PROCEEDINGS OF SCIENCE

Raw Data Reduction in the CMS Experiment for Run-3 and Phase-2

Saswati Nandan for the CMS Collaboration [a,*]

[a] *I.N.F.N. – Sezione Di Pisa Largo B. Pontecorvo, N. 3, 56127 Pisa CF: 84001850589 - PI: 04430461006*

E-mail: saswati.nandan@cern.ch

Reducing event and data sizes is critical for experiments at the LHC, where high collision rates and increased detector granularity rapidly increase storage and processing requirements. In the CMS experiment, a recent development to address this challenge is the Raw' format: a new approach for recording silicon strip data in which only the reconstructed cluster's barycenter and average charge are stored, rather than the analog-to-digital converter counts from every strip. This format was successfully deployed online during Run-3 for PbPb collisions at CMS, achieving an event size reduction by nearly a factor of two and enabling CMS to record almost all hadronic minimum bias PbPb collisions. To further enhance Raw', we optimized the number of bits used to encode the cluster barycenter and total charge, using tracking efficiency and resolution as benchmarks. Comparing standard Raw with Raw' shows that refining the bit precision yields stronger compression while maintaining similar performance. Additionally, we introduce a lossless compression strategy that encodes distances between clusters instead of their absolute positions within a detector module. Unlike absolute positions, the distribution of these distances is peaked around zero, effectively reducing entropy of that variable. Consequently, LZMA compression becomes more efficient, allowing even stronger data reduction than the current Raw' algorithms without losing information integrity. Lastly, we discuss projected data sizes for Phase-2 and explore extending these techniques to other CMS detectors, notably the High-Granularity Calorimeter, which is anticipated to generate a substantial fraction of future data.

The European Physical Society Conference on High Energy Physics (EPS-HEP2025)
6-11 July 2025
Marseille, France

*Speaker

PoS(EPS-HEP2025)558

1. Introduction

At the Large Hadron Collider (LHC) at CERN, proton-proton (pp) bunch crossings occur at 40 MHz, generating approximately 50 TB of data, with each event around 1 MB in size. Storing this enormous data is not feasible due to limited storage capacity, and many multijet events with low momentum are of limited interest. To address this, the CMS experiment [1] employs a two-level trigger system for initial event selection, reducing the event rate from 40 MHz to 100 kHz at the hardware-based Level-1 (L1) trigger, followed by the High-Level Trigger (HLT) system, which further reduces the rate to a few kHz.

For the Phase-2 High-Luminosity LHC (HL-LHC), where the pile-up (PU) is expected to reach approximately 200 compared to 60 during Run-3, the CMS detector will undergo significant upgrades [2]. The tracker granularity will be increased, and its pseudorapidity (η) coverage will be extended up to 4. New RPC and CSC chambers, as well as GEM detectors, will be installed. A completely new sub-detector, the Minimum Ionizing Particle (MIP) Timing Detector, will be installed to record the timing of charged particles. The Electromagnetic, and Hadronic endcap calorimeters will be replaced by the High-Granularity Calorimeter (HGCal).

These upgrades, along with the increased granularity, are expected to increase the event size by a factor of 10 [3], further pushing against the bandwidth limit. Consequently, compressing event data without significant loss of physics information is crucial to store a large number of events, which is essential for discovering new physics phenomena.

2. Reduction of strip detector data size in Heavy Ion collision

As shown in Fig. 1, the strip detector provides the largest contribution to the event size. To mitigate this in heavy-ion (HI) collisions, a novel algorithm known as the Raw' approximated-cluster [4] was implemented during Run-3 data taking to save the strip detector data instead of the Raw data format.

In the strip detector, strip clusters are formed when a particle deposits charge across a set of consecutive strips. In pp collisions, the analog-to-digital Converter (ADC) values of each strip in a cluster are stored as a 8-bit unsigned integer, while the position of the first strip in the cluster is stored as a 16-bit unsigned integer. These clusters will be referred to as Raw clusters.

Clusters are saved in a data collection as a flat vector of clusters; for every module with at least one cluster there is an entry in a vector containing the module ID and the offset in the vector of clusters for that module. The module Id is known as detId, which encodes details such as the sub-detector type, module number, and layer or wheel index etc. The detId is stored as 32-bit unsigned integer. The offset variable is called `beginIndices_`, and stored as a 16-bit unsigned integer.

