Data Compression via Partial Online Processing in CMS: Experience in Heavy Ions and Prospects

Yu-Chen (Janice) Chen

Massachusetts Institute of Technology

On behalf of the CMS Collaboration

156th LHCC Poster Session, November 2023

CMS-DP-2023-X (to appear)

Motivation

Challenge

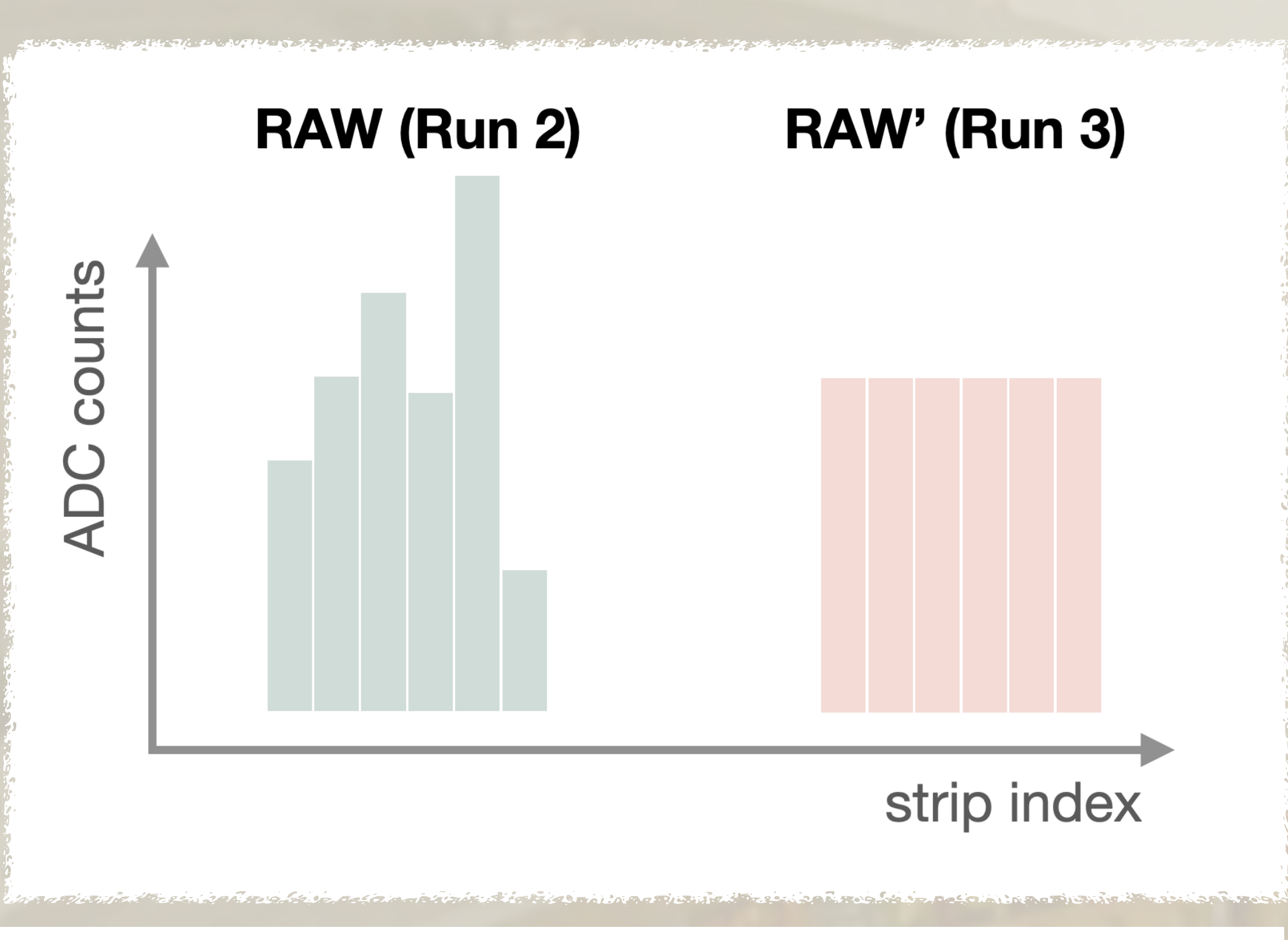
- Raw event data size in high luminosity pp and PbPb collisions at LHC
- Due to the large number of channels, the silicon strip tracker (SST) output is the limiting factor of the DAQ throughput

