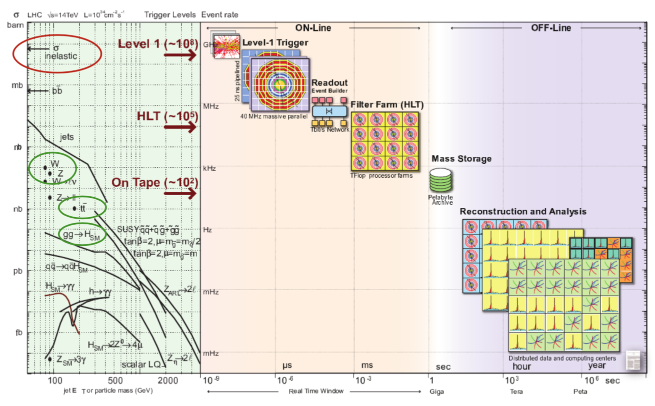
Commissioning CMS online reconstruction with GPUs

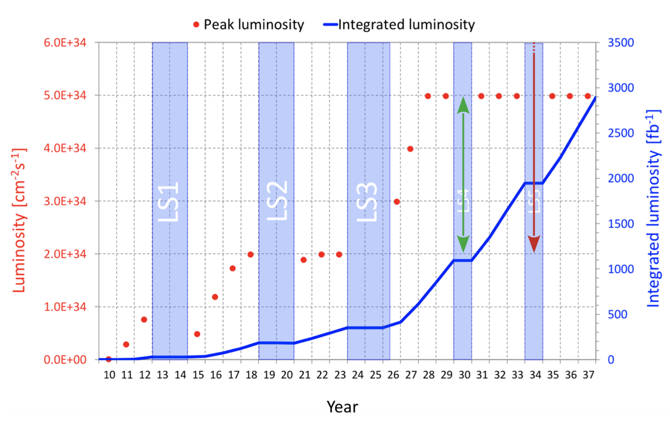
Marc Huwiler on behalf of the CMS collaboration
CERN

The CMS High Level Trigger

- Collisions in the LHC happen at **40MHz**, impossible to save all events
- **Level 1 trigger**: first filtering based on FPGAs and custom electonics reduces the rate to **100 kHz**
- **High Level Trigger (HLT)**: Streamlined version of reconstruction software reduces the rate to about **1kHz** for O(1GB/s) data readout



The High Luminosity challenge



- At *High Luminosity LHC*, the instantaneous luminosity will be increased by a factor 2.5 or more
- Detectors are being upgraded (higher granularity, more readout channels, etc...)
- Need a **factor 30 increase in computing resources** to keep similar physics reach as run 2

GPU reconstruction

- In order to cope with the higher throughput required, and keeping energy consumption and cost under control, CMS has decided to upgrade the trigger farm with GPUs
- Parts of the reconstuction are offloaded to GPUs: **Pixel Tracking**, **ECAL**, **HCAL**
- In addition to releasing the CPU from time consuming tasks, it offers the ability to use global pixel tracks at the HLT **for the first time**

