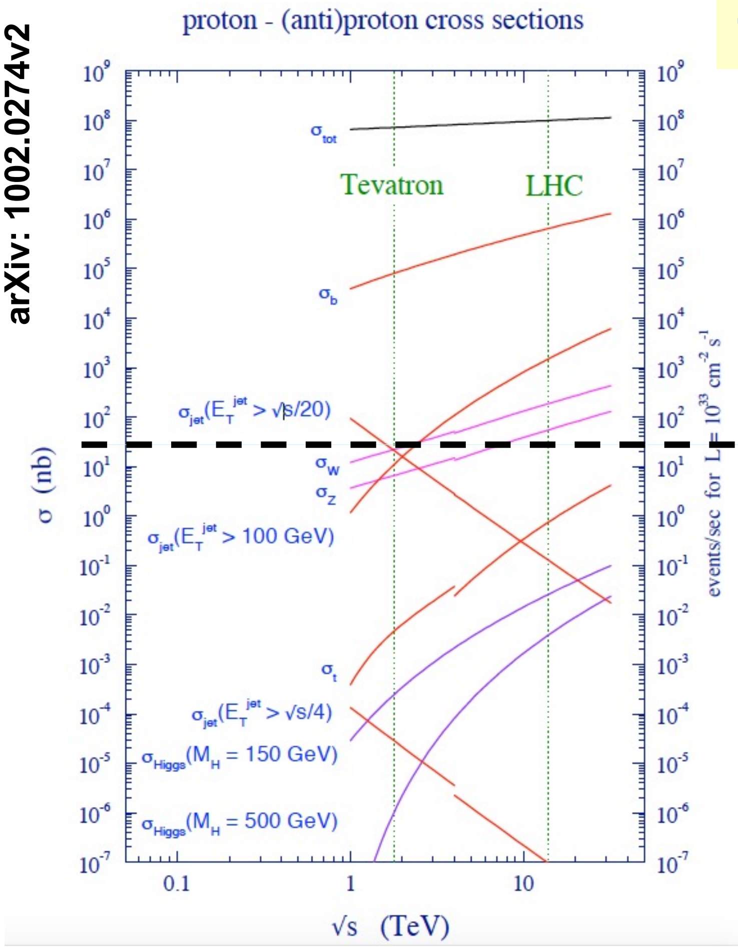
Data Acquisition and Trigger System

- The experiments must selectively read out their data using a real-time data selection system known as the “trigger”
- Why? Too much data to continuously stream to disk for storage and/or computer processing reasons
    - ❑ The LHC collides proton bunches at ~30 MHz with 40 MHz xing rate
    - ❑ An LHC experiment could generate ~50 terabytes per second!
- It is really the *first step of a data analysis*
  - ➢ An online data filtering system: select the collisions you want most
  - ➢ But irreversible – you can’t go back to data you did not record
    - ❑ e.g. Throw away 99.998% of all LHC crossings…
      - Keep only ~1000 Hz out of 40 MHz beam crossing rate
    - ❑ But, don’t throw out the baby (Higgs) with the bath water!

Note, that there is another approach to keep most events, but only a selective fraction of data per event (more later)

LHC Cross Sections and Rates

arXiv: 1002.0274v2



for L =2 x 10^{34} Hz/cm^2

← **Total collision rate: 2 GHz** (at ~30 MHz beam xing)

← **b quark rate: 10 MHz**

← **W boson rate: 4 kHz**

Signals that look like high rate SM processes in the detector are difficult to trigger on (i.e. hadronic final states)

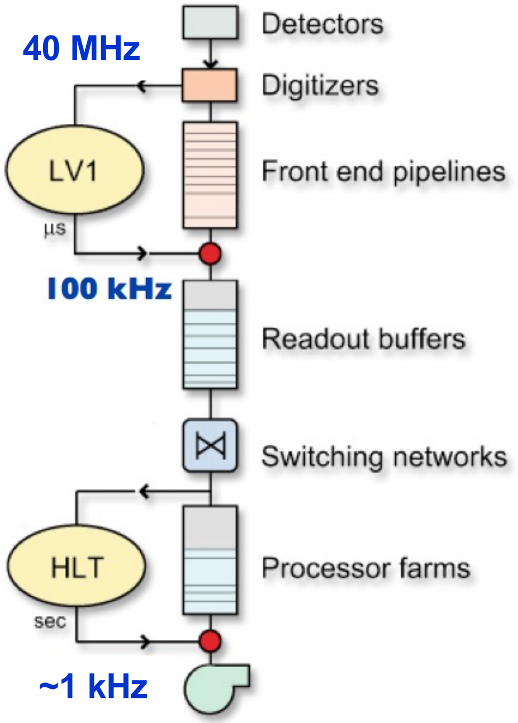
← **Top quark rate: 20 Hz**

← **Higgs boson rate: 1 Hz**

Keep

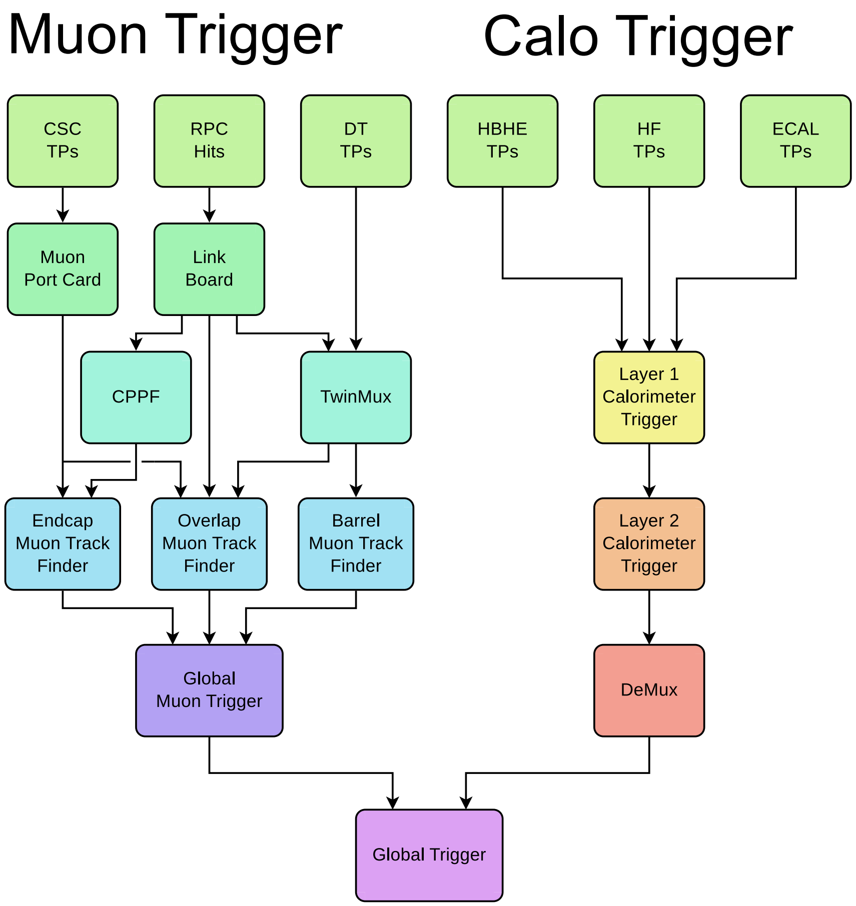
CMS Trigger Architecture

- Two levels:
  - Level-1: custom electronics to reduce the data from a crossing rate of 40 MHz (~30 MHz collisions) to no more than 100 kHz for the detector readout electronics, with only a 4 µs latency (buffer depth)
  - High Level Trigger (HLT): event filter farm comprised of commercial CPUs (and soon GPUs) running software to further reduce event rate to storage to an average of ~1 kHz (for LHC Run 2)
    - ☐ Order of 26k CPU cores



Current Level-1 Trigger Architecture

JINST 15, P10017 (2020)



- Muon Trigger
  - Redundancy of the three muon detection systems used early in the trigger processing chain to improve efficiency, resolution, and reduce rate
- Calorimeter Trigger
  - Improved tau ID, e/gamma isolation, and jet reconstruction with pileup subtraction over previous system
- Global Trigger: apply L1 menu
- No inner tracking (→Phase-2)

Level-1 Trigger Hardware

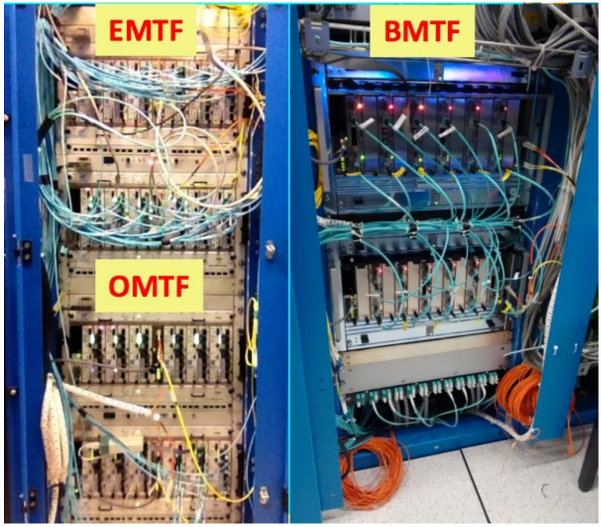
- Up to 10 Gbit/s optical links
- Large Field Programmable Gate Arrays (FPGAs)
  - Extensive logic and multiplier resources
- uTCA telecommunications infrastructure

