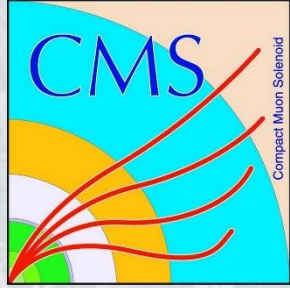


Run-3 Commissioning of CMS Online HLT reconstruction using GPUs

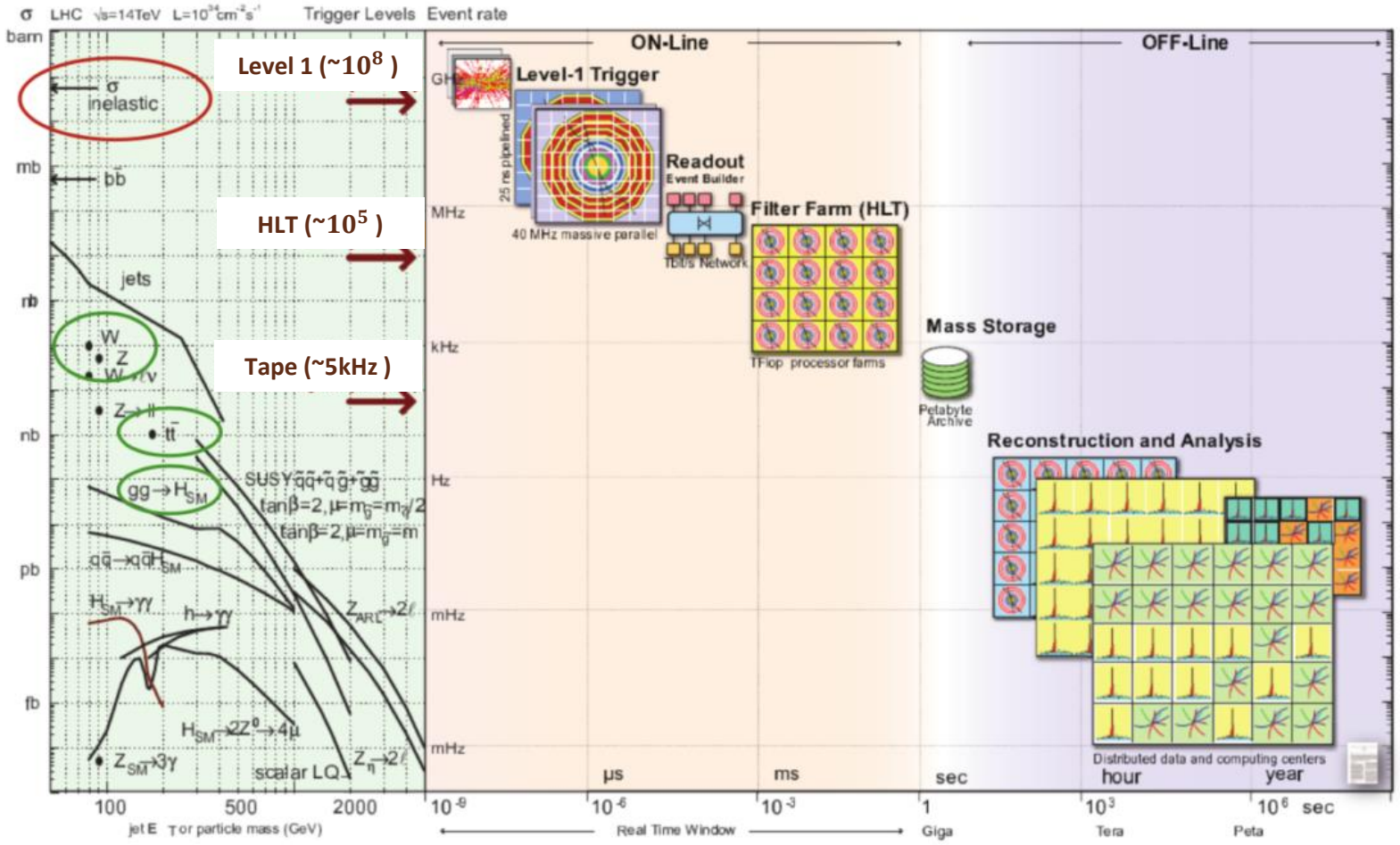
Ganesh Parida, University of Wisconsin-Madison
on behalf of the CMS collaboration

CMS High Level Trigger

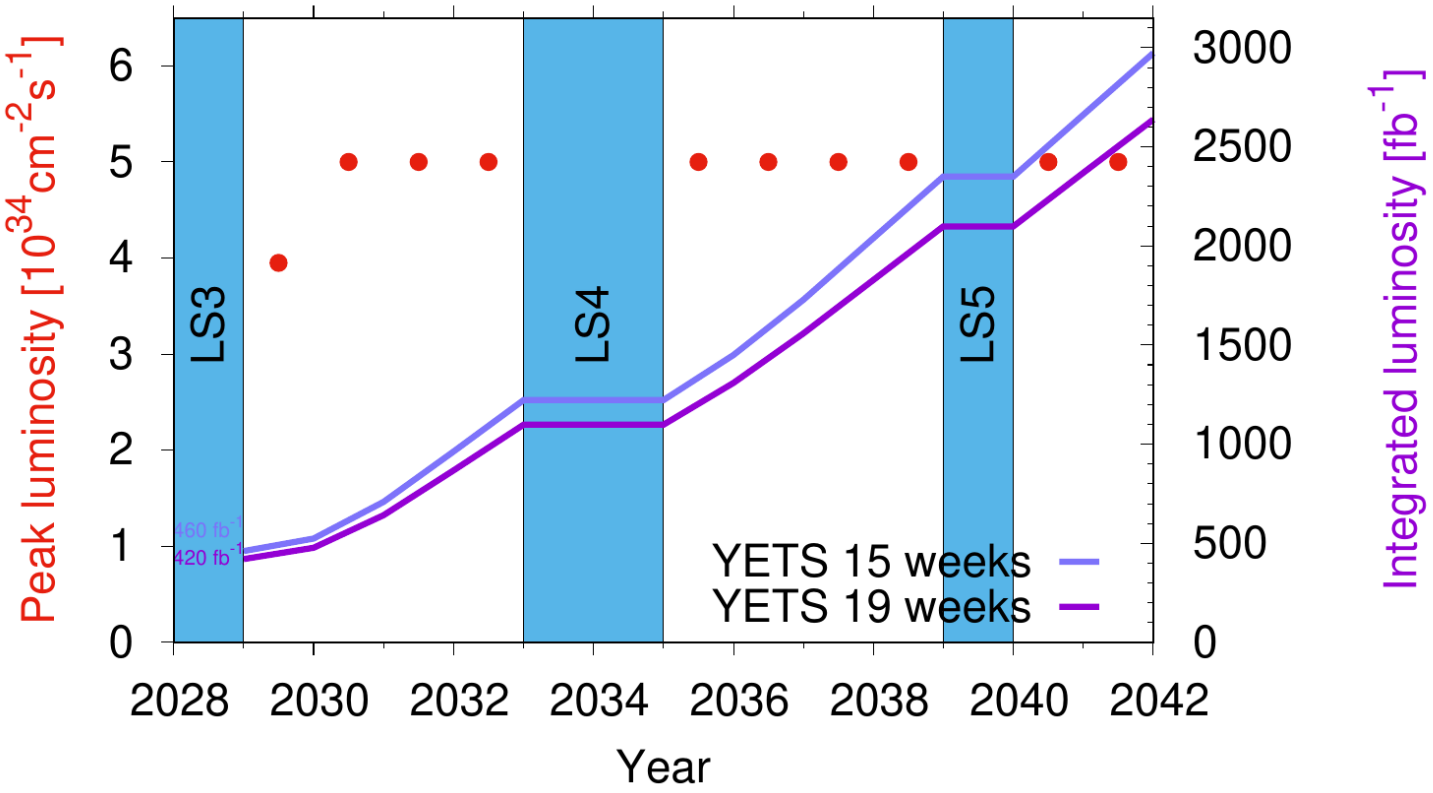
- LHC collision rate is 40 MHz – Impossible read and process all events at full granularity due to resource constraints

