ML-based Trigger Algorithm Design & Deployment on FPGA for *b*-hadron Event Selection in sPHENIX

Yaozhong Chen, Kai Chen
Nuclear Electronics Lab (CCNU)

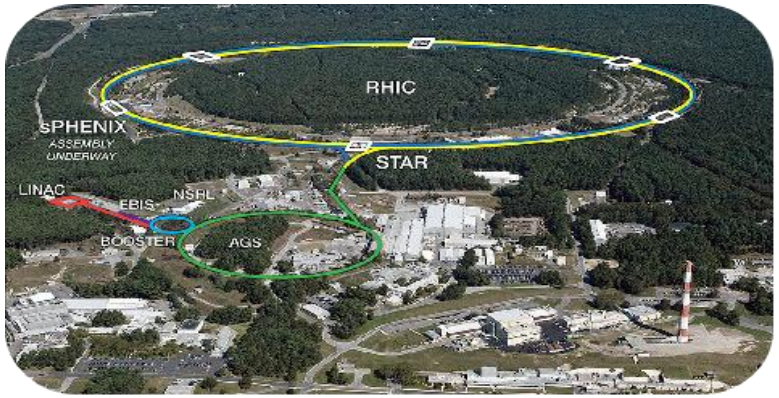
第二十三届全国核电子学与核探测技术学术年会
2026年8月9-12日 登封

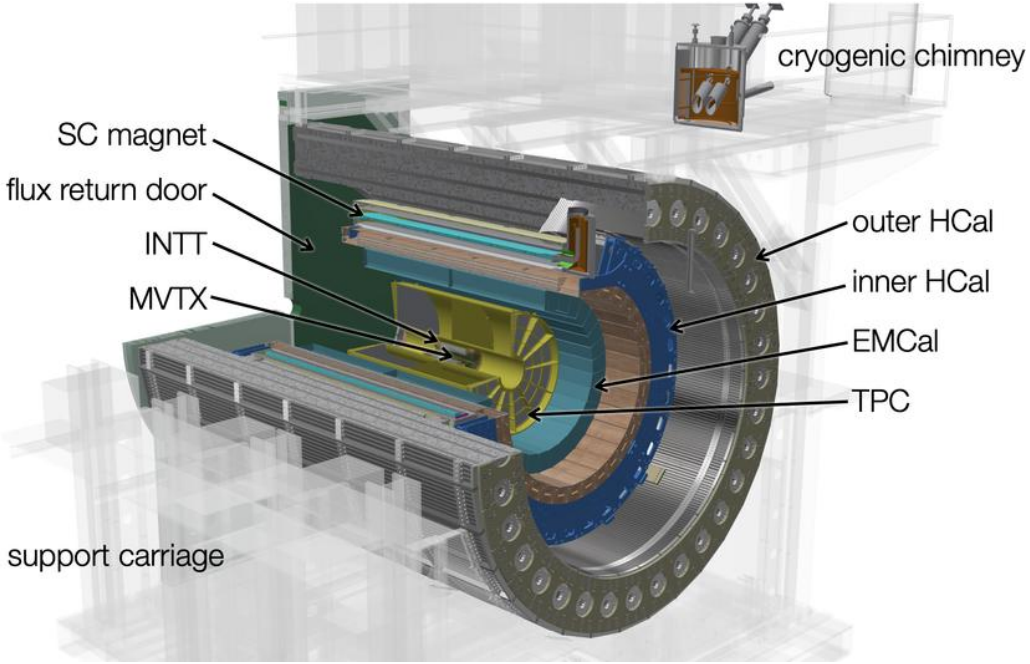
Outline

- Background & requirements
- Algorithms design
- Deployment on FPGA
- Summary

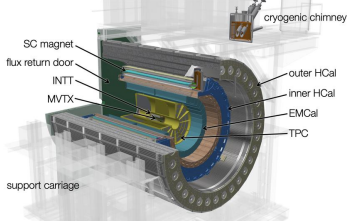
sPHENIX expriment

- Located at RHIC (USA)
- Running period 2024-2026
- Main detectors: tracking detectors (MVTX, INTT, TPC), calorimeters (EMCal, HCal)
- Hybrid trigger scheme
  - Tracking detectors support streaming readout
  - Calorimeters readout is trigger-based: 15kHz event rate

