# b-jet identification at High Level Trigger in CMS

**Eric Chabert**

Institue Pluridisplinaire Hubert Curien, Département Recherches Subatomiques, IPHC/DRS,
23 rue du Loess -BP 28- F67037 Strasbourg

E-mail: `Eric.Chaber@cern.ch`

**Abstract.** The CMS experiment at the CERN LHC has been designed with a 2-level trigger
system. The Level 1 Trigger (L1) is implemented on custom-designed electronics. The High
Level Trigger (HLT) is a streamlined version of the CMS offline reconstruction software running
on a computer farm. Being able to identify b-quark jets (b-tagging) at trigger level will play a
crucial role during the Run II data taking to ensure the top quark, beyond the Standard Model
and Higgs boson physics program of the experiment. It will help to significantly reduce the
trigger output rate which will increase due to the higher instantaneous luminosity and higher
cross sections at 13 TeV. B-tagging algorithms based on the identification of tracks displaced
from the primary proton-proton collision and/or on the reconstruction of secondary vertices
have been successfully used during Run I. We will present their design and performance with
an emphasis on the dedicated aspects of track and primary vertex reconstruction, as well as the
improvements foreseen to meet the challenges of the Run II data taking.

## 1. The CMS experiment
The Compact Muon Solenoid (CMS) detector recorded proton-proton collisions occuring at the
LHC during the first period of data-taking (2010-2013), called Run I. A description of the CMS
detector can be found in [1]. CMS uses a right-handed coordinate system with the x-axis pointing
towards the centre of the LHC, with the y-axis directed upwards along the vertical, and with the
z-axis directed along the counterclock-wise beam direction. The azimutal angle $\phi$ is measured
with respect to the x-axis in the $x - y$ plan and the polar angle $\theta$ is defined with respect to the
z-axis. The pseudo-rapidity is defined as $\eta = -ln[\tan(\theta/2)]$.

## 2. b-jet identification: principles and algorithms
Many channels studied in the CMS physics program involve b-jets. Top quark[1], beyond the
Standard Model[2] and Higgs boson[3] physics strongly rely on the capability to identify jets arising
from the hadronization of b quarks.A good performance in terms of b-jet identification efficiency
and rejection of non-b jets are essential to improve the purity and/or the significance of the
selected events. Moreover, the use of b-tagging at trigger level is helpful if not mandory, to
select efficiently the relevant events at trigger level while keeping the background sufficiently
low to accommodate the online processing and bandwidth constraints.

---

[1] Top quarks decay almost exclusively into a W boson and a b-quark
[2] We can mention as example the searches for bottom or top quark partners, W' bosons, ... decaying to b quarks.
[3] H$\to b\bar{b}$ is the dominant decay mode for the Higgs boson

B-tagging algorithms exploit the properties of the B-hadrons such as their large decay lifetimes ($c\tau \approx 450~\mu m$) or the presence of leptons in the final state to separate them from light hadrons (made of u,d,s-quarks). Several algorithms have been developed and optimized from the simplest ones which are cut-based, to more sophisticated ones using multivariate techniques. The Combined Secondary Vertex (CSV) algorithm [2] is currently one of the most frequently used algorithms within the CMS collaboration. It is based on a Likelihood Ratio technique to combine discriminating variables constructed from displaced track and secondary vertex information as well as jet kinematics. The offline b-jet identification efficiency is about 70% for a misidentification probability for light jets of about 1% for jets with a transverse momentum $p_T$ between 80 and 120 GeV/c. This algorithm has also been deployed in an online version at the High Level Trigger during Run I. Both online and offline versions use the same version of the algorithm. The difference comes from the tracks and primary vertices collections that are used as input. The measured performance shows a b-jet efficiency of about 60% for a misidentification probability of about 2%.

## 3. High Level Trigger

The CMS trigger system is based on two consecutive levels. The Level 1 Trigger (L1) is an hardware level implemented on FPGA and custom-designed ASICs. The goal is to select events at maximum rate of about 100 kHz from the 20 MHz incoming rate delivered by the LHC. L1 treats in parallel information coming from the electromagnetic and hadronic calorimeters as well as from the muons chambers. A global decision is taken based on the presence of energy deposits compatible with the presence of photons, electrons, muons, jets or hadronically decaying tau leptons.