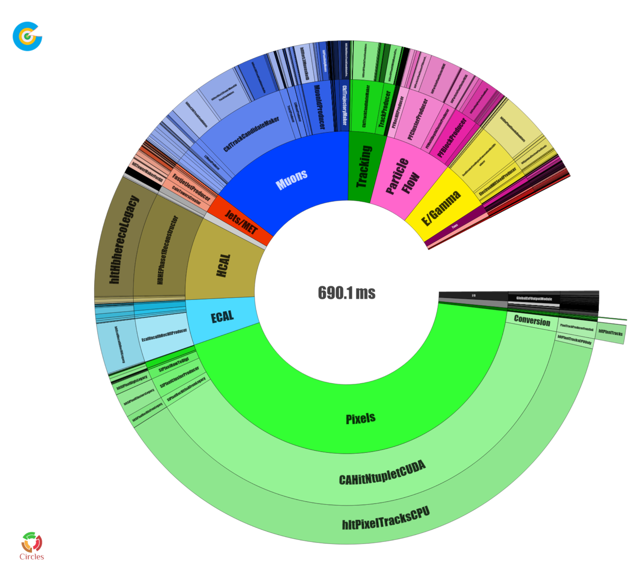
The HLT trigger farm

- The HLT farm consists of 200 machines with 2 sockets, equipped with AMD EPYC 7763 "Milan" 64-core processors for a total of 128 physical cores and 256 hardware threads, as well as 2 low profile Nvidia T4 GPUs
- The Nvidia Tesla T4 has 2560 processing cores running at 1.59GHz, 16GB GDDR6 DRAM and 6MB L2 cache
- The new HLT farm is in service since the start of run 3 (July 4th)
- Currently, the GPU code is running at 90kHz (pixel reconstruction is run on 88% of events, ECAL on 70% and HCAL on 65%)

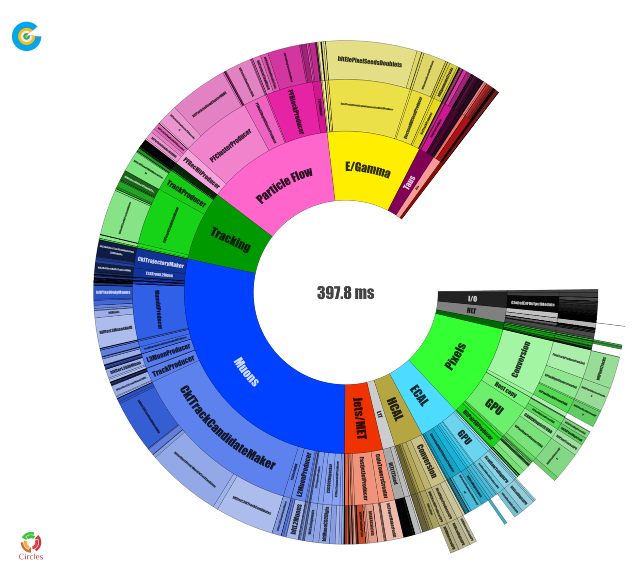
Advantages of using GPUs

- Detector event reconstruction is an embarrassingly parallel problem (data level parallelism), suitable to be run on GPUs
- GPUs offer more throughput for the price
- The energy efficiency of GPUs is increased compared to CPUs, which makes their use a greener and more suitable computing model for energy scarce times