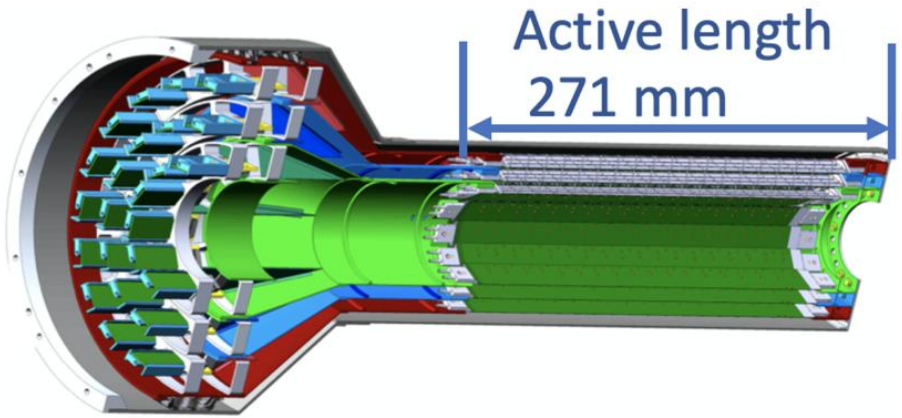


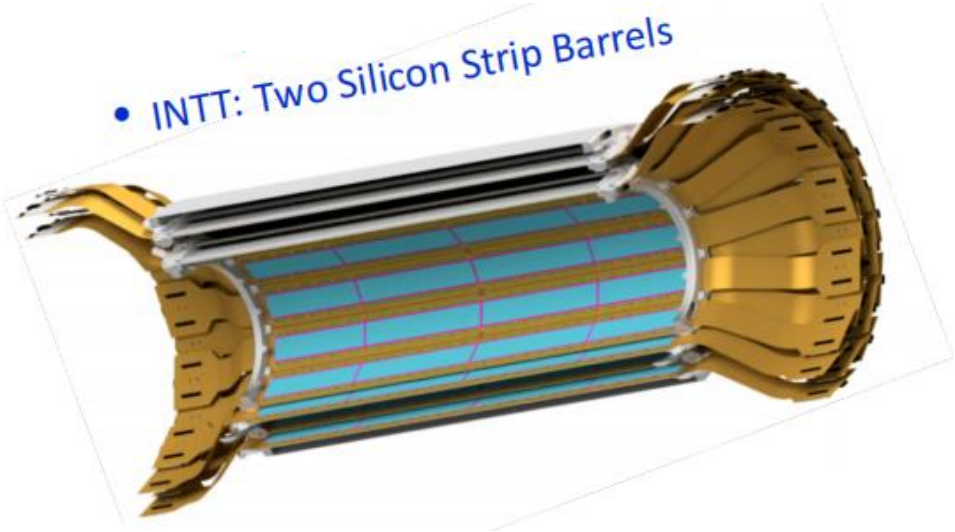


MVTX and INTT



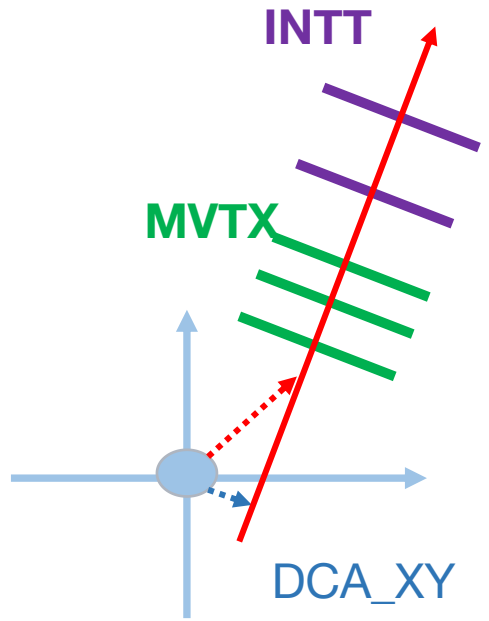
- MVTX
  - Based on ALICE ITS IB detector design
  - Pitch: 27 μm × 29 μm
  - Time resolution: 6-10 μs
  - Active area: ~1,685 cm²
  - 3 layers, 48 staves, 9 chips per stave: ~230M channels
- INTT
  - Silicon strip detector
  - Pitch: 27 μm × 16 (or 20) mm
  - Time resolution: ~ 100 ns
  - 2 layers, 56 ladders