New idea

- Processing strip clustering algorithm on the global SST online
- Approximating properties of the online SST clusters
- Storing out in a compressed format of the approximated cluster properties + SST detector/cluster auxiliary info

Approximated SST clusters — “ RAW’ ”

- PbPb Run2 at 5.02 TeV (2015-2018): SST info is stored as per-strip ADC counts in raw data
- PbPb Run3 at 5.36 TeV (from 2023): new **RAW’** data format is deployed
  - Implemented in the High Level Trigger
  - Rectangular cluster-amplitude approximation, in place of the original per-strip-ADC-count data format:
    - Barycenter: amplitude-weighted strip-index center (10% strip’s width precision)
    - Size: the length of the cluster’s strip sequence (exact info from original cluster)
    - Average charge: average amplitude of the strip sequence (integer precision)
      *Total charge has the precision of (cluster’s size) * (integer precision)*
    - Booleans for the strips’ amplitude saturation and the cluster shape peak filter
  - A list of modules associated to Front End Driver in error state is stored on the event basis



SST cluster format	RAW			RAW’	
Stored content	Strip index		ADC counts (8-bit int)	Approximated cluster properties	[Event-basis] FED modules & readout error info
Example stored tracker data	**First strip** 123 **(16-bit int)**		75	• **Barycenter** = 125.5 **(10-bit int)** (We store 10x barycenter as int) • **Size** = 6 **(6-bit int)** • **Average charge** = 100 **(8-bit int)** • **Cluster shape:** **(1-bit Boolean)** • **Saturated strip** • **Peak filter**	
	(derived by first strip & ADC list)	124	103		
		125	127		
		126	94		
		127	161		
		128	42		
Example total bits per cluster	16 + 8*6 = 64 bits			10 + 6 + 8 + 1*2 = 26 bits + smaller FED error contribution	

An example and sketch of RAW & RAW’ SST cluster data format

Datasets & selections

(in this analysis)

RAW’ comissioning datasets

- Original & RAW’ datasets taken in Sept.-Oct., 2023
- Cluster datasets: 600 matched events, 36M clusters
- Track datasets: 7.7K minimum-bias triggered events, 4.7 M good tracks