- Two Tier Trigger system to filter events
  - Level 1 Trigger : custom electronics(FPGAs etc.) - rate reduction to **100kHz**
  - High Level Trigger (HLT)

- HLT – runs a streamlined version of CMS reconstruction software further reducing the rate to ~5kHz for Run3



Need for GPUs in HLT…



Challenges of HL-LHC

- Approximately two and half fold increase in instantaneous luminosity and detector upgrades (High Granularity Calorimeters, new tracker, Muon chambers)
- Translates to factor 30 increase in computing resources for similar Physics goals
- Need for High throughput requires us to take advantage of parallelism wherever possible – and the event reconstruction in detectors can be parallelized

Keeping Up with the Market trends

- Hardware availability in markets is tilting more towards heterogenous computing
- Making use of GPUs is more efficient in terms of cost – added benefit of being energy efficient
- Adapting the CMS software to use GPUs will make it compatible with resources available in the future

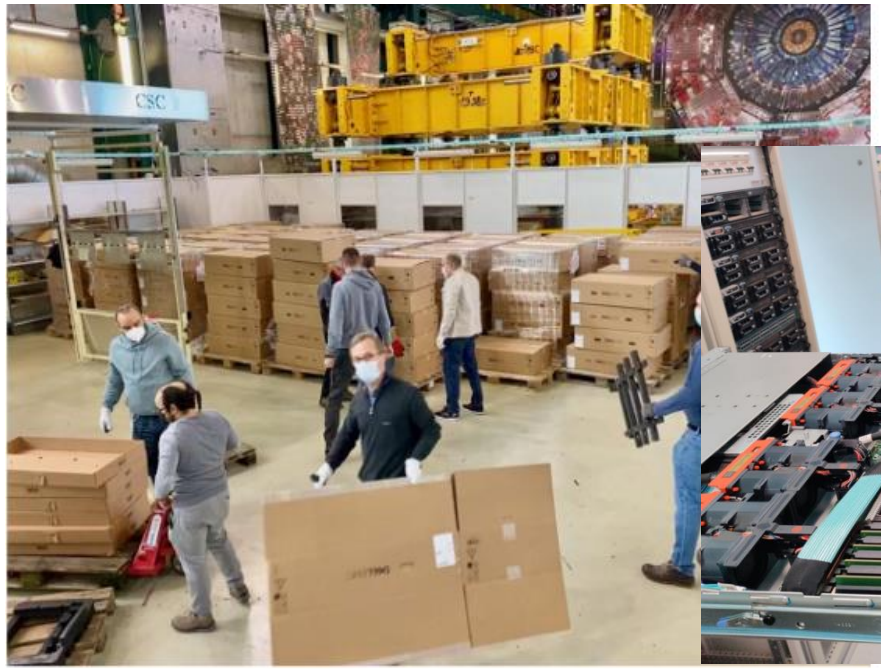
Physics Improvements

- Parallelization requires re-engineering the existing code – and in the process we achieve gains in physics performance

- More computing power allows CMS to invest in accurate methods of reconstruction – high quality physics objects
  - Pixel track seeding
  - Move to single iterative tracking
  - Dedicated tracking for Long Lived Particles(LLP)
  - Soft electron reconstruction

- Scouting : Strategy pioneered by CMS in which we increase event rate (by lowering thresholds) and decrease event size (Trigger Bandwidth being the limiting factor) – save and perform analysis with objects reconstructed at the HLT, instead of offline reconstruction. No RAW data is stored
  - New high level physics objects (electrons, photons and pixel only track in addition to objects stored during Run-2 such as muons, jets and jet constituents)
  - Event storage rate of up to 30% of L1 rate
  - Investigate physics processes that have a rate too large for available offline resources

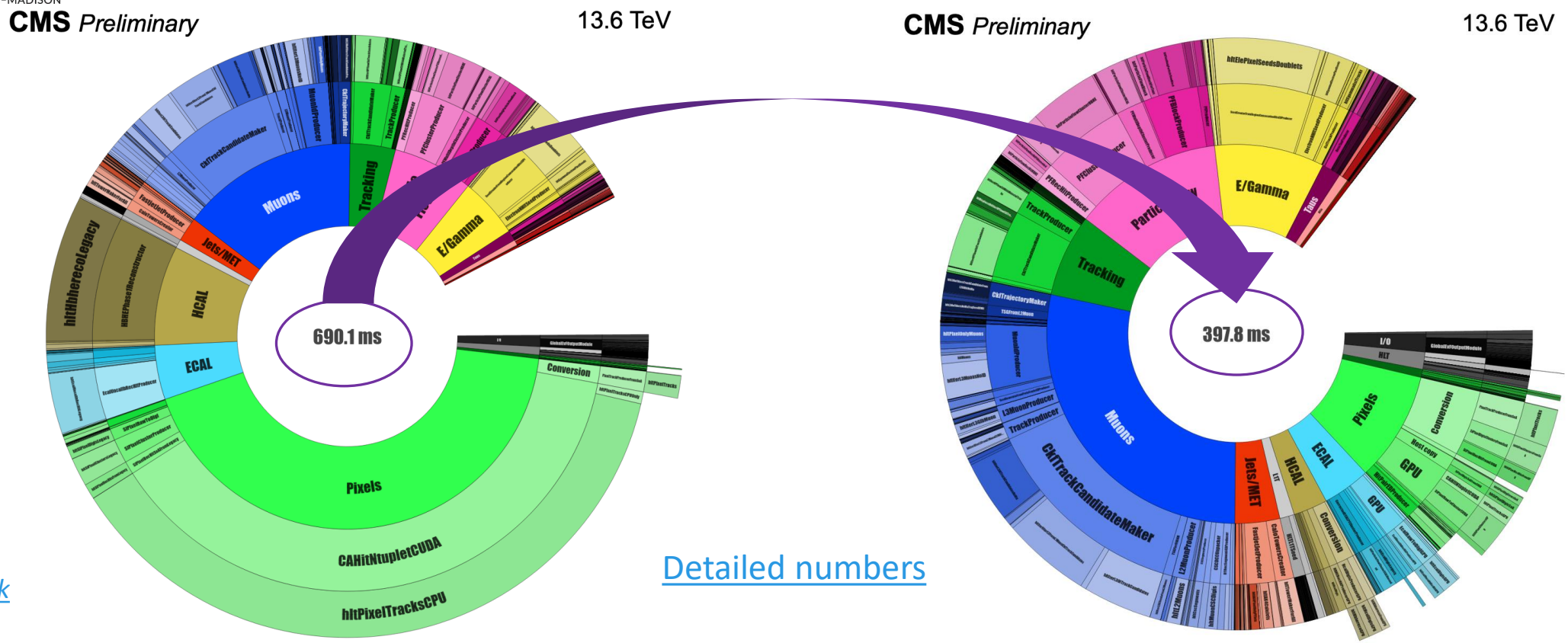
New HLT Farm for Run-3



Source

- 200 machines:
  - 2x AMD EPYC 7763 “Milan” 64-core processors (128 Cores, 256 threads)
  - 2x NVIDIA T4
  - System Memory – 256 GB
- Each T4 has 2560 CUDA “threads” running at 1.59GHz, 16GB GDDR6 DRAM and 6MB L2 cache
- Power Consumption:
  - 2x AMD CPU ~ (2 x 280W) = 660 W
  - GPU ~ (2 x 70W) = 140W
  - System Memory ~96W