✷ Flexibility:

- Modest revisions to data formats between boards is possible
  - Send extra data if useful
- FPGA firmware can be updated, with generally room for new trigger ideas
  - e.g. Kalman Filter algorithm, neural networks
  - Can be written in high level languages (e.g. Vivado HLS)

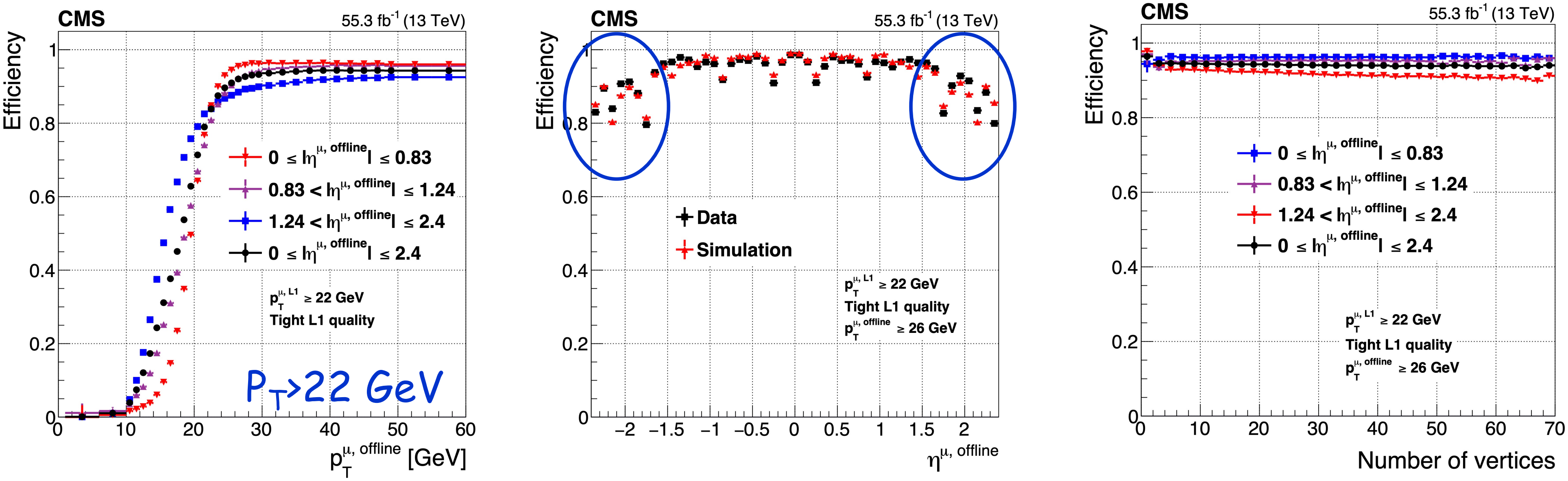
Calo trigger rack

Muon trigger racks



Level-1 Muon Trigger Performance

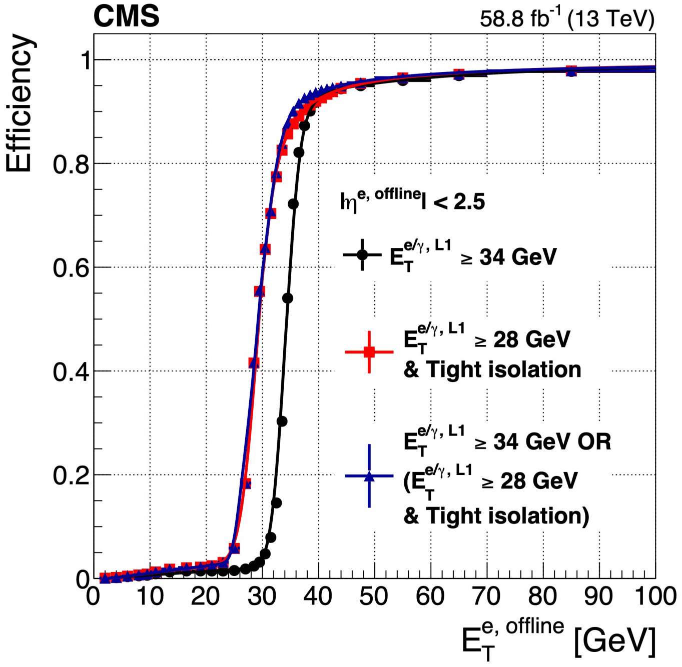
JINST 15, P10017 (2020)

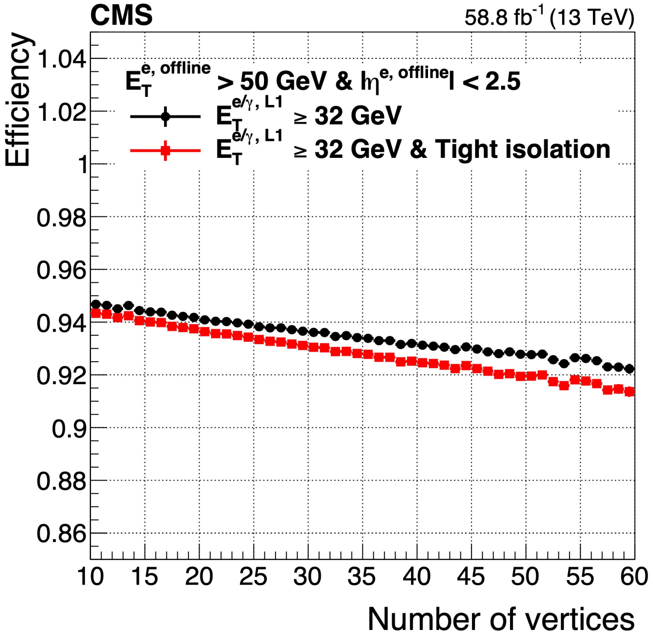


- Efficiency falls slightly in forward region
  - Less redundancy → should improve for Run 3 with new GEM detectors
- Efficiency flat with pileup

Level-1 Electron/Photon Trigger Performance

JINST 15, P10017 (2020)

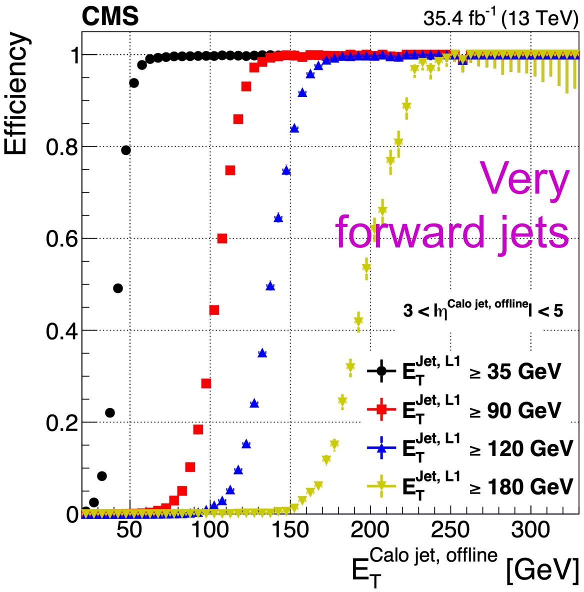


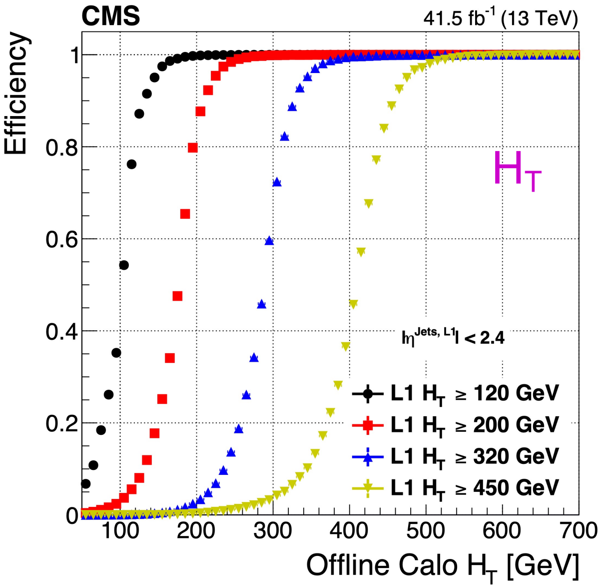


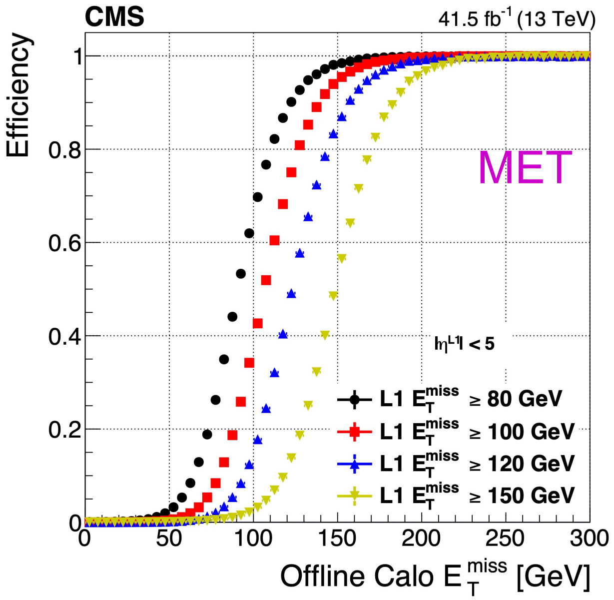
- Sharp efficiency turn-on
- Only modest dependence on pileup