Timing results

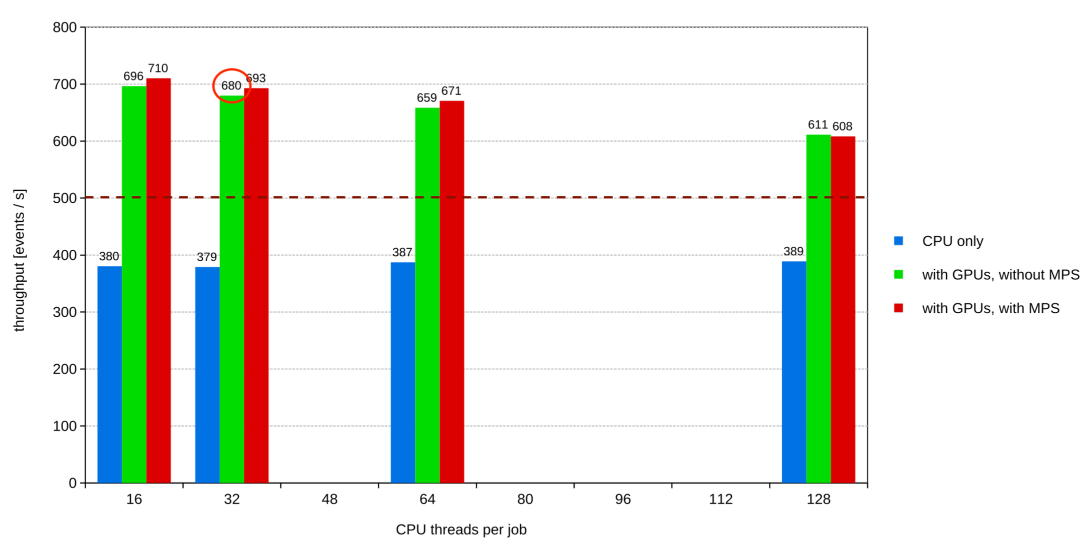


Average time per event on CPU



Average time per event on GPU

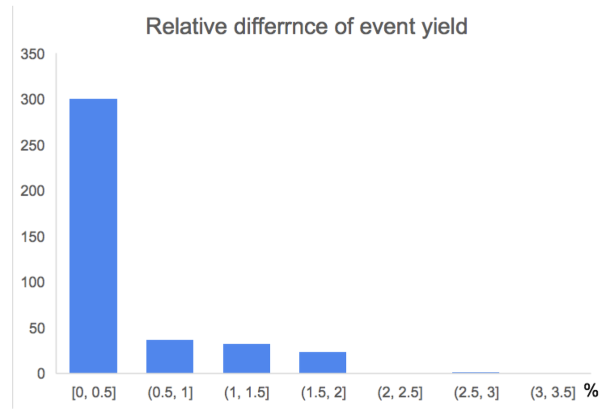
- The **execution time** per event of the HLT step was reduced on average by a **factor 1.7**
- A sizeable fraction of time is spent on **conversion** between SoA and legacy data formats - moving all the algorithms to SoA format gives **room for improvement**
- The speedup translates in increased throughput of the upgraded HLT farm



Plots obtained from pp collisions data at 13.6 TeV in October 2022, with average pileup 55

Effect on trigger paths

- The trigger menu is run with and without GPU reconstruction, and each path's yield compared
- From the ∼ 700 trigger paths in the menu, ∼ 400 show no difference at all
- Looking at paths that accepted >100 events (out of 1'284'337), 99% have a yield difference lower than 2% (plot on the right)



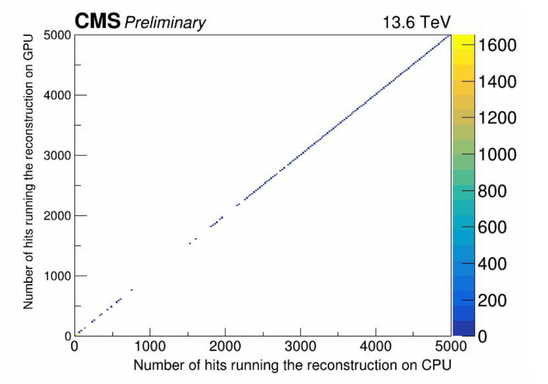
- A dedicated trigger path was added to monitor differences in yield between CPU and GPU after the full ParticleFlow reconstruction, and store these events: **HLT_PFJet40_GPUvsCPU_v1**
- During a run of pp collisions data at 13.6 TeV on 13 October this trigger recorded 5316 events at 0.18Hz, while the corresponding GPU trigger recorded 2'312'690 events, corresponding to a relative difference of:

$$\epsilon = \frac{5316}{2312690} = 0.22\%$$

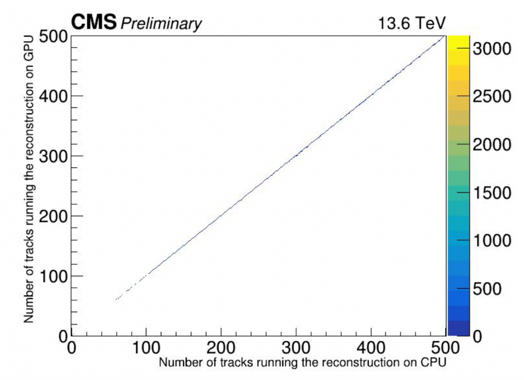
GPU vs CPU reconstruction

- For the validation of the GPU reconstruction, a series of studies were performed to make sure the GPU reconstruction does not introduce regressions
- A few machines of the old HLT farm were equipped with GPUs to take data with cosmics in 2021/2022 and during the 900GeV run in May/June 2022
- A pilot submission to validate the latest pre-release of the reconstruction software with simulated benchmark datasets was launched on grid GPU machines on the Grid
- Event by event comparisons were implemented in the online Data Quality Monitoring (DQM) software. A set of plots with events recorded on a run of pp collisions data at 13.6 TeV on 2nd of October are shown below.