Reconstruction using GPUs



link

Detailed numbers

link

Average time per event for CPU Only Configuration

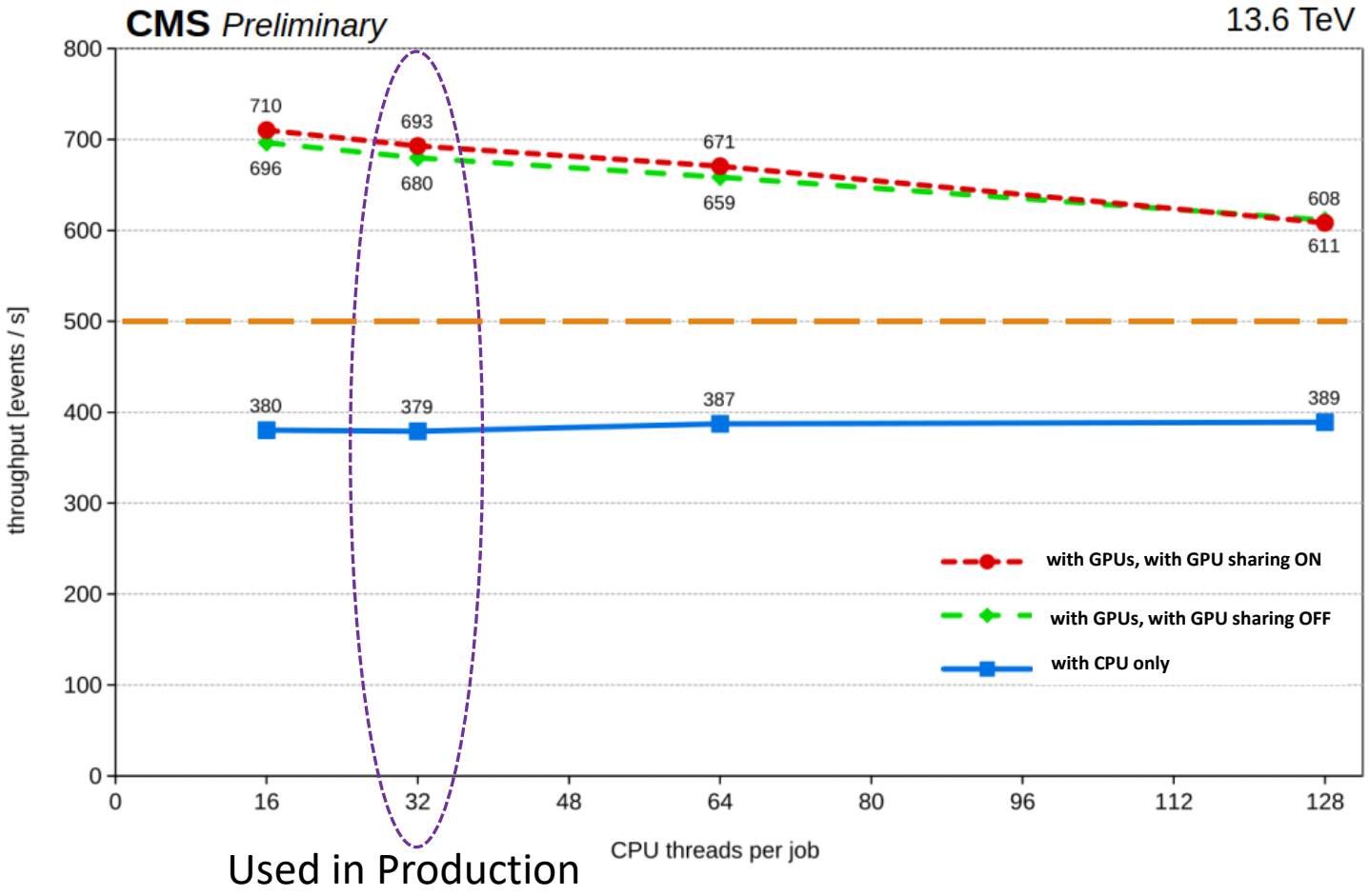
Average time per event for CPU + GPU Configuration

- Currently HCAL, ECAL, Pixel Local Reconstruction, Pixel-Only Track and Vertex Reconstruction (which seeds the full tracking, also used standalone for scouting) are running on GPUs – which has a huge impact on timing
- The execution time per event of was reduced by ~ 40%

HLT Throughput

Rough Estimates of the Requirements

- Input Level 1 rate - 100 kHz
- Machine Throughput requirement:
  - Input Level 1 Rate / number of machines (100 kHz / 200) = 500 Hz (Marked in Orange)



Used in Production

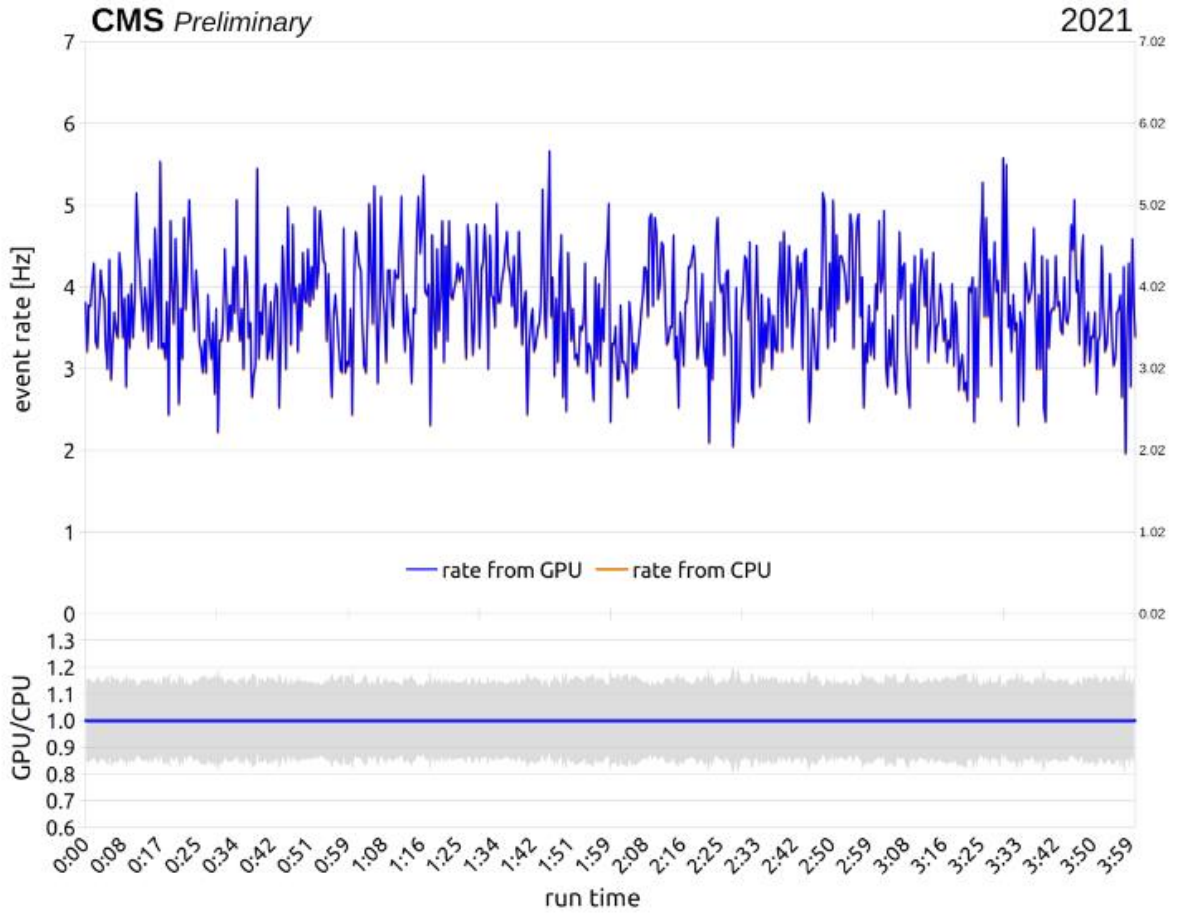
PP collisions data at 13.6 TeV in October, 2022, Average PileUp of 55

Throughput Increases by a factor of ~1.80
Power Consumption (per throughput) reduced by ~30%