In the approximated method, instead of saving the ADC values for each strip, average ADC value known as average charge ($\frac{\sum_{i=1}^{sizeofcluster} \mathrm{ADC}_i}{sizeofcluster}$) is stored as an 8-bit unsigned integer. Since total charge is larger than average charge, it needs more bits compared to average charge, and hence average charge is stored. Also, the barycenter position which is ADC-weighted strip index is stored as a 16-bit unsigned integer and cluster size is stored as an 8-bit unsigned integer. This reduces the event size for the HI collision by around 50% allowing to store almost all the minimum bias events.

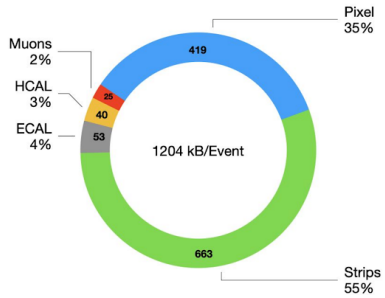

Figure 1: Contribution of each sub-detector to the event size for pp collision at 60 PU scenario. [5]

3. Extension of approximated method

In the approximated method, the barycenter position is stored with a precision of 10% of a strip size, as only one decimal place is retained, and this method will be referred to as HI_raw'.

$$\text{round}(\text{ barycenter's value } \times 10) \tag{1}$$

The data size is controlled by the precision of each variable. The method for storing the barycenter value is generalized in Eq. 2 [6], where the precision of a variable is determined by the number of bits used to encode it. This formula is applied to store both the barycenter, and the average charge variables.

$$\text{round}\left(\text{variable's value} \times \frac{2^{\text{maximum bit}} - 1}{\text{variable's maximum value}}\right) \tag{2}$$

The approximated method is further extended by storing the difference of the barycenter position with respect to the previous cluster rather than the absolute barycenter value. Similarly, the values in the list of detIds is stored as a difference with respect to the previous detId. For each module, the length (in number of strips) of previous module is added to the barycenter difference for the first cluster in the module. In that case, the variable `beginIndices_` becomes redundant and has been removed. The advantage of saving differences is that these variables take much smaller values than the corresponding absolutes, leading to distributions that are sharply peaked at small values, as shown in Fig. 2. This reduces the entropy of the variable. During compression, the most frequent values can be encoded with shorter codes, while less frequent values are assigned longer codes, thereby reducing the average code length and increasing the overall efficiency of the compression algorithm. This extended approach is referred to as the v2 method in the following discussion.

4. Effect on cluster and track reconstruction

For regular cluster analysis, and track reconstruction, the approximated-clusters are converted to the traditional (Raw) cluster. The Raw clusters are matched to those from the v2 method. Clusters are considered matched if they have the same size, and the clusters with the closest barycenters are considered as matched one. The barycenter resolution for matched clusters is shown in Fig. 3.

Tracks are reconstructed separately using Raw clusters and v2 clusters. After reconstruction, tracks are selected with the normalized $\chi^2 < 2$, p_T resolution (p_T error / p_T) $< 10\%$, and total number of hits within tracker > 11. Tracks are then matched based on the closest ΔR separation with

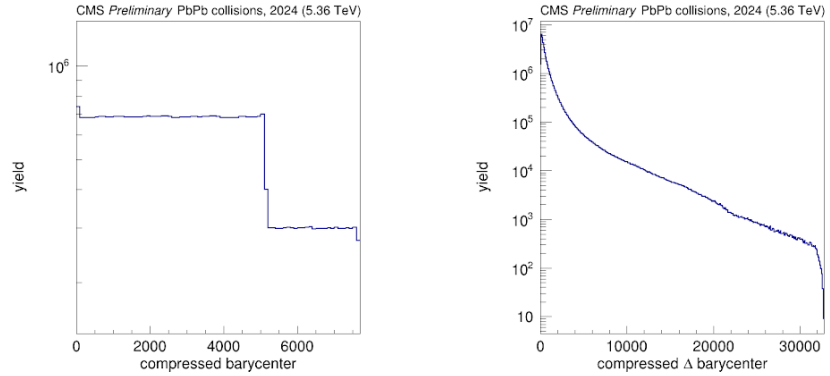

Figure 2: Barycenter distributions in HI_raw' (Left plot), and v2 (Right plot). In both cases, the values are rounded to the nearest integer, and are therefore referred to as compressed. The discontinuity observed in the left plot arises from the fact that some modules contain 512 strips while others contain 768 strips [6].

maximum cut of 0.05. Here ΔR is defined as the distance in (η, ϕ) space:$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$. The relative p_T difference of matched tracks is also shown in Fig. 3. Both the barycenter and track p_T matching distributions are sharply peaked at zero, demonstrating the robustness of the v2 method in preserving cluster positions and track momenta.