Level-1 Jet, H_T, and MET Performance

JINST 15, P10017 (2020)





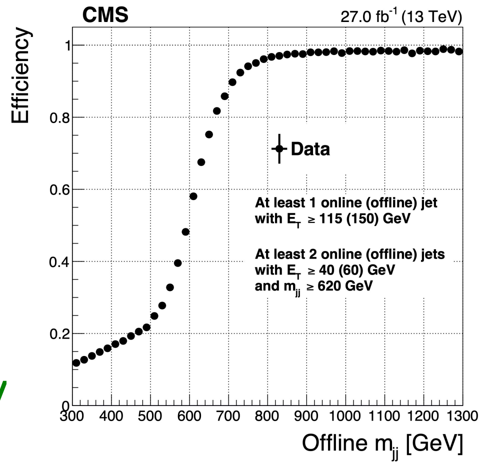


- Pileup mitigation applied

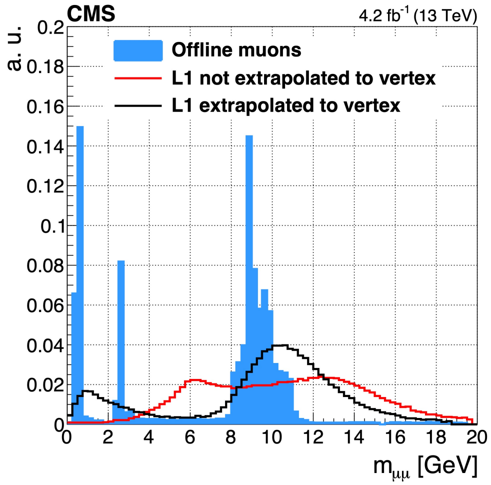
Dedicated Level-1 Triggers for Analyses

JINST 15, P10017 (2020)

- Possibilities with Global Trigger:
- VBF dijet trigger with invariant mass cut
  - At least two jets with E_T > 115 and E_T > 35GeV and at least one pair of jets with E_T > 35 GeV each, and an invariant mass greater than 620 GeV
- Low mass dimuon triggers with invariant mass cut (e.g. B trigger)
  - Apply lower P_T thresholds with dimuon invariant mass cut
  - Muon momenta extrapolated to vertex for mass calc. from standalone muon measurement



Other ideas are possible as well

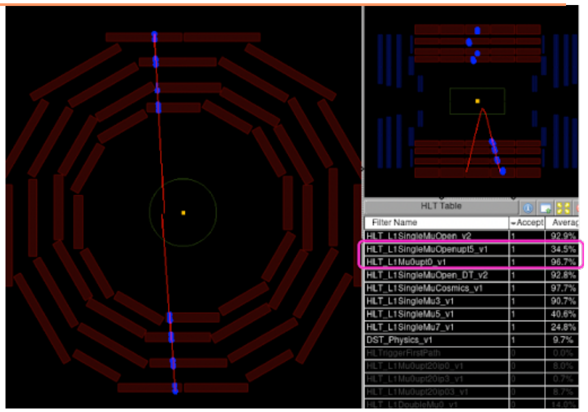


Run 3 Enhancements for Level-1 Muon Trigger

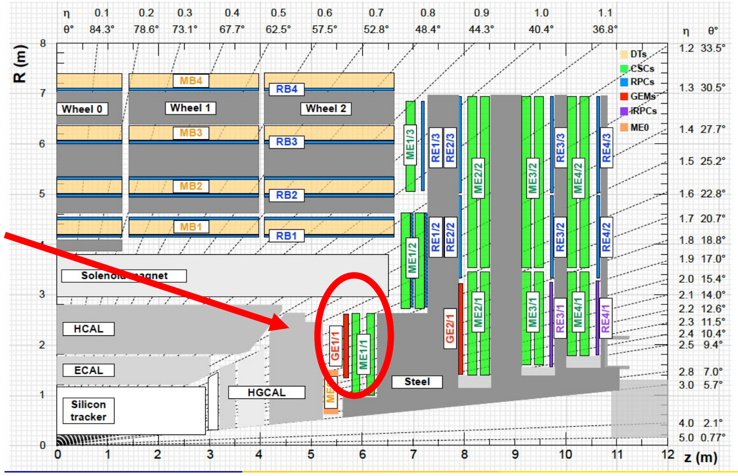
PoS Vol. 343 (2019) 139

https://indico.cern.ch/event/998052/

- Barrel:
  - Kalman filter for muon tracking has been incorporated into the FPGA logic of barrel muon trigger
    - Allows the possibility of triggering on LLPs with transverse displacements up to ~1m
    - Tested in recent cosmic muon data-taking
- Endcap:
  - Neural network for measuring displaced muons in endcap being ported to the FPGA logic of endcap muon trigger
    - Extends trigger on LLPs to endcaps
  - Innermost disk of Phase-2 GEM detectors being added to muon trigger
    - Should improve efficiency in the forward region: $1.6<|\eta|<2.1$
    - Under commissioning
- Potential to trigger on a single LLP leg



Filter Name	Accept	Averag
HLT_L1SingleMuOpen_v2	1	92.9%
HLT_L1SingleMuOpenupt5_v1	1	34.5%
HLT_L1Mu0upt0_v1	1	96.7%
HLT_L1SingleMuOpen_DT_v2	1	92.8%
HLT_L1SingleMuCosmics_v1	1	97.7%
HLT_L1SingleMu3_v1	1	90.7%
HLT_L1SingleMu5_v1	1	40.6%
HLT_L1SingleMu7_v1	1	24.8%
DST_Physics_v1	1	9.7%
HLTriggerFirstPath	0	0.0%
HLT_L1Mu0upt20ip0_v1	0	8.0%
HLT_L1Mu0upt20ip3_v1	0	0.7%
HLT_L1Mu0upt20ip03_v1	0	8.7%
HLT_L1DoubleMu0_v1	0	14.0%



The High-Level Trigger

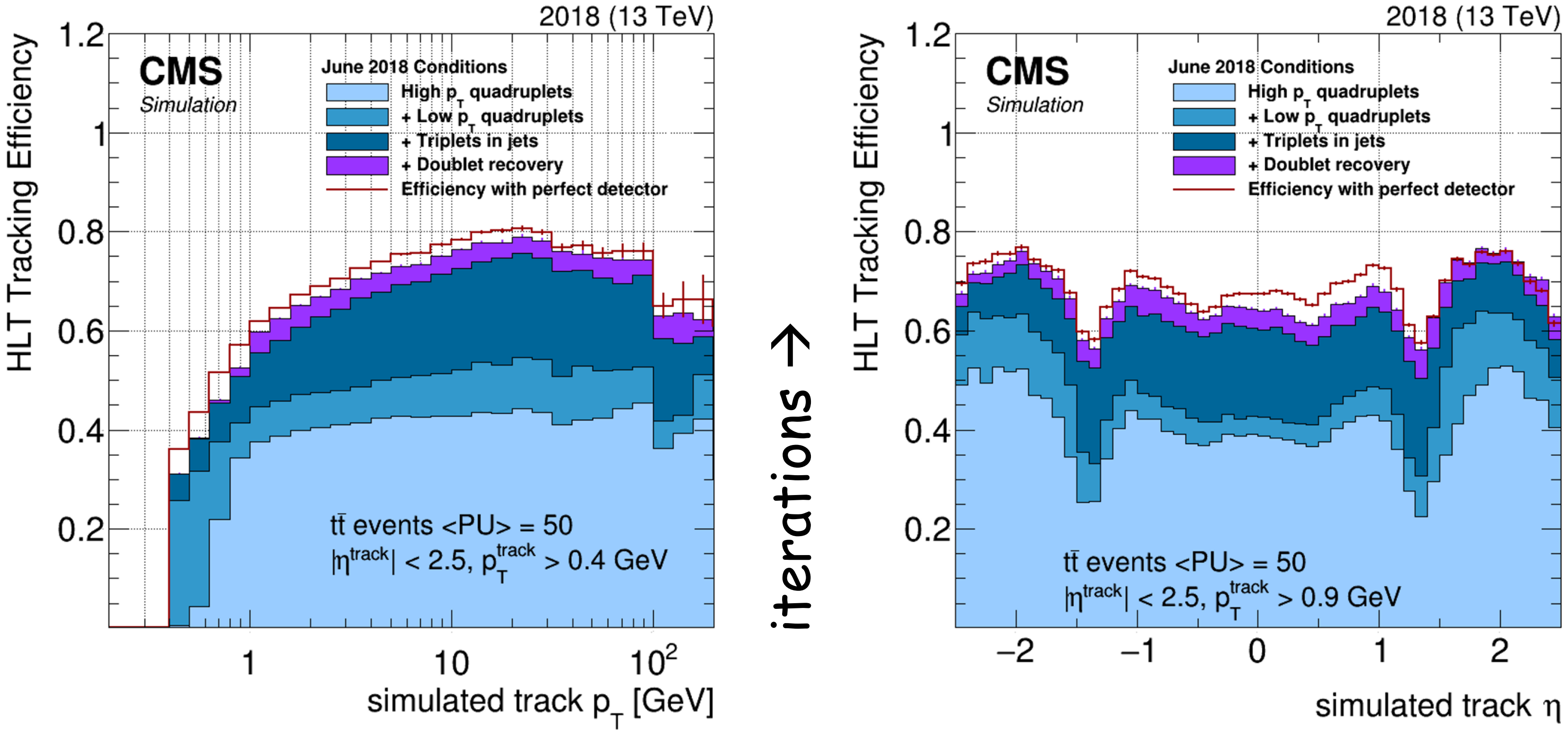
- Must reduce event rate from 100 kHz to ~1 kHz (LHC Run 2)
  - The output rate is set by offline computing data processing constraints and data storage limits
- But HLT starts with a data sample already enriched in physics!
  - Level-1 already applied a factor 400 background rejection, and still need to find another factor 100 reduction
- What else can HLT do?
  - Work with higher precision data than that used by Level-1
    - Uses same raw data as offline analyses
  - Include new detectors not used by L1
    - Silicon strip and pixel tracking
  - Employ more sophisticated algorithms
    - Particle flow, b-tagging
  - Use software very similar to that used offline