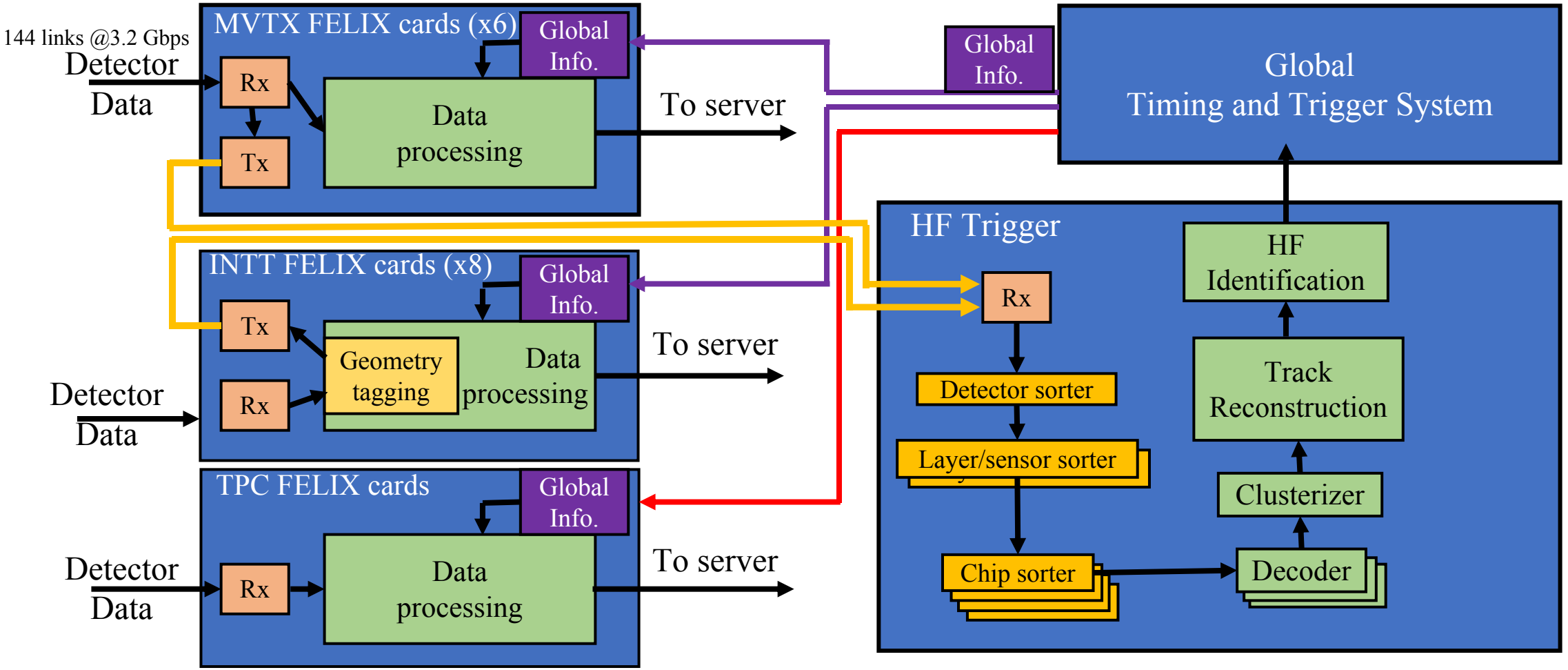


Motivation - Heavy Flavour

- p+p collision rate @200 GeV: ~2 MHz
  - Beauty quark: ~60 Hz
    - $1\ \mu b/42\ mb$: ~0.003%
  - DAQ trigger rate: ~15 kHz
- Tracking detectors are Streamed Readout (SRO) capable
  - But the readout and DAQ can't save all TPC data
  - TPC working in trigger mode, supports > 50 kHz
- A real-time ML trigger system for b-quark is essential to improve its trigger efficiency
  - from about 0.75% to > 90%



Overall diagram for *b*-tagging



The latency constrains for ML-base algorithm

- The TPC buffer can hold up 20 us of data
- Detector readout delay, fiber transmission delay, data encoding/decoding
  - MVTX readout window ~8 us
  - Interaction Region (IR) ->Counting house ~0.3 us (100 m cables)
  - FELIX data forward, decoder buffers ~0.6 us (@240 MHz)
  - Global level 1 Trigger decision latency + counting house -> IR ~0.3 us
  - ...
- The goal is to achieve **10 us** latency for the trigger algorithm