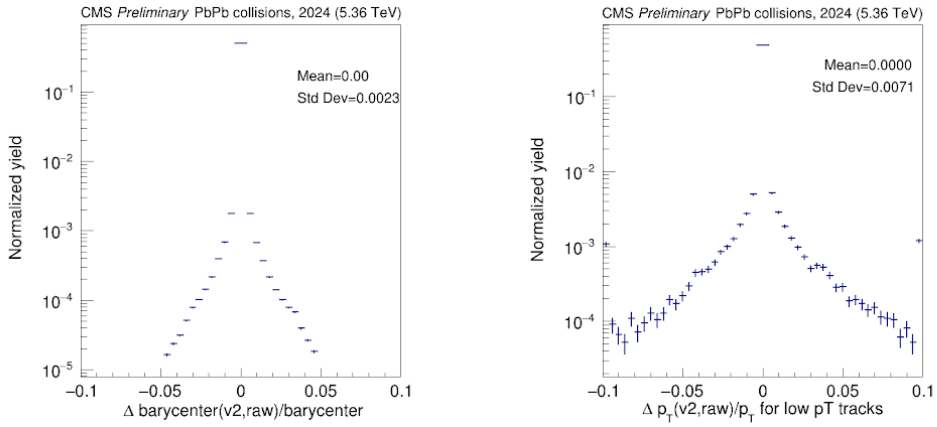

Figure 3: Resolution distributions for matched clusters and tracks using the v2 method. The left plot shows the resolution of matched barycenter positions, while the right plot shows the resolution of matched track p_T. The X-axis represents the resolution of either the barycenter position or the track p_T, and the Y-axis shows the normalized yield. An upper p_T cut of 1 GeV is applied on the right plot [6].

5. Tight charge cut on strip cluster

At the HLT stage, strip clusters with a signal-to-noise ratio above a given threshold are retained in order to suppress contributions from detector noise [7]. Starting from Run-2 LHC operations with 25 ns bunch spacing, HLT and full/offline tracking algorithms apply a minimum cluster charge requirement to remove clusters originating from pp interactions in an earlier bunch crossing (out-of-time pileup or PU) [8]. A loose (1620 ADC/cm) and a tight (1945 ADC/cm) working points are commonly used. Here we propose to apply the tight working point. The impact of this cut is

illustrated in Fig. 4. For HI collisions, which occur under nearly zero-PU conditions, the effect of this requirement is negligible. In contrast, for pp collisions, the cut reduces the strip cluster data size by approximately 13%, with minimal impact on track reconstruction performance.

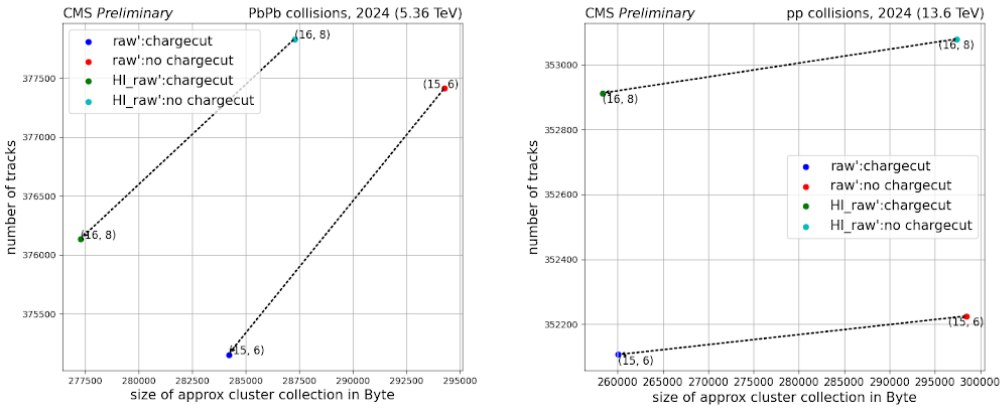

Figure 4: Effect of the tight charge cut on strip clusters at the HLT stage. The left plot shows results for HI collisions and the right plot for pp collisions. The X-axis corresponds to the size of the strip cluster collection in byte, while the Y-axis indicates the number of selected tracks reconstructed from clusters passing the tight charge cut. The numbers in parentheses denote the maximum number of bits used to encode the barycenter and the average charge, respectively [6].