HLT Tracking

- Tracking in the HLT is performed in 3 (+1) iterations
  - Start with tight requirements for the track seeds, and loosen for each subsequent iteration
    - ❑ Hits in the tracking detectors already used in a track are removed at beginning of the next iteration
  - The first two iterations require the maximum of four consecutive hits in the pixel detector to seed the tracking (first high P_T, then low)
  - The third iteration relaxes the number of hits in the track seeds to three, restricted to the vicinity of jets identified from calorimeter information and the tracks from the two previous iterations
  - An additional iteration was added using pixel doublets for $p_T > 1.2$ GeV as protection against DC/DC failures as observed in 2017

HLT Tracking Efficiency

CMS-DP-2018-038

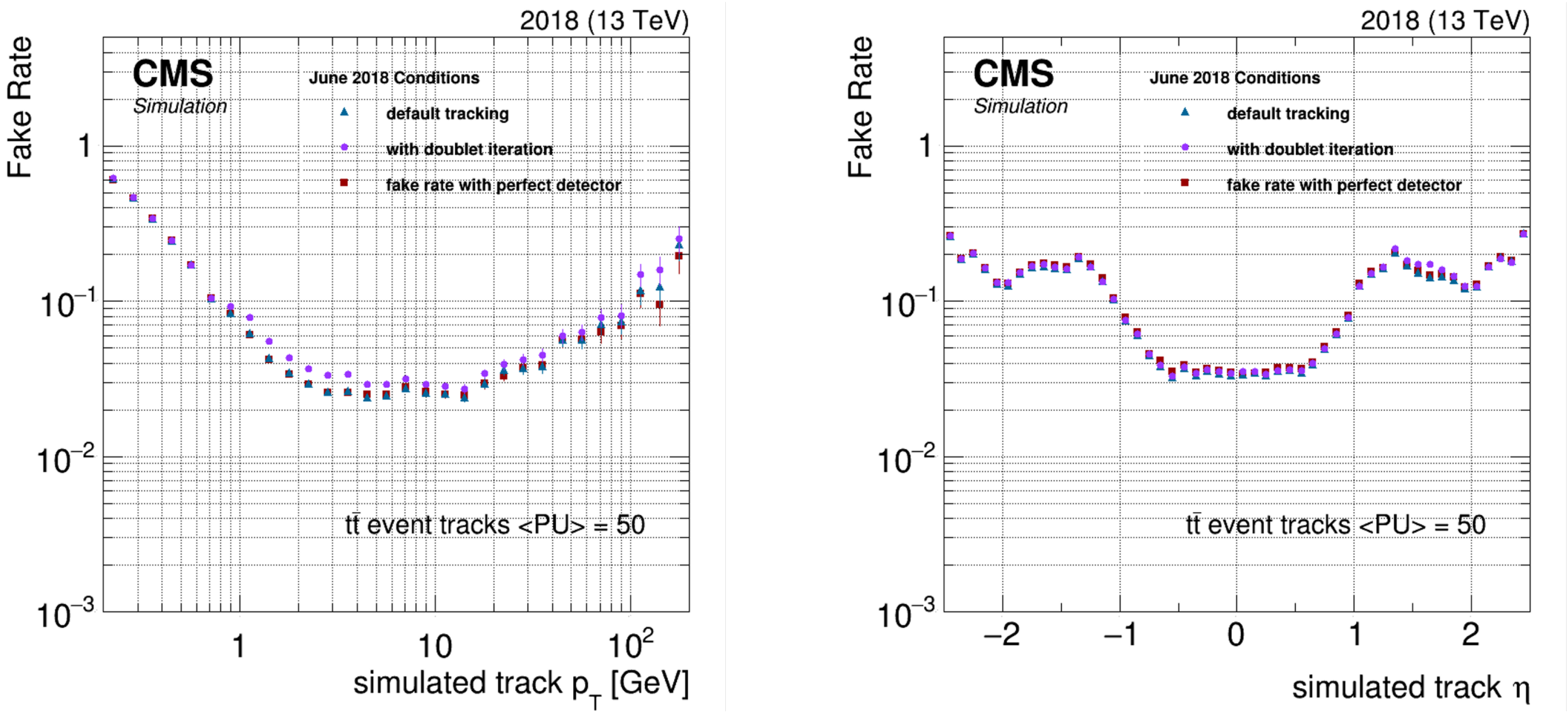


iterations →

- After all iterations, efficiency is close to that expected for perfect detector
- Efficient for tracks as low as $p_T = 0.4$ GeV