The High Level Trigger (HLT) is a streamlined version of the CMS offline reconstruction software running on a cluster of commercial rack-mounted computers which consists of about 13000 processors. The goal is to reduce the rate to about 800 Hz for offline data storage. During Run I, only half of the selected events has been used for prompt reconstruction (within 48 hours) and reconstructed later on for physics purpose. The second half of the rate has been parked and the reconstruction has been delayed by several months. The parked data have been reconstructed during the first period of the LHC shut-down and are used in specific analyses where the gain of lowering the online conditions is substantial.

## 4. b-tagging sequence at HLT during Run I

HLT paths involving b-tagging have been used by many analyses during Run I. Many Higgs boson analyses[4] involving b-jets made the choice of using a trigger path based on the CSV algorithm like the analyses focusing on the Higgs boson decaying into two b-quarks [3][4] or on a Higgs boson produced in association with a Z boson [4]. Similarly the search of an invisible Higgs boson produced with a Z boson decaying into two b-quarks [5] used also b-tagging at HLT. Those examples show the interest for b-tagging at HLT to reduce the bandwidth when selecting hadronic events where the thresholds on the jet kinematics are not sufficient to reduce the background. The sequence used by the above mentioned analyses rely on the b-tagging sequence presented in the following:

- **Jets from the L1 trigger (L1 jets)**
  Before entering in the b-tagging sequence, events have to successfully pass a L1 trigger path requiring two central calorimeter jets with $p_T >$40 GeV/c and $|\eta| <$1.6. Jets using anti-$K_T$ algorithm with R = 0.4 are used.

[4] Higgs boson channels have been studied in the context of the SM or SUSY models like the MSSM.

- **Pixel clustering**
  A regional clustering is performed in the regions of interest determined by the direction of the L1 jets that have fired the L1 trigger.

- **Fast Primary Vertex reconstruction**
  The Fast Primary Vertex (FastPV) reconstruction is a procedure that allows to reconstruct a primary vertex based on pixel cluster information only. No tracks are involved in the determination of the vertex. The goal is to determine the position of the vertex along the beam axis (z-axis). Clusters in the pixel layers must have the same $\phi$ coordinate as the jet. Cluster sizes along the y-axis have to be compatible with the pseudo-rapidity $\eta$ of the jet, *i.e.*, the size increases for jet at high $\eta$. Cluster sizes along the x-axis have to be small to select only high $p_T$ tracks. The selected clusters are projected along the z-axis using the jet $\eta$ direction. The peak in the z projection direction is used to determine the position of the Fast Primary Vertex.

- **Pixel track reconstruction**
  The reconstruction of the pixel tracks is performed using the pixel clusters that are within a distance of 1.5 cm of the FastPV position obtained in the previous step.

- **Pixel Primary Vertex**
  The reconstruction of the primary vertex is performed using the pixel tracks reconstructed in the previous step.

- **Full track reconstruction**
  A regional track reconstruction using all the pixel and strip clusters is performed. The Combinatorial Track Finder (CTF) algorithm [6], based on Kalman filter method, is used. The regional track reconstruction is only performed when the distance between the pixel tracks and the pixel primary vertex is less than 1.5 cm.

- **Application of the b-tagging algorithm: Combined Secondary Vertex**
  A CSV discriminant is computed for each jet based on the information of the pixel primary vertex, and the regional CTF tracks. Then a cut on the discriminator is used to make a decision to select or reject the event.

### 5. Improvements of the b-tagging sequence at HLT for Run II

The b-tagging sequence has been revisited and optimized for the Run II. The main goal is to reduce the execution time of the sequence. This was achieved thanks to a more extensive use of regional reconstruction. The second goal is to improve the performance of the online version of the CSV b-tagging algorithm. This was also achieved thanks to the smart iterative tracking reconstruction and to a better resolution on the position of the primary vertex. The new sequence consists of the following steps:

- **Pixel clustering**
  This step is similar to the 2012 version.