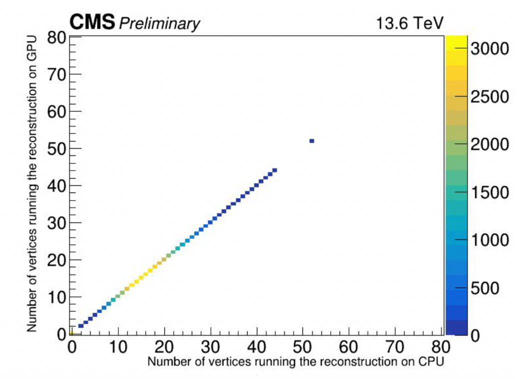
Tracking



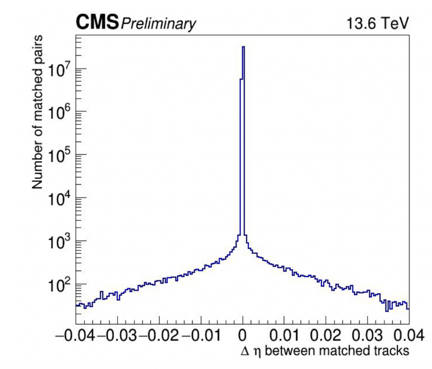
Comparison of the number of reconstructed hits (rechits) in the pixel detector per event



Comparison of number of pixel tracks per event



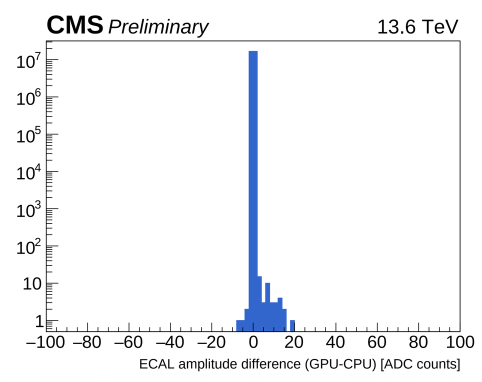
Comparison of number of pixel vertices per event



Difference in η of a track reconstructed on CPU with the track reconstructed on GPU, matched within a geometrical acceptance of ΔR < 0.2

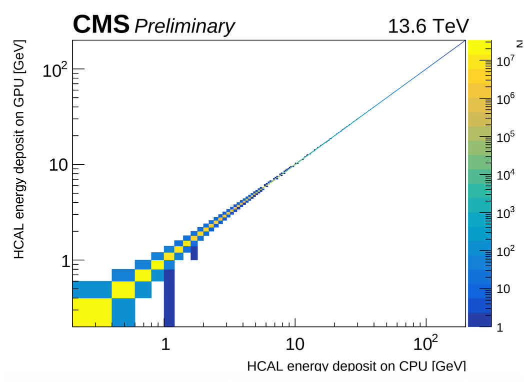
- The number of pixel rec hits perfectly agree
- Mismatch of ∼ 0.1% for the number of tracks
- Amplitude difference ≠ 0 in 10^{-6} fraction of ECAL rec hits
- Good agreement (small differences can be explained by usage of *float* on GPU and *double* on the CPU)

ECAL



ECAL barrel: difference of amplitude of same pulse when the fit is run on GPU and on CPU

HCAL



HCAL barrel + endcap: Energy response of the same energy deposit reconstructed on GPU and CPU

Conclusion

- The CMS trigger is being upgraded with GPUs to get ready for HL-LHC. 2 Nvidia T4 GPUs per machine have already been installed in the trigger farm, and enable to run and commission the GPU reconstruction from the beginning of run 3.
- The throughput has already increased by about a factor 1.7, while some room for further improvement resides in the usage of dedicated data structures throughout the reconstruction and avoiding conversions.
- Additional algorithms are being ported to GPU (e.g. Particle Flow: dedicated poster by Felice on Tuesday)
- The validation shows no significant discrepancy between GPU and CPU results, and residual differences are being investigated.
- In order to reduce dependency on a particular architecture, the heterogeneous reconstruction is being moved to Alpaka for the future (dedicated talk by Andrea on Thursday)