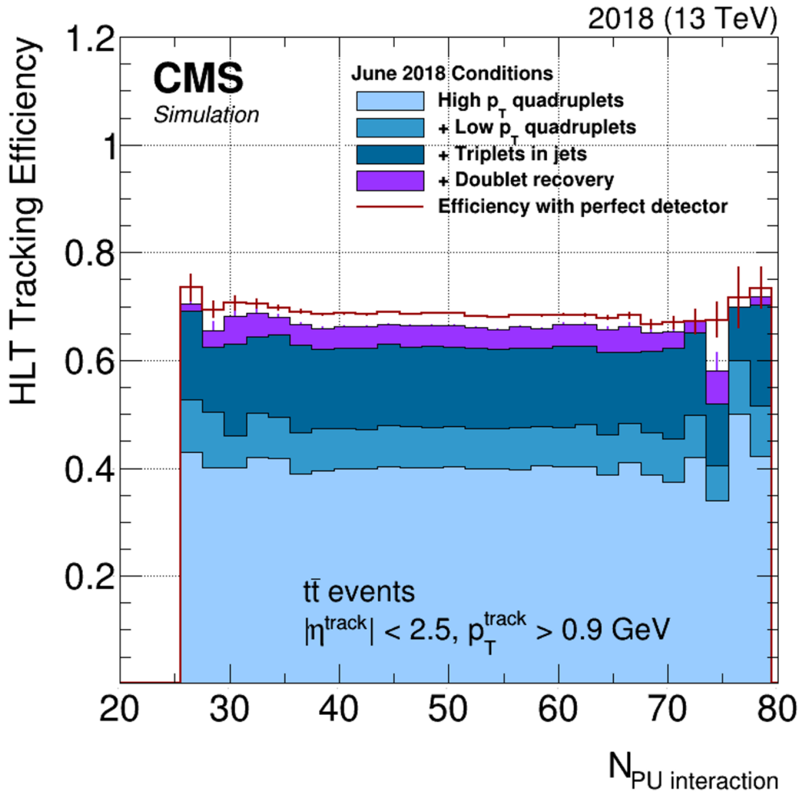
HLT Tracking Fake Rate

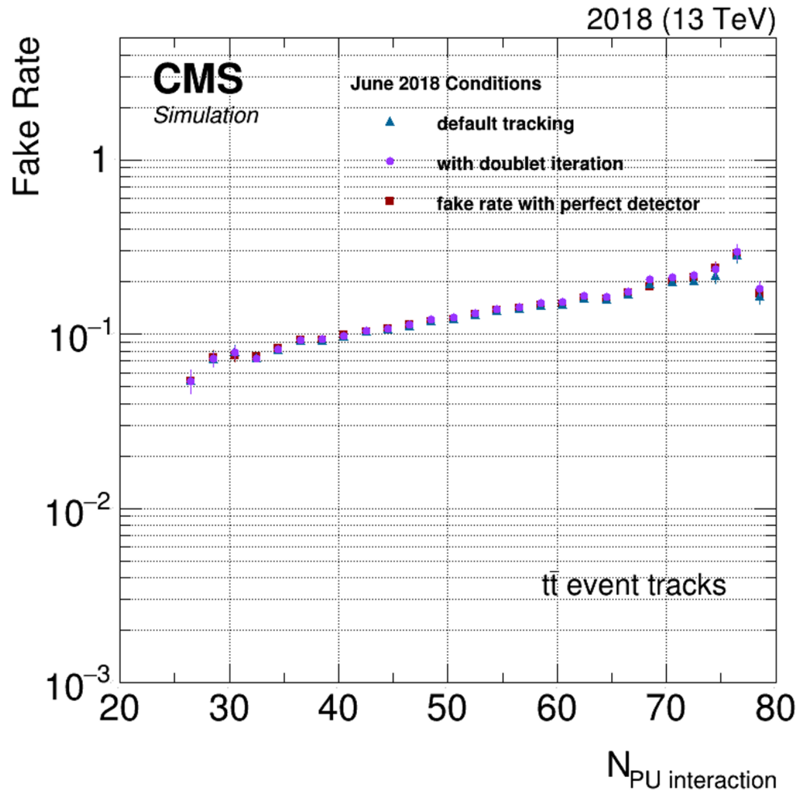
CMS-DP-2018-038



- Fake rate largest in forward regions, grows at low and high P_T

HLT Tracking Pileup Dependence CMS-DP-2018-038



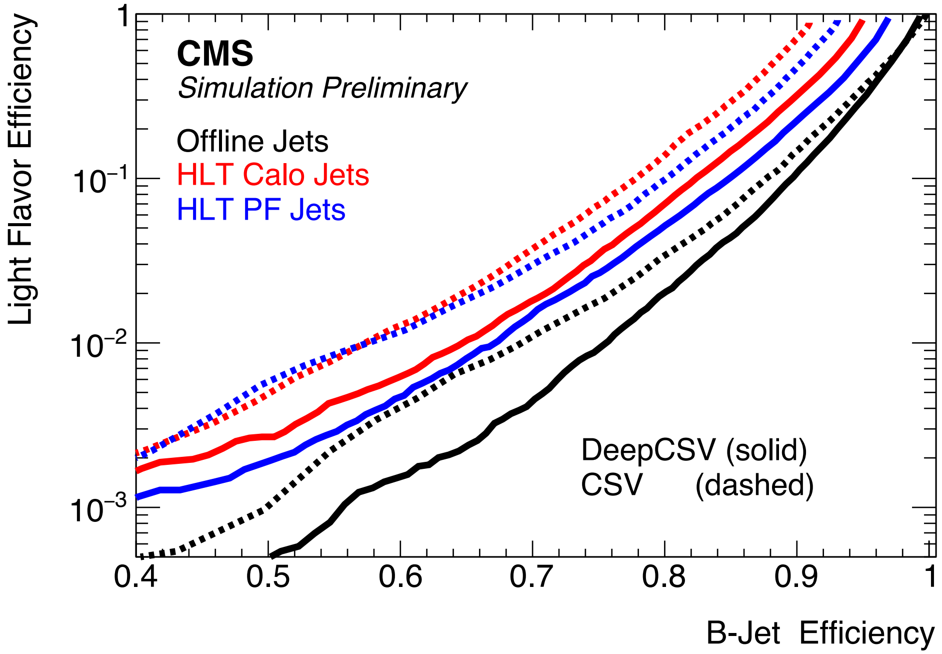


- Efficiency is stable with pileup, though fake rate grows slightly

B-Tagging

CMS-DP-2019-042

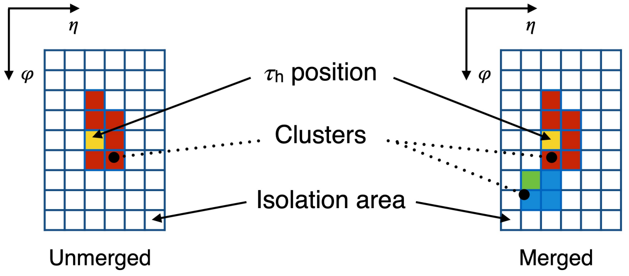
- Since 2017 a neural network classifier, DeepCSV, was used to identify b-jets (AK-4 jets)
- Improves b efficiency over previous CSV algorithm by 5-15% for fixed light flavor efficiency

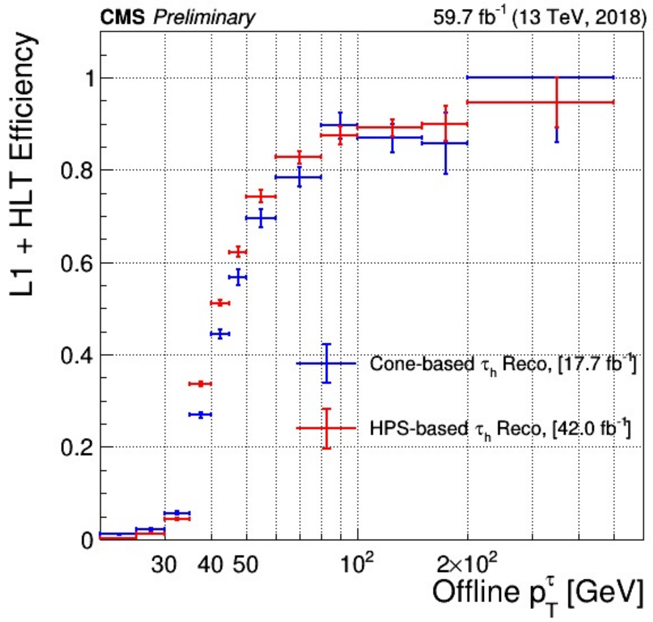


Di-Tau Trigger, Hadronic Leg

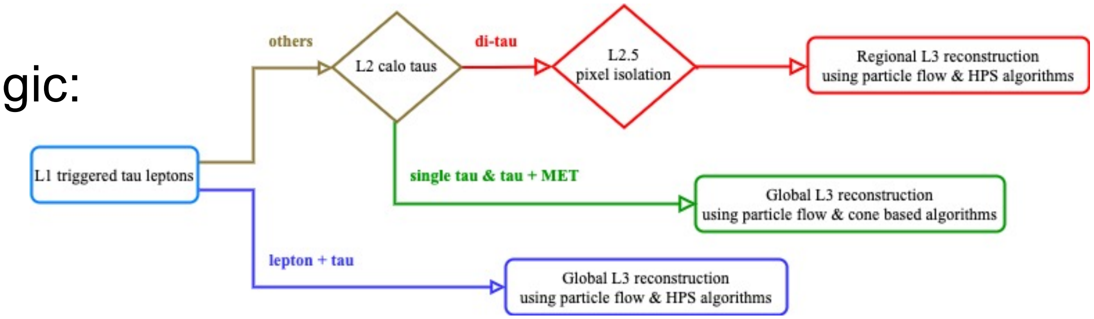
CMS-DP-2019-012

- Level-1 uses topological clustering of calorimeter deposits
- HLT reconstruction
  - Cone based reconstruction prior to mid-2018
  - “Hadron Plus Strips” reconstruction of exclusive decay modes used afterward (as offline)
- Used in combination with e, μ or another τ_h





HLT Logic:



Displaced Di-Muon Trigger

CMS-DP-2019-028

- Double muon trigger using only muon system (without tracker) and without a vertex constraint, for LLP searches
  - $p_T > 23$ GeV each leg
- Performance measured using cosmic rays and compared with BSM signal MC:
  - Efficient for very large impact parameters

Selecting Your Data: Trigger Menus

- A trigger "menu" represents the large set of selection criteria for the broad physics program of the experiment, e.g.
  - Entrée 1: 2 muons of opposite charge, one with $P_T > 17$ GeV and one with $P_T > 8$ GeV
  - Entrée 2: "jets" with a total scalar jet $H_T > 500$ GeV
- Crudely speaking, about one trigger per analysis topic
  - Some triggers are very general and serve many analyses (preference)
  - Some are very specific (needed to keep manageable rates)
  - And some are "backup" triggers required for the control regions of specific analyses
- There are separate L1 and HLT menus
  - Level-1 menu has ~300 items for CMS
  - HLT menu has ~600 items for CMS

Note, however, that there are efforts to go **triggerless**! Keep all the events, but only a selective fraction of data per event