- **Fast Primary Vertex**
  In the newest version, an improvement has been obtained by extending the pseudo-rapidity range for the considered jets ($|\eta| <2.4$), and by using weights for the pixel clusters. The weight of a pixel cluster depends on its size in the x-direction, the cluster charge, $\rho_{cluster}$, the azimuthal angle between the jet and the cluster, $\Delta\Phi_{jet,cluster}$, the difference between the cluster size in the z-direction and the one expected from the knowledge of $\eta_{jet}$. The impact of the use of weights on the efficiency was estimated and is shown in the Figure 1. A gain of around 10% is observed on the efficiency to find the FastPV in a window of 2 cm around the true vertex. For jets with $p_T >80$ GeV, the efficiency exceeds 90%. The use of the end-cap clusters also helps to significantly improve the efficiency for jets with

$|\eta| >1$, leading to an improvement of about 20%. Moreover, the fraction of events where no FastPV is found decreased. Finally, the resolution on the z-coordinate of the primary vertex also improves as shown in Figure 2.

- **Pixel track reconstruction**
  The pixel track reconstruction is now regional only, based on the regions of interest around the jet axis. The constraint applied using the FastPV position remains the same. As one could expect, this speeds up the pixel track reconstruction.

- **Pixel Primary Vertex**
  A great feature of the use of regional tracking is the improved efficiency to find the real primary vertex. The constraint on the FastPV is loose due the relative poor resolution. Using regional track reconstruction suppresses the impact of tracks from pile-up collisions when reconstructing the primary vertex. The resolution on the position of the pixel primary vertex along the z-axis is about 100 $\mu m$.

- **Full track reconstruction**
  An iterative tracking procedure is performed based on pixel and strip clusters. The procedure is composed of three steps. First, high-$p_T$ tracks are reconstructed using pixel track seeds and applying a constraint on the pixel primary vertex. Then lower $p_T$ tracks are reconstructed using pixel triplets as seeds. The third step consists in recovering high $p_T$ tracks seeded by pixel pairs. The use of regional iterative tracking helps to reduce the execution time as it reduces the combinatorics.

- **Primary vertex reconstruction**
  In the current version, an additional step is performed in order to reconstruct a primary vertex using the reconstructed iterative tracks. The resolution obtained on the position of the pixel primary vertex along the z-axis is about 20-30 $\mu m$. The vertex finder algorithm is based on a method using deterministic annealing and referred to as "3D Adaptive Vertex Finder".

- **Application of the b-tagging algorithm: Combined Secondary Vertex**
  A CSV discriminant is computed for each jet based on the information of the primary vertex and the regional iterative tracks. Then a cut on the discriminator is used to make a decision to select or reject the event.

## 6. Comparison of the performance of b-tagging at HLT for Run 1 with the new sequence

The performance of the new sequence has been compared to the previous version. The studies have been performed with top-quark pair-produced events at a centre-of-mass energy of 13 TeV and with an average of 25 pile-up collisions. The Figure 3 presents the performance of the CSV algorithm in terms of b-jet identification efficiency and the misidentification probability for light jets. For an efficiency of 60%, the misidentification probability decreases from 2% to 1.3% with the new b-tagging sequence. The Figure 4 shows the processing time for the HLT path HLT_DiCentralPFJet30_PFMet80_BTagCSV07 using the old and the new b-tagging sequence. The mean processing time is reduced by about 15% and the tail at high values is strongly reduced.

## 7. Conclusion

B-jet identification algorithms have been successfully used at HLT during Run 1 for many analyses, in particular those involving $b\bar{b}$ resonances such as the Higgs boson decaying to $b\bar{b}$. Improvements with respect to the processing time and the performance have been developed during the last year leading to a promising use during Run II. The improvements are mainly due to a multi-step approach based on an intensive usage of regional reconstruction techniques.
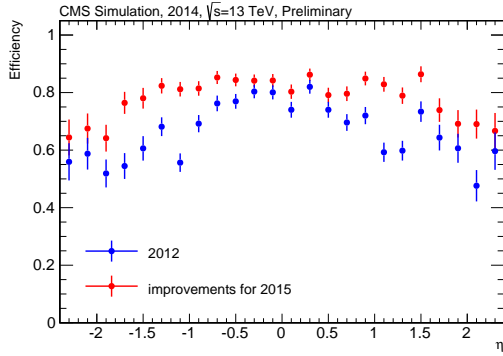
**Figure 1.** The efficiency to find a simulated primary vertex within 2 cm of the primary vertex reconstructed with the FastPV algorithm as a function of the pseudo-rapidity. The blue and red curves show respectively the performance before and after the improvements foreseen for 2015. This result is obtained using simulated $Z(\nu\nu)H(b\bar{b})$ events at $\sqrt{s} = 13$ TeV with on average 60 pile-up collisions.
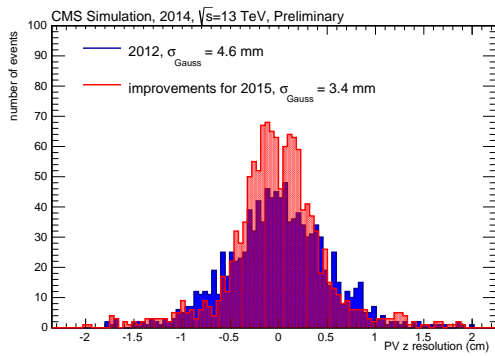