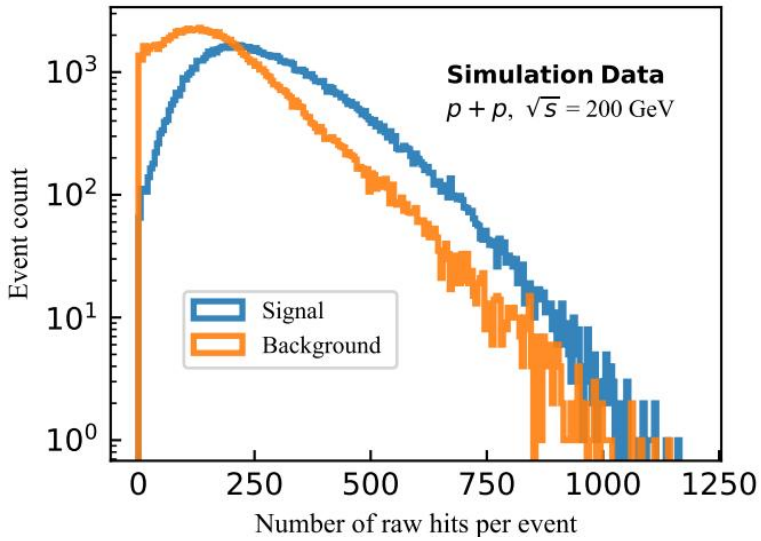
Outline

- Background & requirements
- Algorithms design
- Deployment on FPGA
- Summary

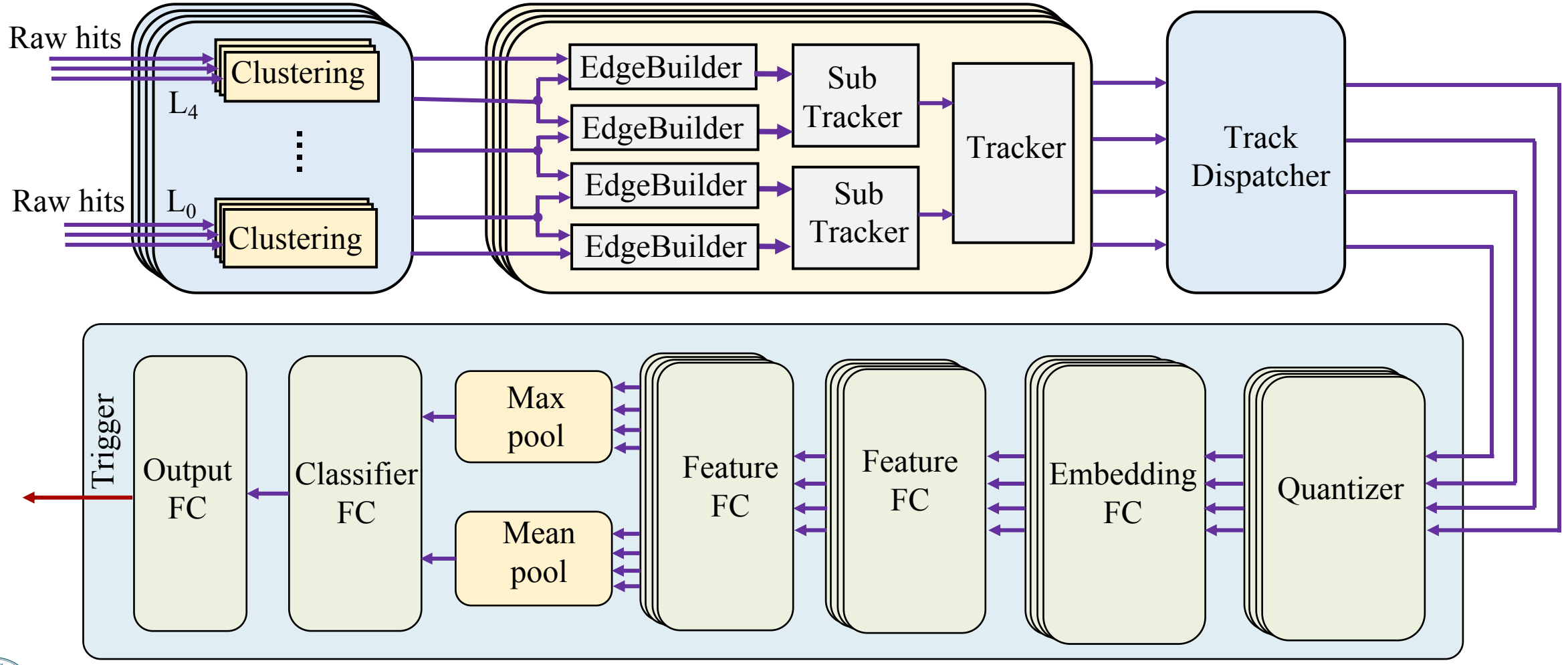
Simulation dataset

- **Pythia8** + **Geant4** for simulation data generation
  - Introducte some: random noise for tracking detectors; fake raw hits
- Features of simulation data
  - **Sparse distribution** in tracking detectors
- The number of hit per event is comparable between simution data and experimental data

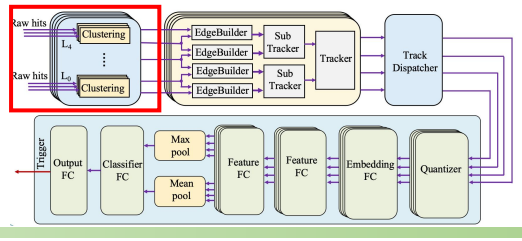
Collision System	p+p
Collision Energy	200 GeV
Event Generator	PYTHIA8
Signal Channals	p+p $\rightarrow b\bar{b} \rightarrow$ B-hadron $\rightarrow X$
Data Size	200,000
Signal : Background	1:1