Commissioning Timeline

- Integration of the GPU code into the CMS Software Framework in 2020-21
- A few machines of the old HLT farm were equipped with GPUs to take data with cosmics in 2021-22 and during the 900 GeV run (Oct, 2021)
- Integration in the central HLT menu in December 2021
- A pilot submission to validate the latest pre-release of the reconstruction software with simulated benchmark datasets was launched on GPU machines on the Grid. Commissioning during 900 GeV collisions (May, 2022)
- Migration to the new HLT Farm with all machines equipped with GPU by the end of June 2022

Source

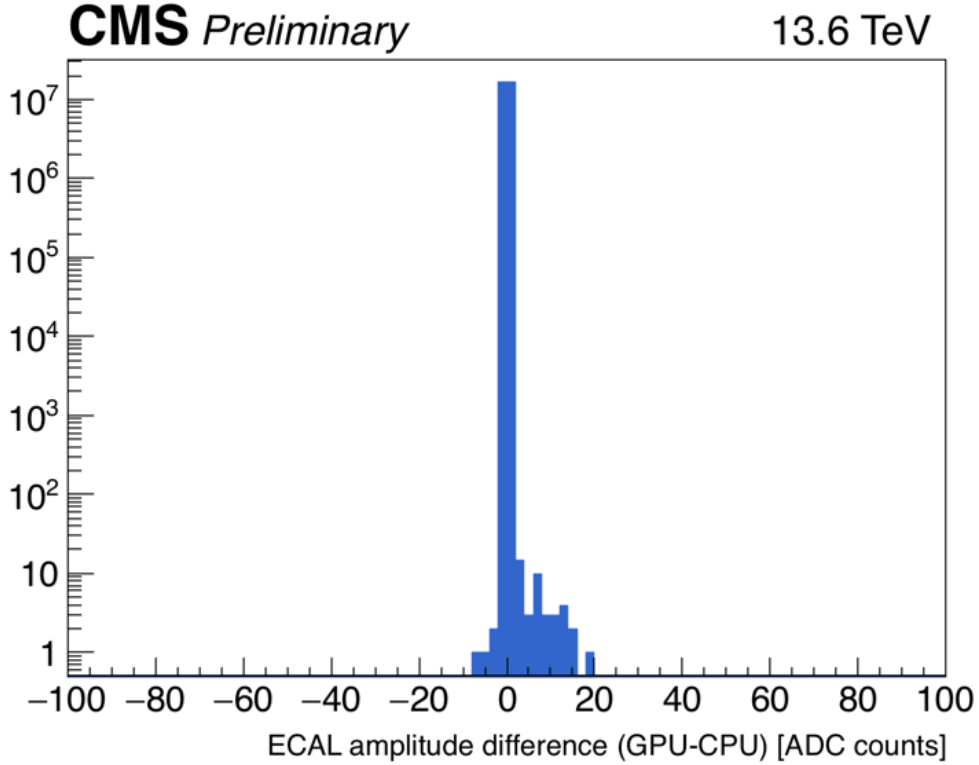


The CPU and the GPU rates were identical (comparison run in real time on stable beam collisions at the end of Oct,21): Events with at least one calorimetric Jet reconstructed above a threshold are shown

GPU Reconstruction Validation

Online Data Quality Monitoring

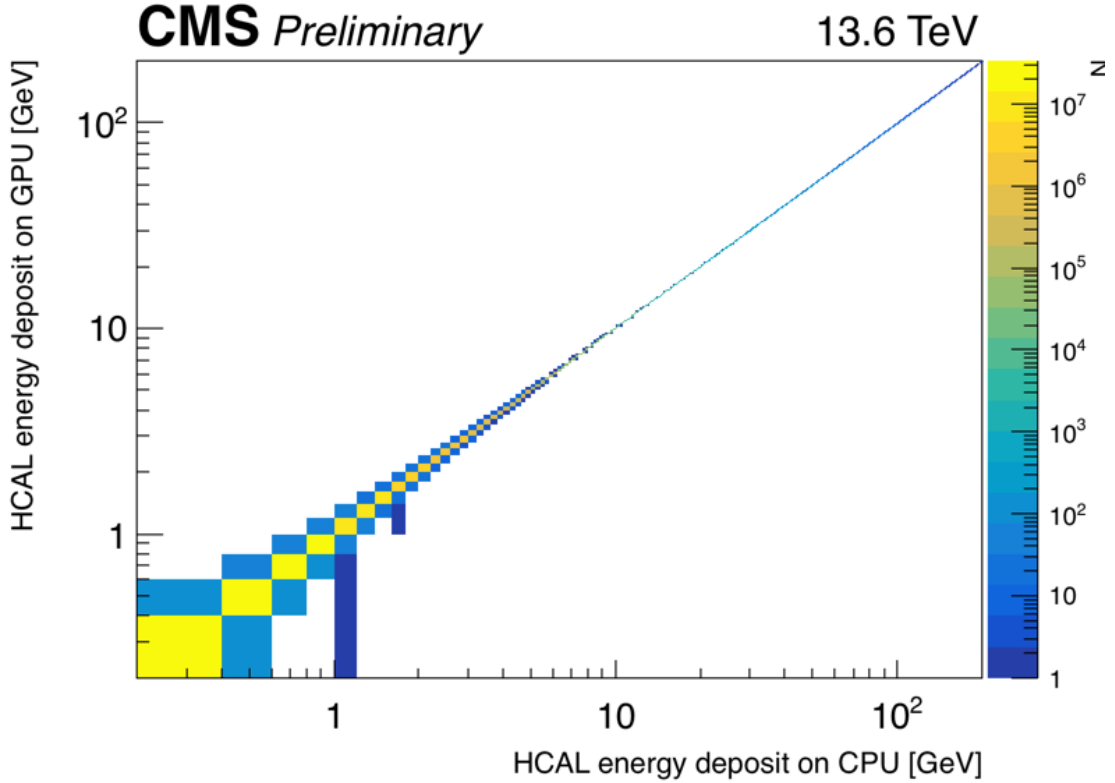
ECAL Reconstruction



- Difference in pulse amplitude of ECAL barrel when run on CPU and GPU – A fraction of pulses of the order of 10^{-6} pulses show a difference

Comparison done online on 13.6 TeV collisions events on Oct,22. CPUvsGPU Online DQM runs on 1 out of every 3000 events