**Figure 2.** Resolution along the z axis of the primary vertex reconstructed with the FastPV algorithm, with respect to the simulated primary vertex (cm). The blue and red histograms show respectively the resolution distribution before and after the improvements foreseen for 2015. This result is obtained using simulated $Z(\nu\nu)H(b\bar{b})$ events at $\sqrt{s} = 13$ TeV with on average 60 pile-up collisions.
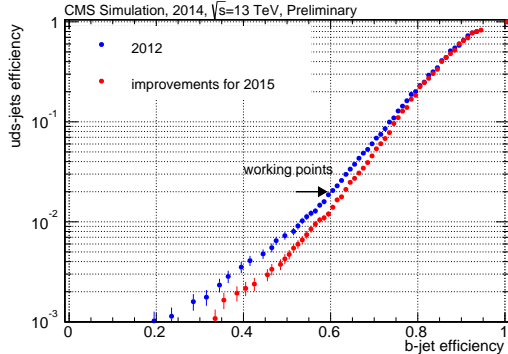


**Figure 3.** Efficiency of the Combined Secondary Vertex (CSV) b-tagging algorithm at HLT for u,d,s-jets vs b-jets. The blue and red curves show respectively the performance before and after the improvements foreseen for 2015. Jets from simulated $t\bar{t}$ events at $\sqrt{s}$=13TeV with on average 20 pile-up collisions and a bunch spacing of 25 ns are used.
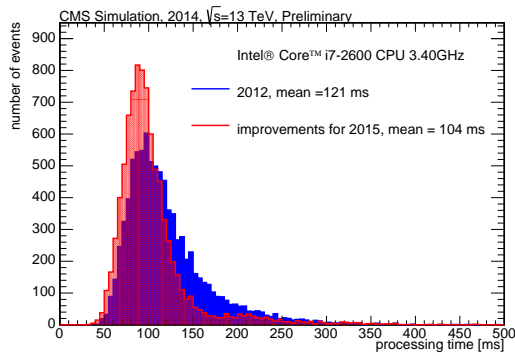


**Figure 4.** Timing of the HLT trigger path HLT_DiCentralPFJet30_PFMet80_BTagCSV07 designed for the 2012 $Z(\nu\nu)H(b\bar{b})$ analysis. The blue and red histograms show respectively the timing distribution before and after the improvements foreseen for 2015. Simulated $t\bar{t}$ events at $\sqrt{s} = 13$ TeV with on average 20 pile-up collisions and bunch spacing 25 ns are used. $Intel\ Core^{TM}$ i7-2600 CPU 3.40 GHz processors are used.

## References

[1] CMS Collaboration 2008 *Journal of Instrumentation* **3** S08004

[2] CMS Collaboration 2013 Performance of b-tagging at $\sqrt{s}$=8 tev in multijet, $t\bar{t}$ and boosted topology events Tech. Rep. CMS-PAS-BTV-13-001

[3] Collaboration C 2010 *Physics Letters B* **722** 207 – 232 ISSN 0370-2693

[4] CMS Collaboration 2013 Higgs to $b\bar{b}$ in the vector boson fusion channel Tech. Rep. CMS-PAS-HIG-13-011

[5] CMS Collaboration 2013 Search for invisible higgs decays in zh with z$\rightarrow b\bar{b}$ Tech. Rep. CMS-PAS-HIG-13-028

[6] CMS Collaboration 2010 *The European Physical Journal C* **70** 1165–1192 ISSN 1434-6044