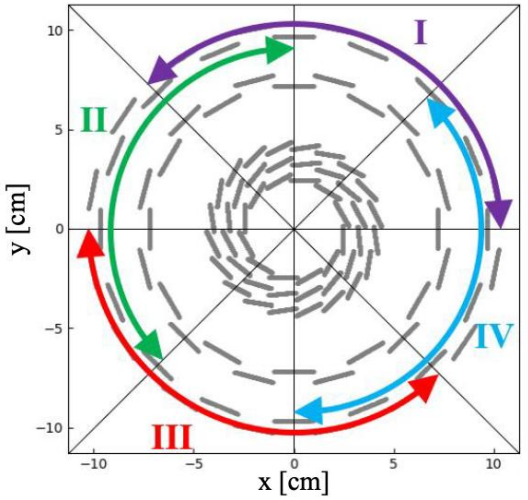
ML Trigger Pipeline

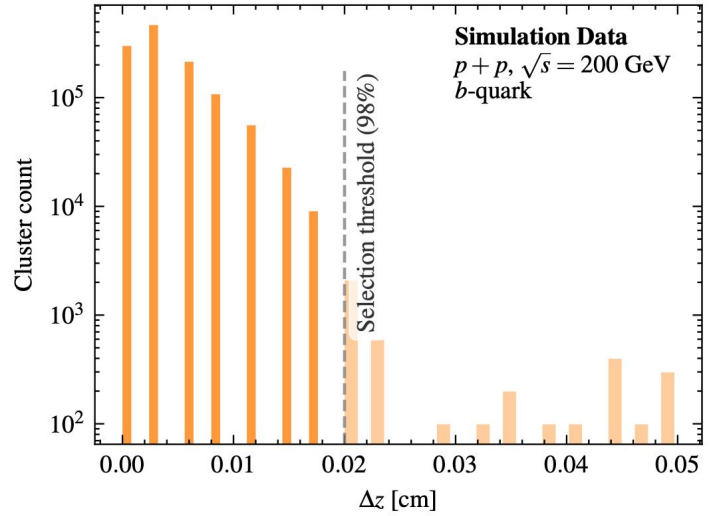


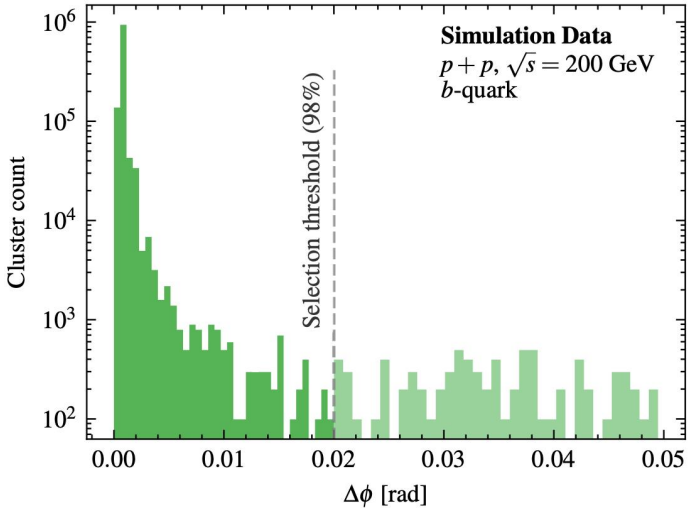
Clustering

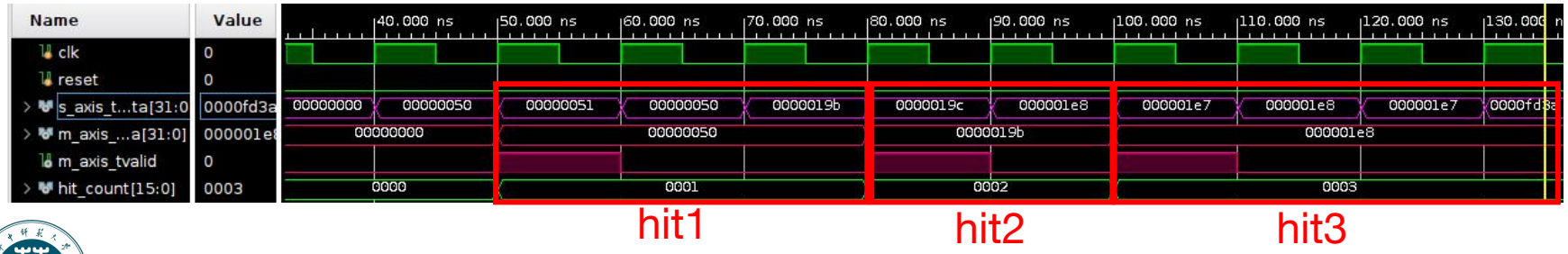


- Independent raw hit streams
  - Stave/Ladder-based readout
- Parallel architecture
  - 8 detector sectors, 4 processing groups
  - 15 clustering engines per group
- Geometry-based clustering
  - z–ϕ spatial constraints
  - Low latency & high throughput