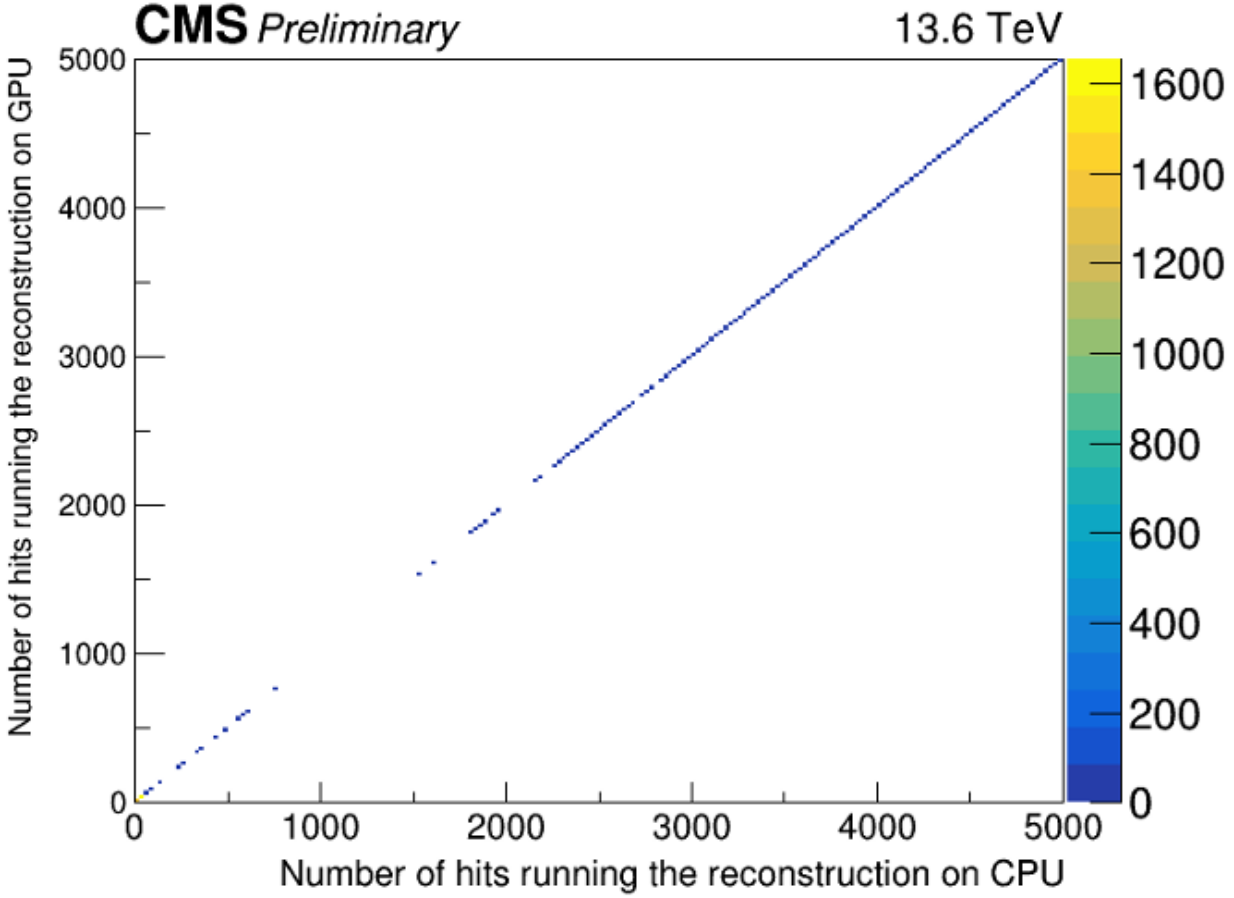
HCAL Reconstruction



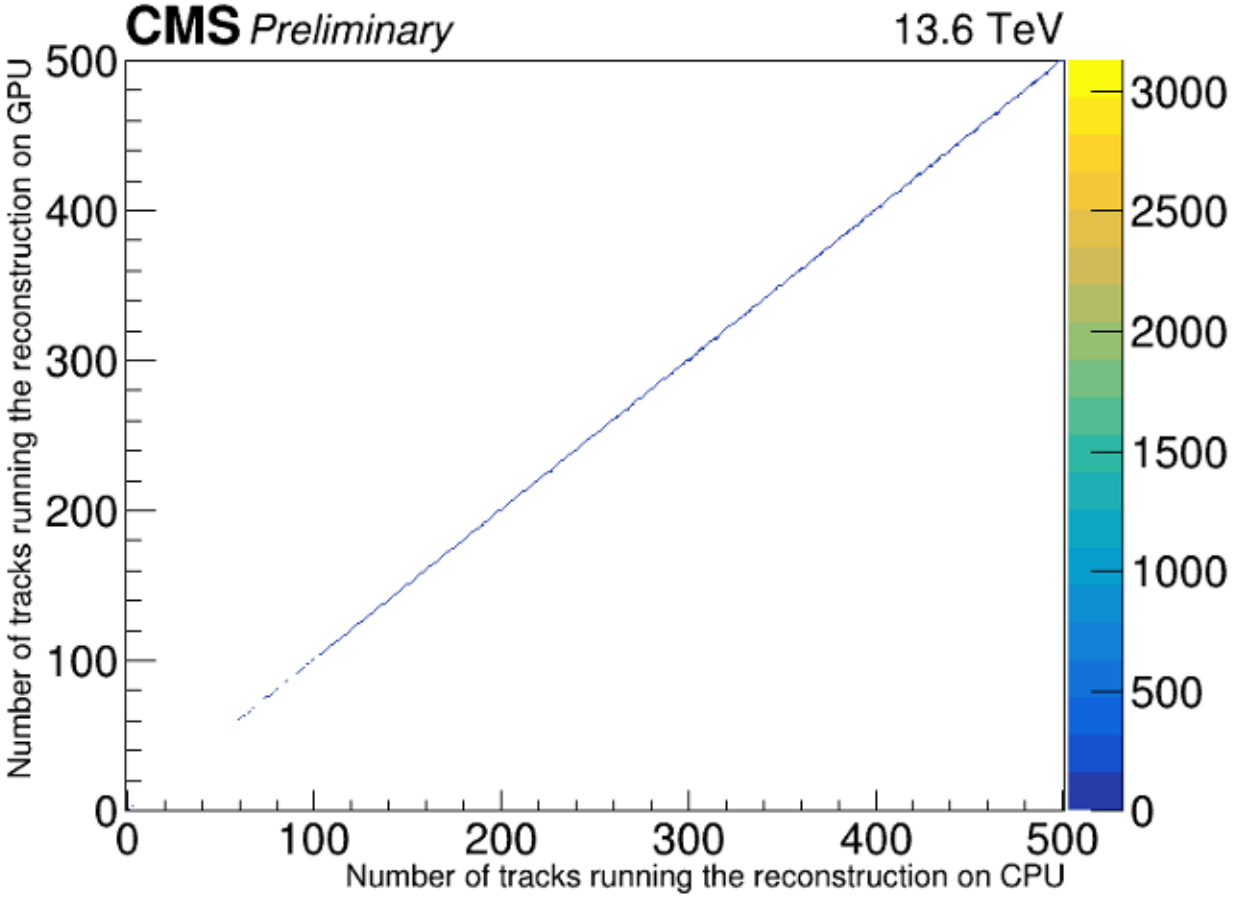
- Response of the same energy deposit (in GeV) reconstructed on GPU and on CPU, for both barrel and endcap. Off diagonal hits are O(10^{-6}) lower than diagonal