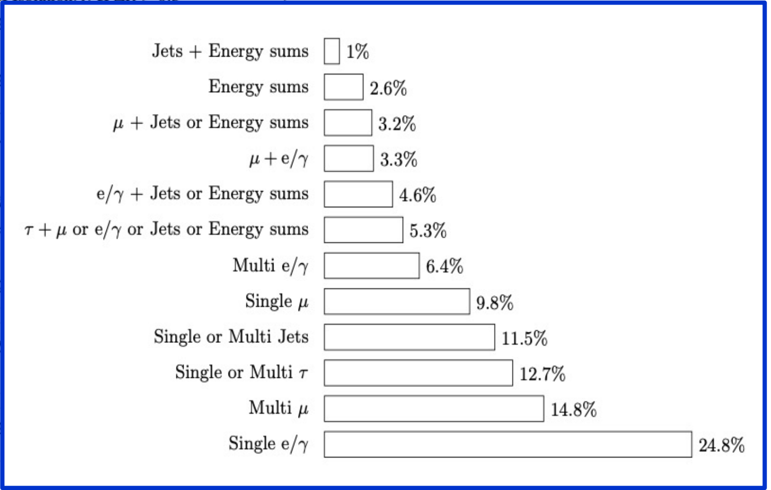
Level-1: Most Used Triggers for Run 2

Algorithm	Requirements (p_T, E_T, $m_{\mu\mu}$, and m_{jj} in GeV)
Muons	
Single μ	$p_T > 22$ & Tight quality
Double μ	$p_T > 15, 7$ & Medium quality
Double μ	$p_T > 15, 5$ & Tight quality
Double μ	$p_T > 8, 8$ & Tight quality
Double μ + mass	$p_T > 4.5$ & $\lvert\eta\rvert < 2.0$ & Tight quality & OS & $m_{\mu\mu} > 7$
Double $\mu + \Delta R$	$p_T > 4$ & Tight quality & OS & $\Delta R < 1.2$
Double $\mu + \Delta R$	$p_T > 0$ & $\lvert\eta\rvert < 1.5$ & Tight quality & OS & $\Delta R < 1.4$
Double μ + BX	$p_T > 0$ & $\lvert\eta\rvert < 1.4$ & Medium quality & Non-colliding BX
Triple μ	$p_T > 5, 3, 3$ & Medium quality
Triple μ	$p_T > 3, 3, 3$ & Tight quality
Triple μ + mass	$p_T > 5, 3.5, 2.5$ & Med. qual.; two μ OS & $p_T > 5, 2.5$ & $5 < m_{\mu\mu} < 17$
Triple μ + mass	Three μ any qual.; two μ & $p_T > 5, 3$ & Tight qual. & OS & $m_{\mu\mu} < 9$
Electrons / photons	(e/γ)
Single e/γ	$p_T > 60$
Single e/γ	$p_T > 36$ & $\lvert\eta\rvert < 2.5$
Single e/γ	$p_T > 28$ & $\lvert\eta\rvert < 2.5$ & Loose isolation
Double e/γ	$p_T > 25, 12$ & $\lvert\eta\rvert < 2.5$
Double e/γ	$p_T > 22, 12$ & $\lvert\eta\rvert < 2.5$ & Loose isolation
Triple e/γ	$p_T > 18, 17, 8$ & $\lvert\eta\rvert < 2.5$
Triple e/γ	$p_T > 16, 16, 16$ & $\lvert\eta\rvert < 2.5$
Tau leptons (τ)	
Single τ	$p_T > 120$ & $\lvert\eta\rvert < 2.1$
Double τ	$p_T > 32$ & $\lvert\eta\rvert < 2.1$ & Isolation
Jets	
Single jet	$p_T > 180$
Single jet + BX	$p_T > 43$ & $\lvert\eta\rvert < 2.5$ & Non-colliding BX
Double jet	$p_T > 150$ & $\lvert\eta\rvert < 2.5$
Double jet + $\Delta\eta$	$p_T > 112$ & $\lvert\eta\rvert < 2.3$ & $\Delta\eta < 1.6$
Double jet + mass	$p_T > 110, 35$; two jets $p_T > 35$ & $m_{jj} > 620$
Double jet + mass	$p_T > 30$ & $\lvert\eta\rvert < 2.5$ & $\Delta\eta < 1.5$ & $m_{jj} > 300$
Triple jet	$p_T > 95, 75, 65$; two jets $p_T > 75, 65$ & $\lvert\eta\rvert < 2.5$
Energy sums	
E_T^{miss}	$E_T^{miss} > 100$ (Vector sum of p_T of calorimeter deposits with $\lvert\eta\rvert < 5.0$)
H_T	$H_T > 360$ (Scalar sum of p_T of all jets with $p_T > 30$ and $\lvert\eta\rvert < 2.5$)
E_T	$E_T > 2000$ (Scalar sum of p_T of calorimeter deposits with $\lvert\eta\rvert < 5.0$)

Algorithm	Requirements (p_T, E_T, $m_{\mu\mu}$, and m_{jj} in GeV)
Two objects	
Single μ + Single e/γ	$p_T(\mu) > 20$ & Tight quality(μ) & $p_T(e/\gamma) > 10$ & $\lvert\eta(e/\gamma)\rvert < 2.5$
Single μ + Single e/γ	$p_T(\mu) > 7$ & Tight quality(μ) & $p_T(e/\gamma) > 20$ & $\lvert\eta(e/\gamma)\rvert < 2.5$
Single μ +	$p_T(\mu) > 18$ & $\lvert\eta(\mu)\rvert < 2.1$ & Tight quality(μ) &
Single τ	$p_T(\tau) > 24$ & $\lvert\eta(\tau)\rvert < 2.1$
Single $\mu + H_T$	$p_T(\mu) > 6$ & Tight quality(μ) & $H_T > 240$
Single e/γ +	$p_T(e/\gamma) > 22$ & $\lvert\eta(e/\gamma)\rvert < 2.1$ & Loose isolation(e/γ) &
Single τ	$p_T(\tau) > 26$ & $\lvert\eta(\tau)\rvert < 2.1$ & Isolation(τ) & $\Delta R > 0.3$
Single e/γ +	$p_T(e/\gamma) > 28$ & $\lvert\eta(e/\gamma)\rvert <$ 2
Single jet	$p_T(\text{jet}) > 34$ & $\lvert\eta(\text{jet})\rvert <$ 2.
Single $e/\gamma + H_T$	$p_T(e/\gamma) > 26$ & $\lvert\eta(e/\gamma)\rvert <$ 2
Single $\tau + E_T^{miss}$	$p_T(\tau) > 40$ & $\lvert\eta(\tau)\rvert < 2.1$ &
Single jet + E_T^{miss}	$p_T(\text{jet}) > 140$ & $\lvert\eta(\text{jet})\rvert <$ 2.
Three objects	
Single μ	$p_T(\mu) > 12$ & $\lvert\eta(\mu)\rvert < 2.3$ &
Double jet + ΔR	$p_T(\text{jet}) > 40$ & $\Delta\eta(\text{jet,jet})$
Single μ +	$p_T(\mu) > 3$ & $\lvert\eta(\mu)\rvert < 1.5$ &
Single jet + E_T^{miss}	$p_T(\text{jet}) > 100$ & $\lvert\eta(\text{jet})\rvert <$
Double $\mu + H_T$	$p_T(\mu) > 3$ & Tight quality(μ
Double μ +	$p_T(\mu) > 0$ & Medium quality
Single jet + ΔR	$p_T(\text{jet}) > 90$ & $\lvert\eta(\text{jet})\rvert <$ 2.
Double μ + Single e/γ	$p_T(\mu) > 5$ & Tight quality(μ
Double e/γ + Single μ	$p_T(e/\gamma) > 12$ & $\lvert\eta(e/\gamma)\rvert <$ 2
Double $e/\gamma + H_T$	$p_T(e/\gamma) > 8$ & $\lvert\eta(e/\gamma)\rvert <$ 2.
Four objects	
Double μ + Double e/γ	$p_T(\mu) > 3$ & Medium quality(μ) & OS(μ) & $p_T(e/\gamma) > 7.5$
Double μ + Double e/γ	$p_T(\mu) > 5$ & Medium quality(μ) & OS(μ) & $p_T(e/\gamma) > 3$
Five objects	
Double $\mu + E_T^{miss}$ +	$p_T(\mu) > 3$ & Tight quality(μ) & $E_T^{miss} > 50$ &
Single jet OR	($p_T(\text{jet}) > 60$ & $\lvert\eta(\text{jet})\rvert < 2.5$) OR
Double jet	($p_T(\text{jet}) > 40$ & $\lvert\eta(\text{jet})\rvert < 2.5$)
H_T + Quad jet	$H_T > 320$ & $p_T(\text{jet}) > 70, 55, 40, 40$ & $\lvert\eta(\text{jet})\rvert < 2.4$

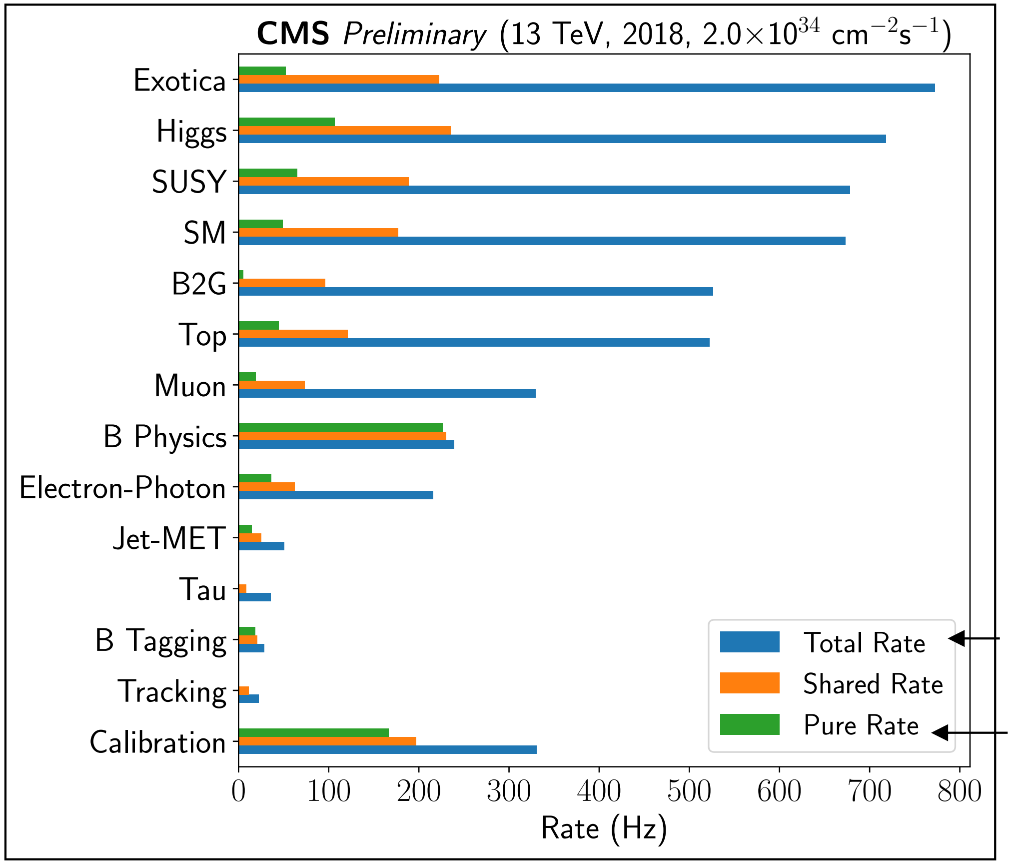
JINST 15, P10017 (2020)