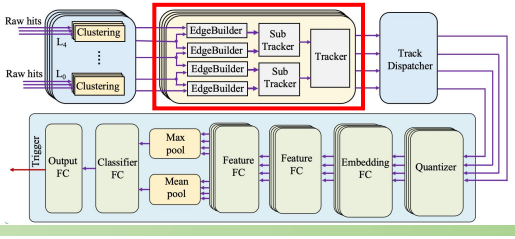




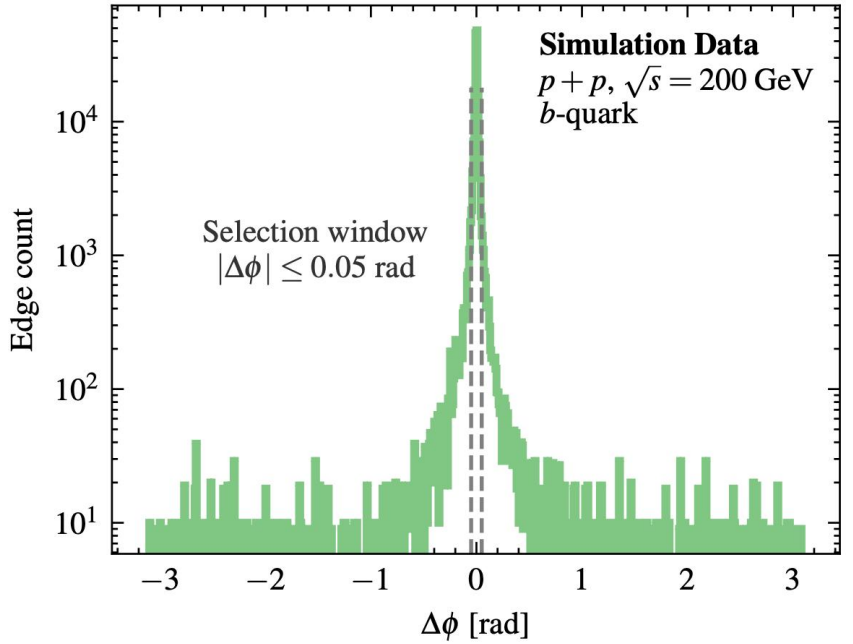


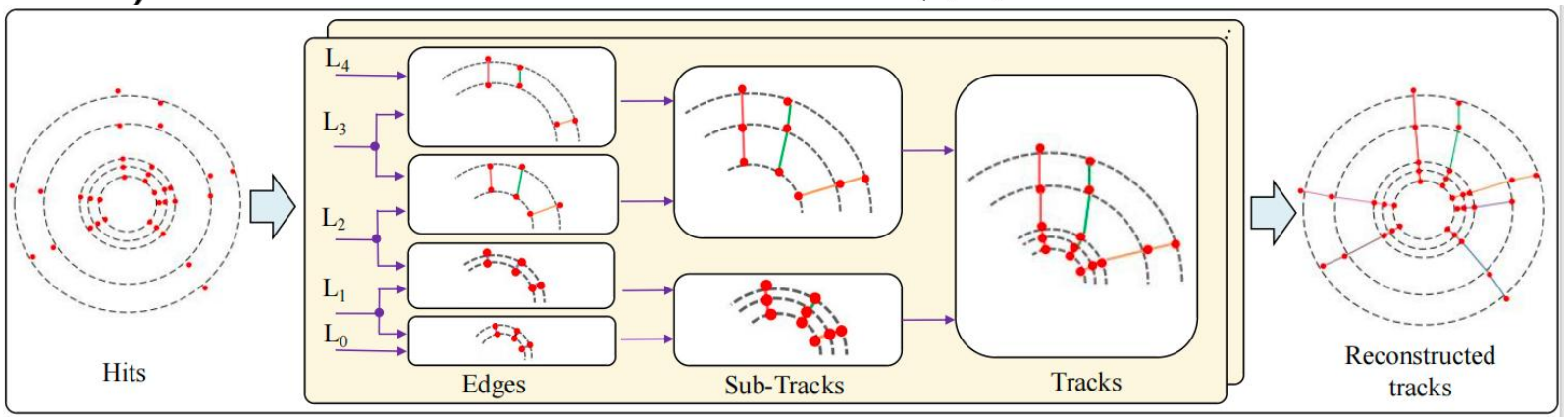


Track reconstruction

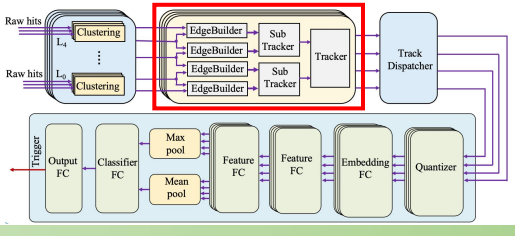


- ϕ-based geometric constraint
  - Strong inter-layer spatial correlation