GPU Reconstruction Validation

Online Data Quality Monitoring



Reconstructed Pixel Hits

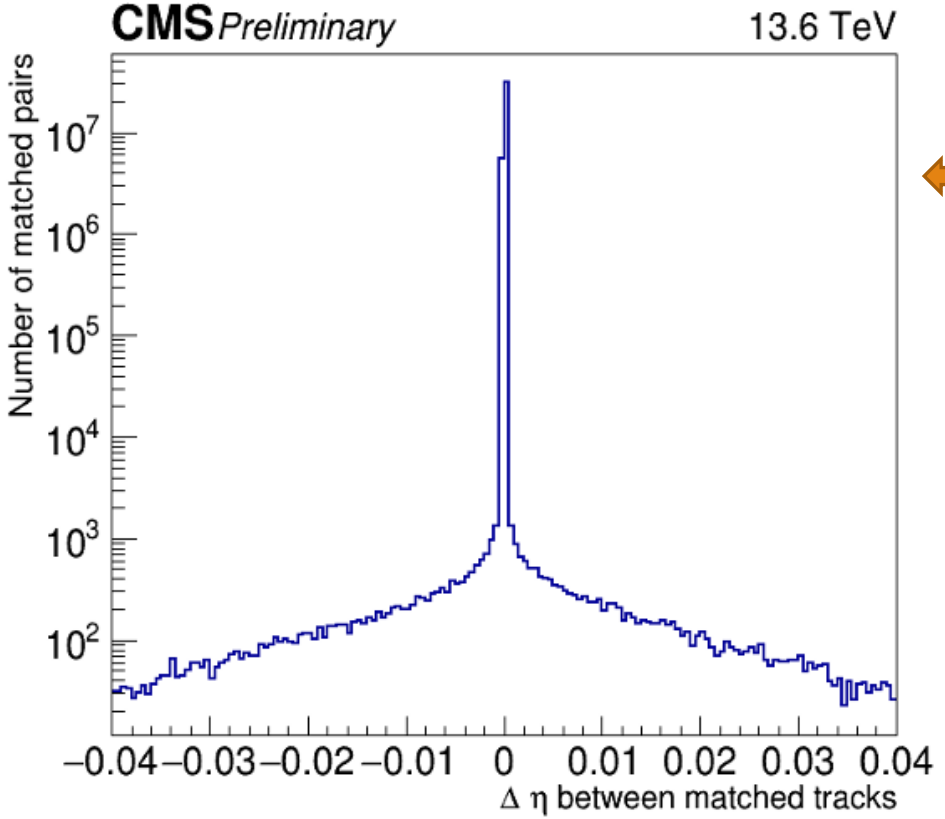


Pixel Tracks

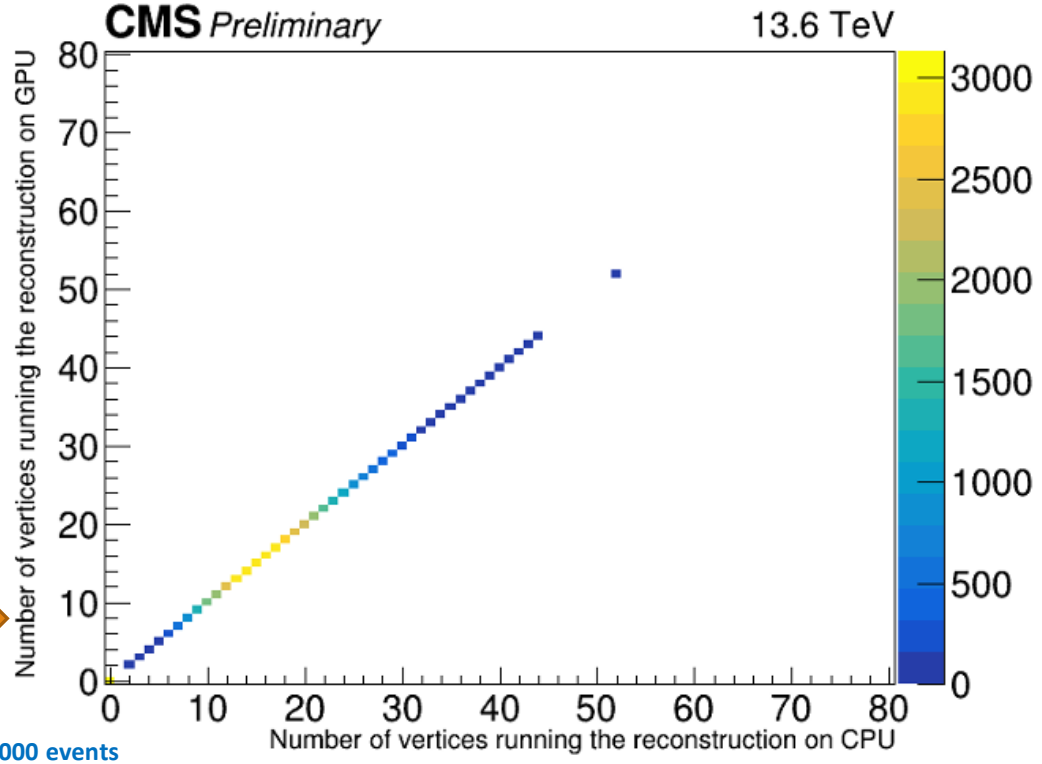
Comparison done online on 13.6 TeV collision events on Oct,22. CPUvsGPU Online DQM runs on 1 out of every 3000 events

GPU Reconstruction Validation

Online Data Quality Monitoring



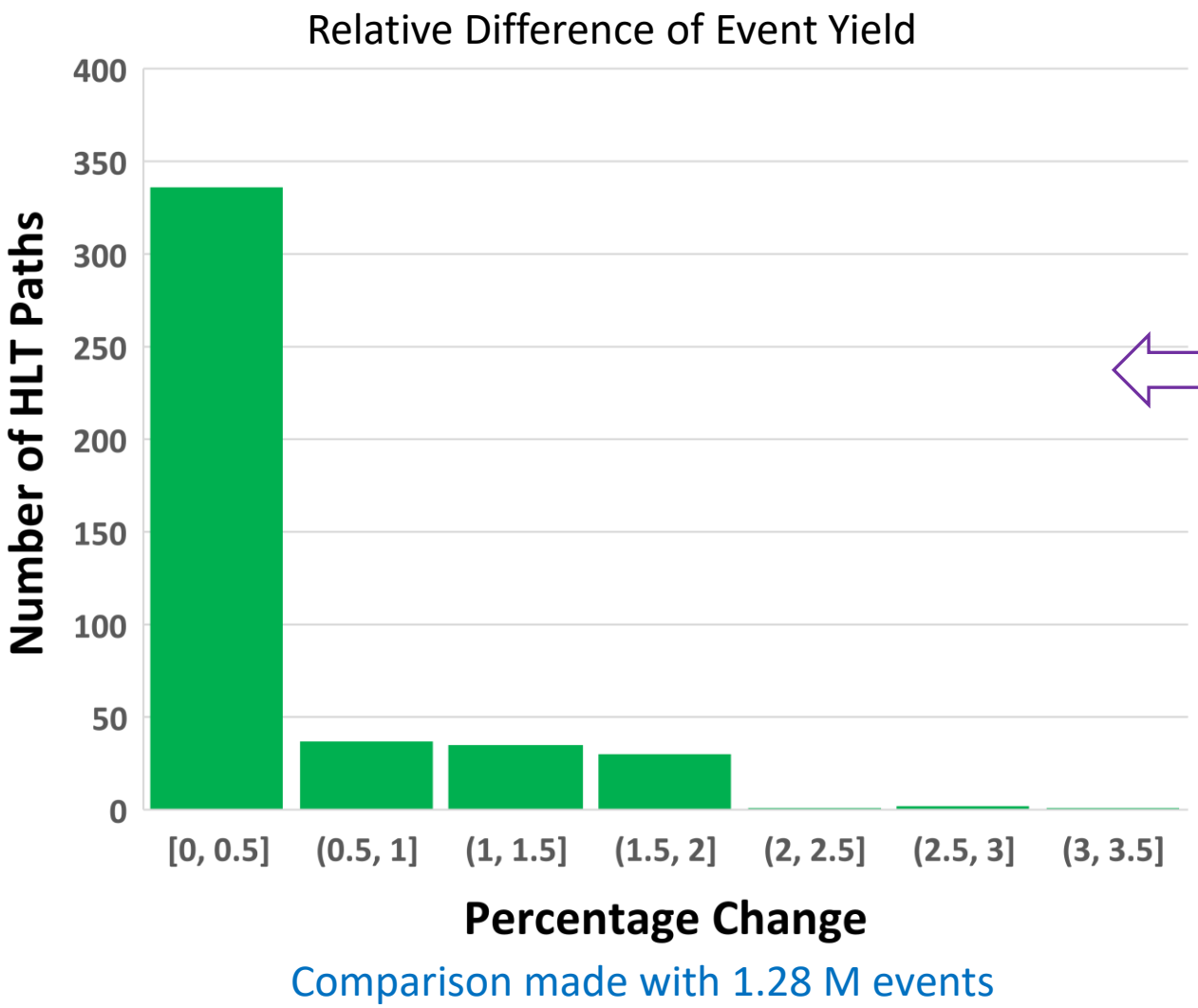
$\Delta\eta$ distributions of a CPU and closest GPU track



Pixel Vertices : Good Agreement

Comparison done online on 13.6 TeV collision events on Oct,22. CPUvsGPU Online DQM runs on 1 out of every 3000 events

Effect on Trigger Results



Comparison made with 1.28 M events

- Event by event difference in HLT trigger result – obtained after running the menu with and without GPUs (studies performed offline)
  - Out of ~700 paths : No differences are observed in around 400 paths
  - Considering paths with accept **counts above 100** (out of 1.28 M) : 99 % of the paths have a difference lower than 2%
- These differences are arising from slightly varying inputs between CPU and GPU (such as Pixel, HCAL, ECAL) to Particle Flow (PF) – A trigger path was introduced to monitor this behavior, **HLT PFJet40 GPUvsCPU v1** (5316 event accepts out of 2,312,690)
  - $\epsilon = {}^{5316}/_{2312690} = \mathbf{0.22}\%$ (very small fraction of events)
- Trigger Results have small differences, and the residual differences are being investigated