Track selections

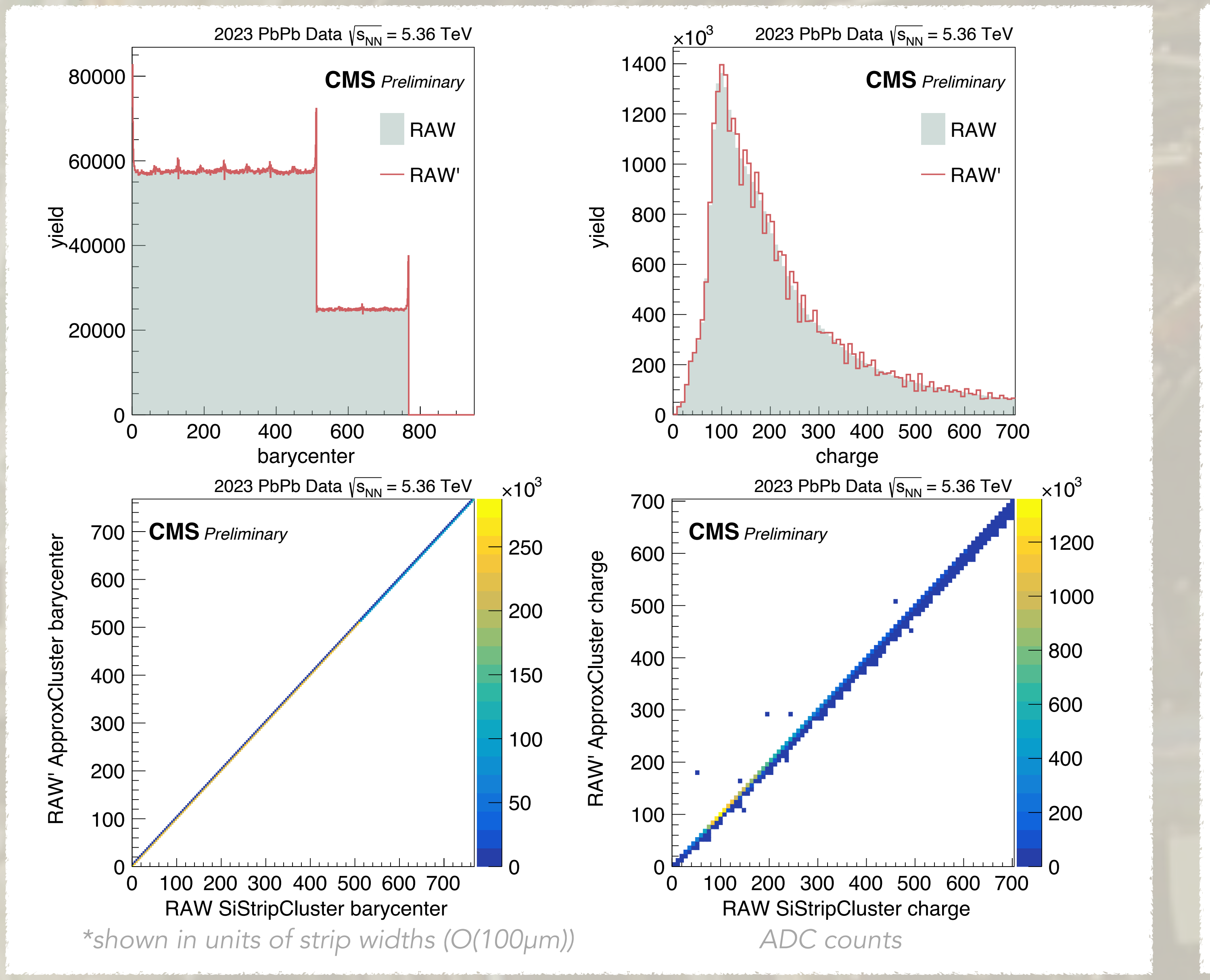
(Similar criteria as Run2 analyses)

- #(hits) ≥ 11
- $\sigma(p_T)/\ p_T < 10\%$
- Normalized $\chi^2 <$ 0.18*#(SST layers)
- $|DCA\ z/\sigma(DCA\ z)| < 3$

(DCA is the distance of closest approach between the primary vertex and the track trajectory)

SST cluster property validations

- Good agreement on cluster barycenter & charge btw original RAW v.s. RAW’ data format!