Rate breakdown



CMS Run 2 HLT Menu Physics Allocation

CMS-DP-2018-057



inclusive

exclusive

- Measured at L=1.2E34 but scaled to L=2.0E34
  - 510 Hz (~1/3) to searches
  - 530 Hz (~1/3) to Standard Model
  - 230 Hz to B-physics
  - 200 Hz to objects
  - 200 Hz to calibration

☐ Note that calibration events are of reduced size

HLT: Some Representative Triggers

- Selected triggers from 2018 HLT menu
  - Sums to ~50% of menu rate
  - Recall ~600 trigger lines

Trigger	Threshold (GeV)
single muon	50
single muon (isolated)	24
double muons	37, 27
double muons (isolated)	17, 8
single electron (isolated)	28
double electrons	25, 25
single photon	200
single photon (isolated)	110, barrel only
double photons	30, 18
single taus	180
double taus	35, 35
single jets	500
multijets with b-tagging	75, 60, 45, 40 (HT>330)
total transverse energy	1050
missing transverse energy	120

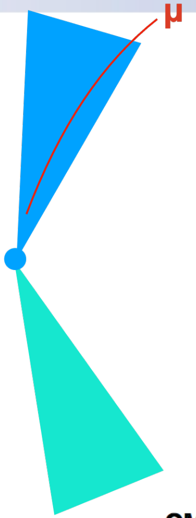
How to Increase Data Storage Rate? Parking

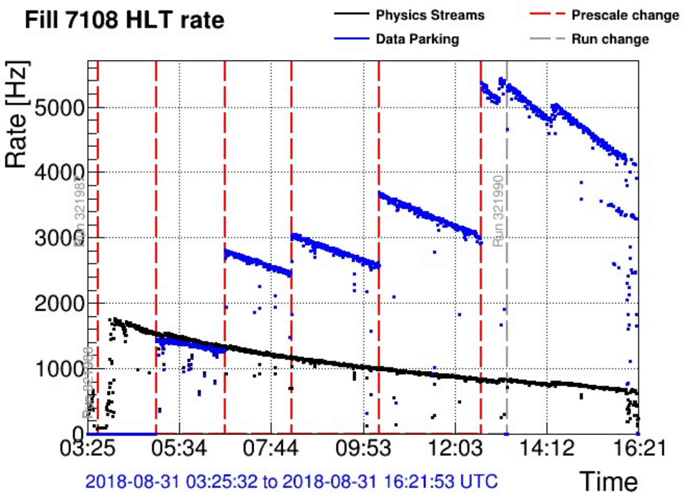
- Recall why a trigger is needed
  - Too much data to continuously stream to disk for storage and/or computer processing reasons
- If we can at least handle the storage rate, we can record data at an even higher rate to disk, but postpone processing that data until experiment is no longer running
- This is known as "data parking"
  - CMS does this to record more data for B physics studies, etc.
    - 10 billion events from Run 2
  - Plan to continue this for Run 3
- Can take advantage of year-end or long shutdown periods
  - For example, LHC has been in a 3-year long shutdown for upgrades

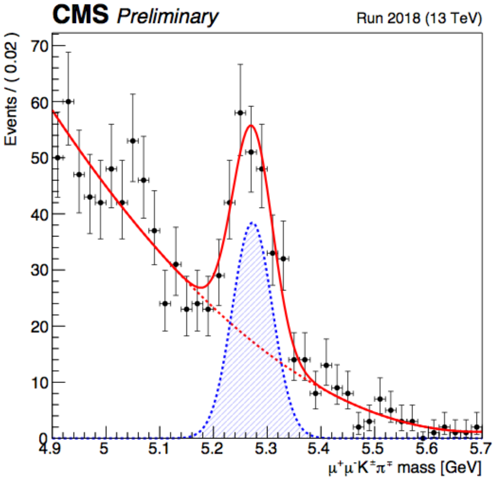
B Data Parking

CMS-DP-2019-043

- Tag on muon from one side of event, other side unbiased
- The rate for the CMS physics streams (black curve) falls from ~2 kHz during this LHC fill in 2018
- The rate for the B Parking stream (blue curve) increases in steps at changes in the prescale column (loosening of trigger thresholds) during a CMS run, reaching as high as ~5 kHz.







B→J/ψ(μμ)K^*(πK)

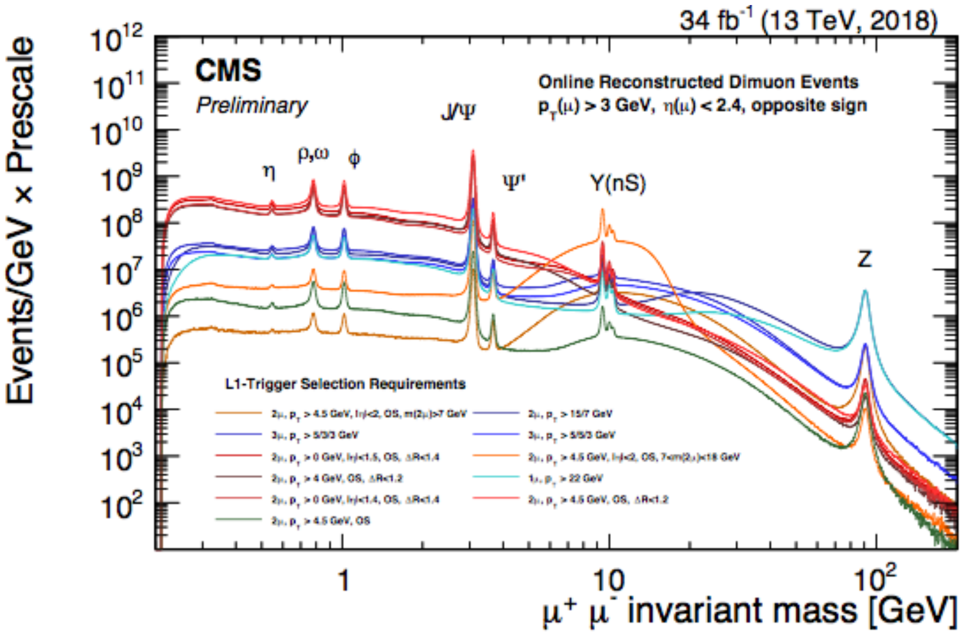
Another Approach: Data Scouting

- The limit on DAQ is the bandwidth of the data to record to disk (few GB/s), not really the event rate per se
- If the amount of data to store per event is small enough, you could store in principle all events that occur, or at least a very high rate of them
  - e.g. ~10 kB vs. ~1 MB per event
  - CMS does this for calibration data, for example
- "Data scouting" (aka Trigger Level Analysis, or Turbo Stream) is a recent invention by the LHC experiments to store only a small summary of reconstructed event quantities, and not all the raw data, in order to record a higher rate of events
  - Allows much lower trigger thresholds, and thus a higher acceptance of a physics process
  - Generally does not allow reprocessing data afterward (from new calibrations or alignments)
    - ❑ Therefore only reliable for a mature experiment
  - Can induce additional HLT processing time from a higher input rate

Run 2 Di-Muon Scouting Trigger

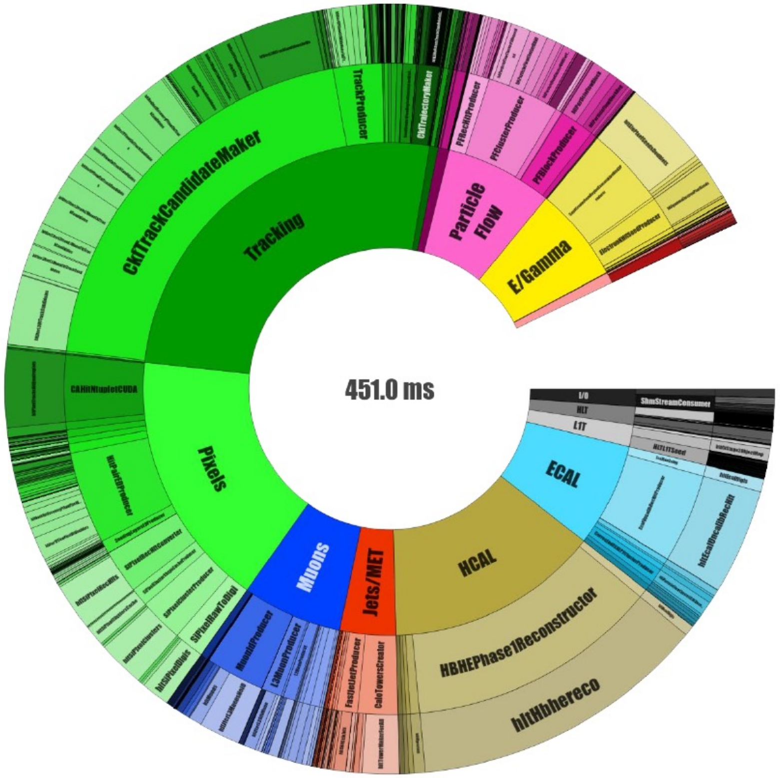
CMS-DP-2018-055

- Events required to have two muons with $p_T > 3$ GeV and $|\eta| < 2.4$
- HLT is seeded by various L1 triggers
- Events able to be recorded at a rate of up to 3 kHz for instantaneous luminosities of $L=2\times10^{34}$ Hz/cm^2
- Sensitive to low invariant masses
  - ❑ Target dark photon searches