6. Comparison among different configurations

A comparison among different methods is shown in Fig. 5. The reduction in strip detector data size is evaluated for different choices of bit configurations used to encode the barycenter and average charge values in each method. The standard deviation of the relative transverse momentum residual is used as the benchmark to determine the optimal configuration. The configuration with 15 bits for the barycenter, and 6 bits for the average charge value in the v2 scheme achieves more than 10% reduction of strip cluster collection data size in both HI and pp collisions relative to the current HI_raw' method with slight improvement in track reconstruction.

7. Conclusions and outlook

The CMS experiment is entering the HL-LHC era, where the event size is expected to increase by about 10 times due to higher pileup conditions, and increased number of detector channels. In HI collisions during Run-3, dedicated measures have been developed to reduce event size by compressing the strip detector data. This is achieved by storing the barycenter position, the average charge, and the cluster size of the strip cluster. This approach has been further extended by applying tight charge cut on strip cluster, saving the difference of the barycenter position (instead of absolute values), encoding the barycenter difference, and average charge using fewer bits compared to the previous scheme. With these improvements, overall strip detector data size has been reduced by approximately 23% for pp collisions and 16% for HI collisions compared to the default raw configuration. Additional reductions are anticipated by exploiting unused bits of barycenter difference, and average charge to encode cluster shape properties. This method will be implemented for HI data-taking in 2025. In Phase-2, the strip clusters will be entirely binary, indicating only

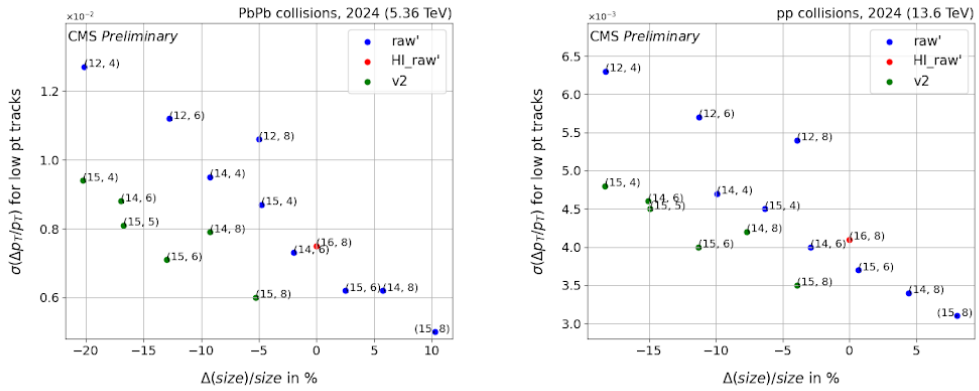

Figure 5: This plot compares different bit configurations, and encoding methods. The two numbers in parentheses indicate the maximum number of bits used to encode the barycenter (or its difference), and the average charge, respectively. The X-axis shows the relative reduction in strip detector data size with respect to the current HI_raw' method, highlighted in red. The blue points correspond to the Raw' method, where the barycenter position is stored directly, while the green points correspond to the v2 method, where the difference of the barycenter position and the difference of detId are encoded. In all cases, a tight charge cut on strip clusters is applied at the HLT stage [6].

whether a hit occurred on a strip. The barycenter position, and detId can be stored using a method similar to v2. In Phase-2, the maximum contribution to the event size comes from the HGCAL detector [3]. The same compression strategy could be extended to the HGCAL, and also other sub-detectors for Phase-2. This work has been supported by the European Union-NextGeneration EU Mission 4 Component 1 CUP I53D23001070006

References

[1] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST*3(2008)*S*08004.361*p*

[2] CMS Collaboration, "Technical Proposal for the Phase-II Upgrade of the CMS Detector", CERN-LHCC-2015-010 ; LHCC-P-008 ; CMS-TDR-15-02

[3] CMS Collaboration, "The Phase-2 Upgrade of the CMS Data Acquisition and High Level Trigger" CERN-LHCC-2021-007;CMS-TDR-022

[4] CMS Collaboration, "Data compression via partial online processing in CMS: experience in heavy ions and prospects", CMS-DP-2024-007 ; CERN-CMS-DP-2024-007

[5] Next Generation Triggers 1st Technical Workshop, 25–27 Nov 2024 CERN

[6] CMS Collaboration, "Data Reduction of the strip detector of the CMS Experiment for heavy ion collisions and proton-proton collisions", CMS-DP-2025-031 ; CERN-CMS-DP-2025-031

[7] CMS Tracker Collaboration, TRK-20-002, JINST 20 (2025) P08027

[8] M. Rovere, "CMS reconstruction improvements for the tracking in large pileup events", CHEP 15 proceedings