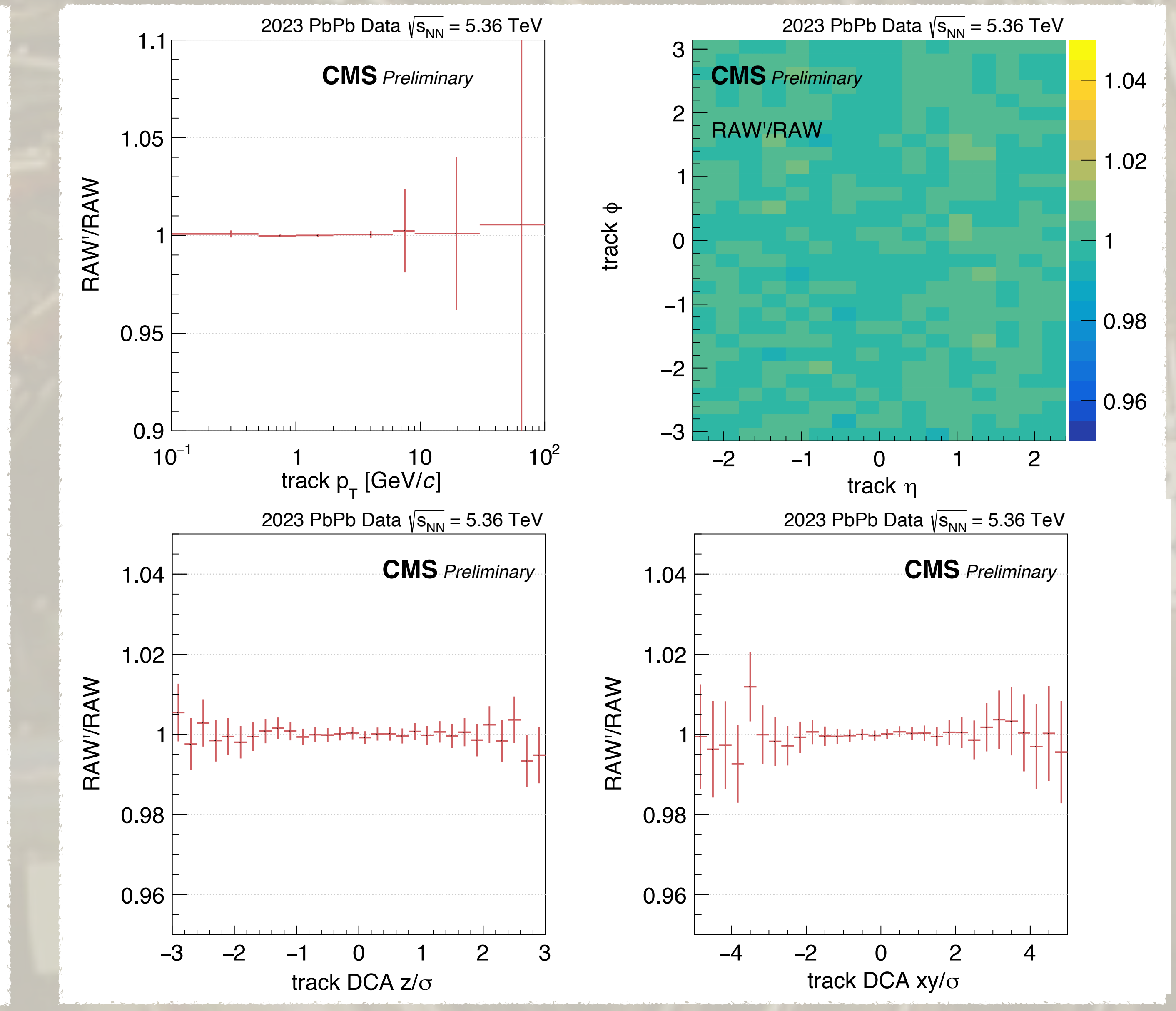
**shown in units of strip widths (O(100μm))*

ADC counts

- Preserving barycenter accuracy within a deviation of 10% the strip’s width
- Outliers in the cluster charge scatter plot are impacts from noisy and dead SST channels

Performance check on tracks

- An agreement better than 2% btw RAW & RAW’ is achieved!



** The uncertainties shown in the plots are statistical errors*

** Analysis-level track selections are applied*

Performance gain & Summary

- Minimum-bias event’s size w/ different cluster formats & compression schemes

Cluster format	Event size
RAW	1.2 MB
RAW’ ZSTD-3	0.77 MB
RAW’ LZMA-4	0.55 MB

- Leading to a substantial reduction of the overall raw event size and a comparable increase in the throughput
  - RAW’ SST cluster approximation reduces **35%** of event size
  - Along with LZMA-4 compression scheme, yielding a **54%** reduction performance
    ⇒ **Doubling** the capacity for minimum-bias events data-taking