HLT CPU Processing Time per Event

The computing cost of running the menu; limits what processing can be done online



- Distribution of CPU time in different instances of CMSSW modules (outermost ring), and grouped by physics object or detector (innermost ring).
- The HLT configuration is based on the 2018 definition, with minimal updates to the local reconstruction to reflect preparation for Run 3
  - Full HLT node (2x Intel Skylake Gold 6130) with HT enabled, running 16 jobs in parallel with 4 threads each

Tracking algorithms dominate

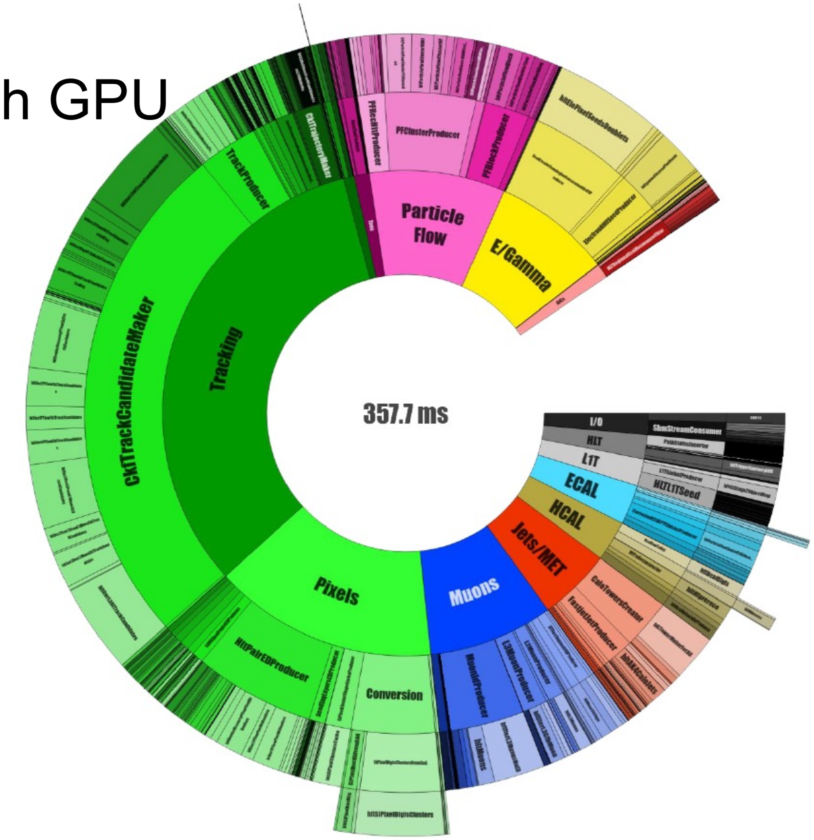
Run 3 Preparation: GPU Acceleration

- One constraint on the HLT is the processing time per event
  - More computing time allows for tracking to be run on a larger fraction of events (e.g. for data scouting) and for more sophisticated algorithms
- Effort launched several years ago to port CMS algorithms to GPUs to accelerate the computations
  - Pixel-based tracking and vertexing ("Patatrack")
  - ECAL and HCAL local reconstruction
    - Written in NVidia CUDA language
- For Run 3, CMS plans to equip the HLT farm with GPUs
  - Configuration to be determined later this year
  - Heterogeneous computing is essentially a must for HL LHC, so will need the experience to port even more algorithms

Frontiers in Big Data 3 (2020), no. 10, 49

HLT Processing Time with GPU Acceleration

with GPU



- Distribution of CPU time with GPU offloading in different instances of CMSSW modules (outermost ring), and grouped by physics object or detector (innermost ring).
- Offloaded pixel-based tracking (Patatrack) and vertex reconstruction ECAL and HCAL local reconstruction
  - Full HLT node + NVIDIA Tesla T4 GPU

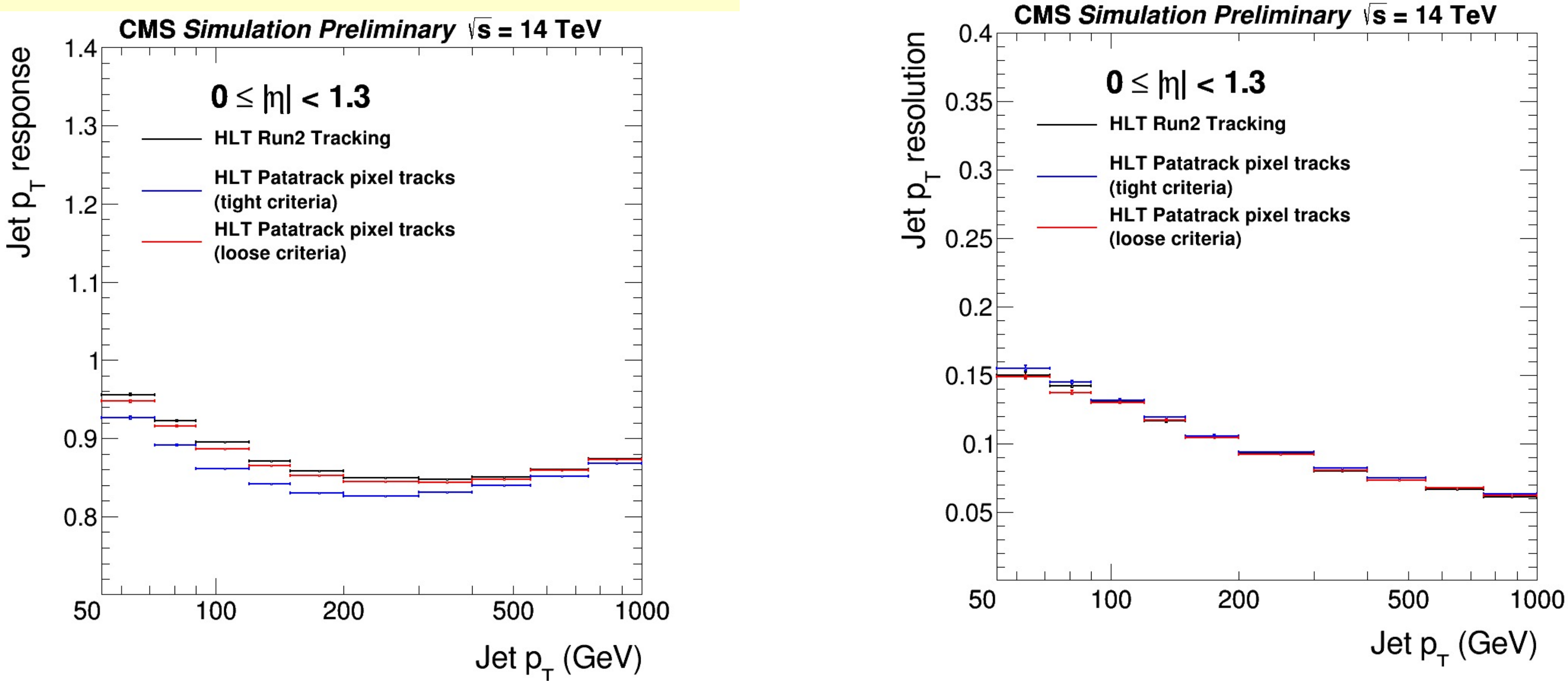
21% offloaded

Run 3 Prep: Particle Flow with Pixel Tracks

- The pixel-based tracking can be run on every input event to a GPU-equipped HLT (100 kHz), unlike standard tracking
  - But not entirely a "free lunch", as the pixel tracker has a much smaller radius than the strip tracker → degraded p_T resolution at high p_T
- Nevertheless, offers an interesting new ingredient!
- A study of the performance of particle flow (an optimized combination of tracks and calorimeter objects) using only pixel tracks has recently been done
  - Could benefit Scouting triggers, for example

Pixel-based PF Jet Response and Resolution

More in Jakob Salfeld-Nebgen's talk



- Promising! Quite comparable to PF jets with full tracking

Other Things Possible for Run 3

- The longitudinal segmentation of HCAL with new QIE readout fully was completed during LS2 (barrel, endcap done in 2018), and could benefit LLP triggers
  - 4 depths in barrel, 6-7 in endcap
  - Also timing at 0.5 ns resolution
- The Phase-2 upgrade of the trigger systems for HL LHC is underway. Often ideas spill over to the current systems…
  - e.g. Offloading of CPU intensive algorithms to GPUs
  - Machine learning inferences(*) for regression or categorization in FPGAs or GPUs
    - ☐ (*) Just keep in mind that the trigger must be robust, with a well understood performance for physics measurements

Summary/Outlook

- CMS has a mature and flexible trigger system proven for physics analyses over Runs 1 and 2
  - Includes data scouting, parking, and other innovations
- Additions for Run 3 include
  - Highly displaced muon triggers at Level-1
  - Additional muon coverage in forward region
  - Pixel-based tracking for all events using GPUs at HLT
  - Continued/expanded use of data scouting triggers and data parking
- Focus indeed should be on opening up new trigger phase space!