Conclusion & Outlook

- A new CMS HLT farm installed with each machine now equipped with 2 NVIDIA T4 . HCAL, ECAL, Pixel Local Reconstruction, Pixel Only Track and Vertex Reconstruction (used to seed the full tracking and standalone for scouting) offloaded to GPUs for Run3 – and ran successfully during 2022 data taking

- Average time spent on each event reduced by 40% and throughput increases by 80%. We also achieve 30% reduction in power consumption

- Online Reconstruction validation and offline comparison of CPU-GPU trigger results shows no significant discrepancy between GPU and CPU results – with those residual differences being actively investigated

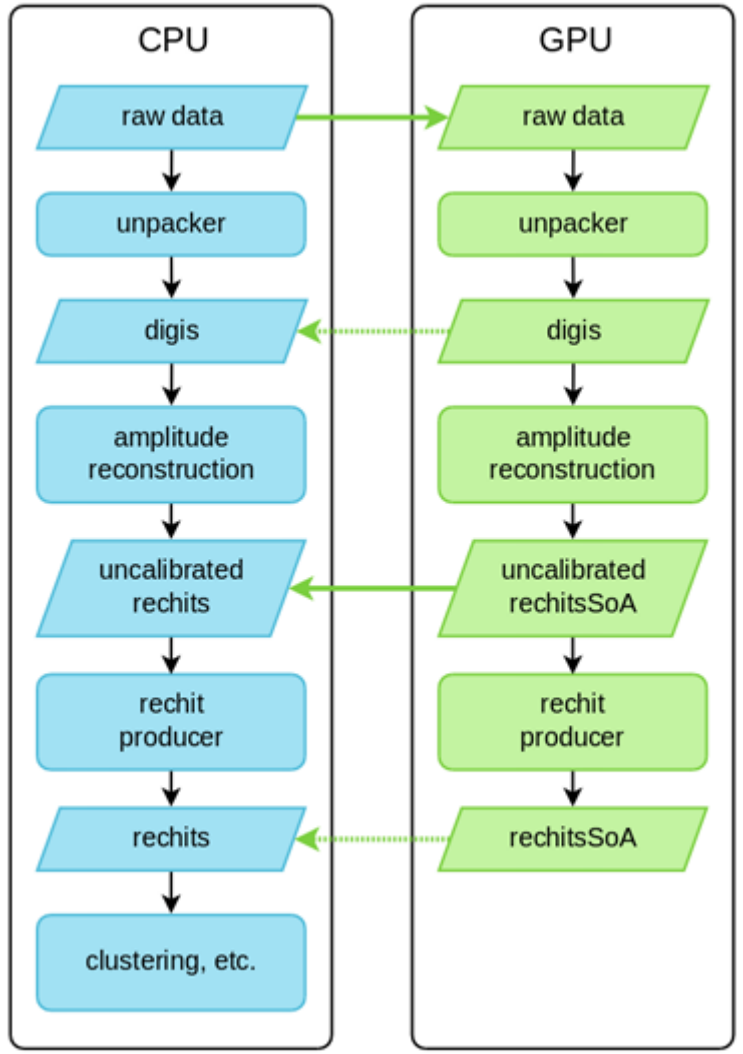
- Room for improvement:
  - Migration from traditional CMS data formats to Structure of Arrays (SOAs) for better utilization of CPUs and GPUs
  - Porting of other pieces of algorithm such as Particle Flow reconstruction to GPUs

- Porting Heterogenous Code to Alpaka (performance portability library) – to reduce dependency on a particular architecture and prevent code duplication is underway [talk by Andrea]

Back Up Slides

ECAL Reconstruction on GPU

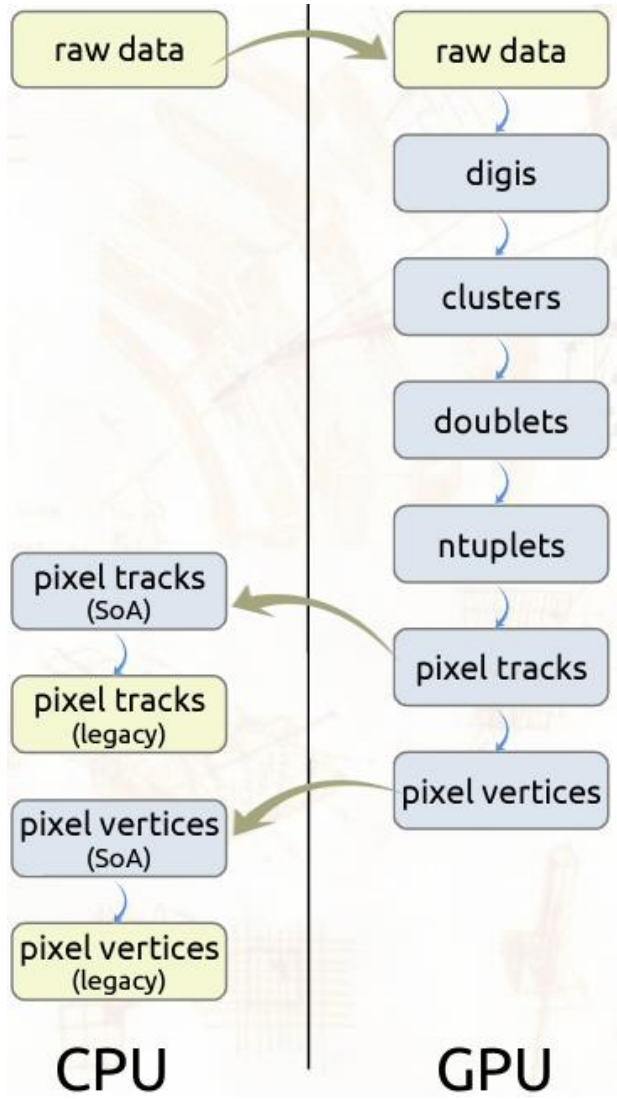
- Unpacking of the raw data from the detector into digitized hits (digis) - packed raw data from the 54 ECAL front end drivers (FEDs) is accumulated on the host in one piece of memory that is then transferred to the device
- Amplitude reconstruction from uncalibrated reconstructed - minimization and amplitude reconstruction is split into several kernels, taking care to avoid computation of common variables
- Calculation of energies from the uncalibrated RecHits by applying various scale factors and calibrations – currently not implemented fully on GPUs



Pixel Reconstruction on GPU

WORKFLOW

- Copying over the RAW data to GPU
- Kernal launch to perform tasks in parallel
  - unpack the raw data
  - cluster the pixel hits
  - form hit doublets
  - form hit ntuplets (triplets or quadruplets)
  - fit the track parameters and apply quality cuts
  - reconstruct vertices
- Move the results back to the host to be used further downstream



Reconstruction using GPUs

CPU

Element	Time	Fraction
AlCa	3.1 ms	0.4 %
B tagging	0.5 ms	0.1 %
CTPPS	0.0 ms	0.0 %
DQM	1.0 ms	0.1 %
E/Gamma	34.8 ms	5.0 %
ECAL	31.4 ms	4.6 %
Framework	0.0 ms	0.0 %
HCAL	57.7 ms	8.4 %
HLT	4.7 ms	0.7 %
I/O	6.0 ms	0.9 %
Jets/MET	14.1 ms	2.0 %
L1T	5.0 ms	0.7 %
Muons	102.4 ms	14.8 %
Particle Flow	46.5 ms	6.7 %
Pixels	295.7 ms	42.9 %
Taus	6.1 ms	0.9 %
Tracking	26.3 ms	3.8 %
Vertices	2.2 ms	0.3 %
event setup	0.7 ms	0.1 %
other	52.0 ms	7.5 %
total	*690.1 ms*	*100.0 %*

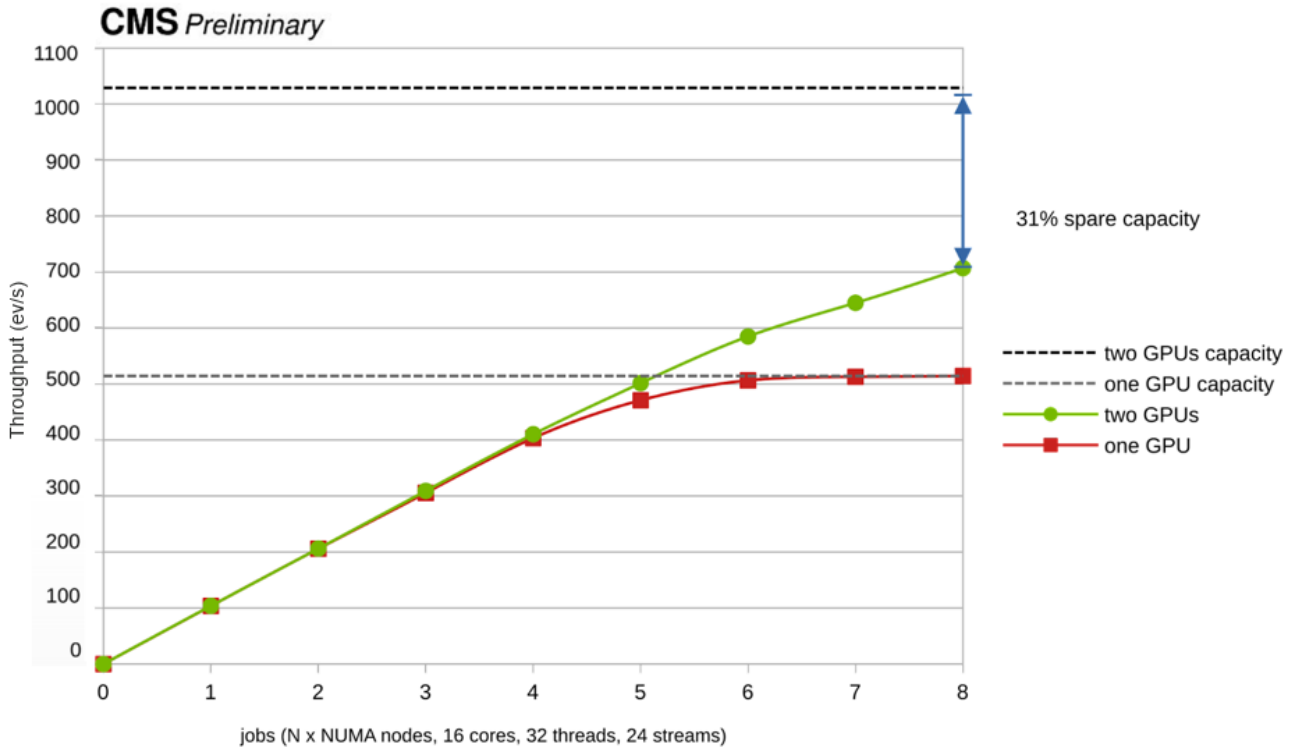
GPU

Element	Time	Fraction
AlCa	3.1 ms	0.8 %
B tagging	0.5 ms	0.1 %
CTPPS	0.0 ms	0.0 %
DQM	1.1 ms	0.3 %
E/Gamma	37.3 ms	9.4 %
ECAL	16.2 ms	4.1 %
Framework	0.0 ms	0.0 %
HCAL	11.6 ms	2.9 %
HLT	5.8 ms	1.5 %
I/O	6.1 ms	1.5 %
Jets/MET	15.2 ms	3.8 %
L1T	5.0 ms	1.2 %
Muons	112.3 ms	28.2 %
Particle Flow	50.9 ms	12.8 %
Pixels	37.3 ms	9.4 %
Taus	6.5 ms	1.6 %
Tracking	28.0 ms	7.0 %
Vertices	2.8 ms	0.7 %
event setup	0.8 ms	0.2 %
other	57.4 ms	14.4 %
total	*397.8 ms*	*100.0 %*

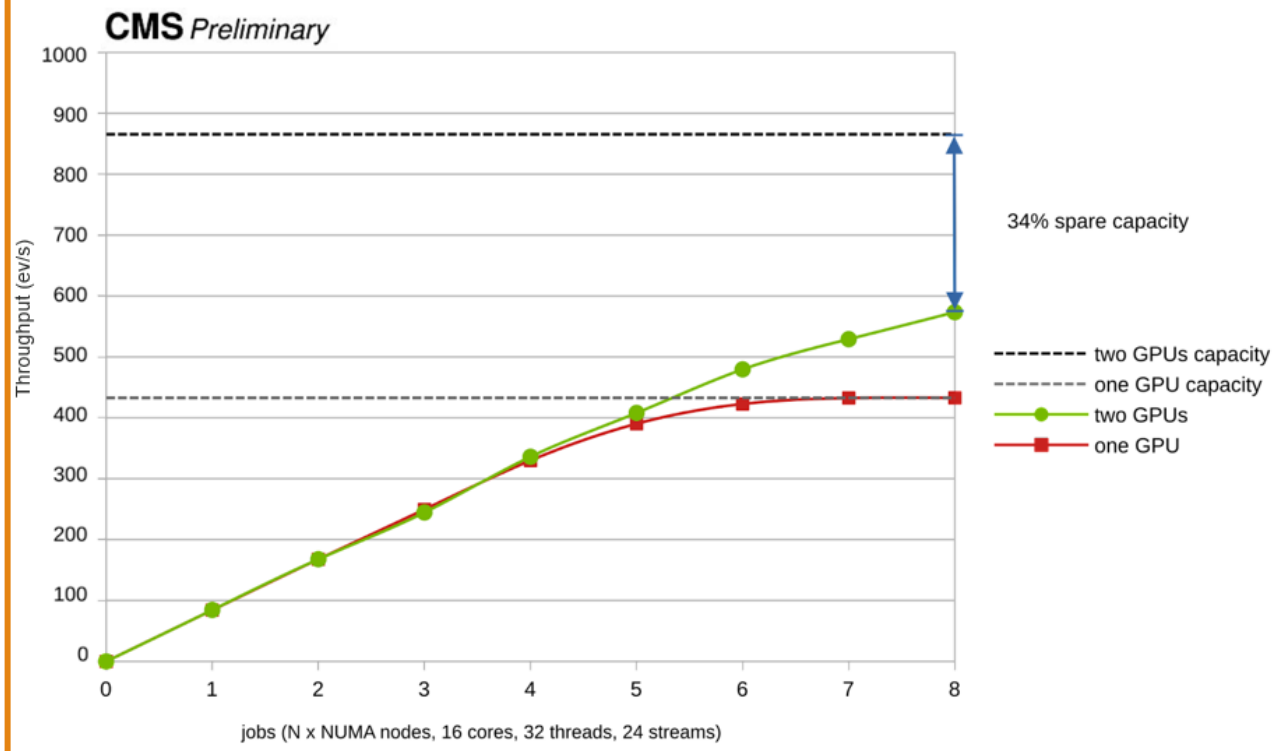
GPU Capacity Estimation

- Estimation of the throughput as a function of HLT jobs (N) running on a single GPU, each job launched with 16 cores, 32 threads and 24 streams
- Finding the maximum throughput with a single GPU
- Another scan of the throughput as a function of HLT jobs (N) with 2 GPUs
- To estimate the spare GPU capacity: Compare the throughput of the fully loaded machine (8 jobs each with 16 cores, 32 Threads and 24 streams running with 2 GPUs) to twice the maximum of single GPU throughput
- Measurements with the HLT menu 2023 v1.0 over L1-skimmed data at pileup 61 (20k events from run 362616) and pileup 70 (30k events from run 362439) show that there is about **30% spare GPU capacity**

GPU Spare Capacity



pileup 61



pileup 70