- Threshold-based edge construction
  - Simple comparison-based algorithm
  - FPGA-friendly streaming design
- Parallel track reconstruction
  - Data stream replication(L1–L3)
  - Parallel edge builders

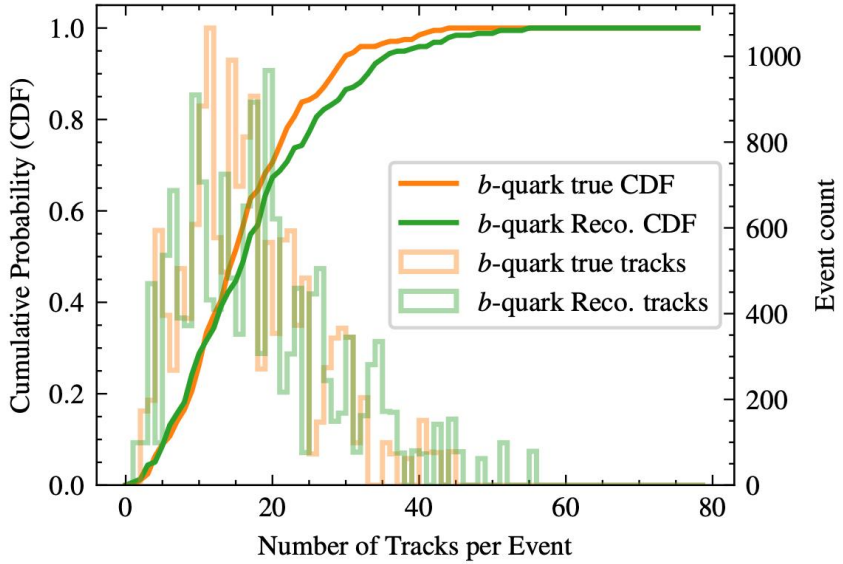


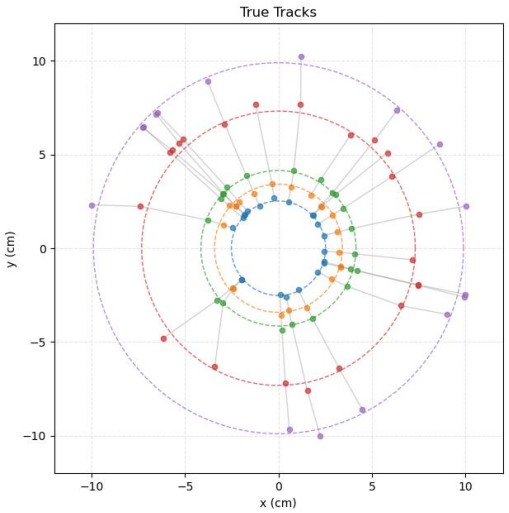


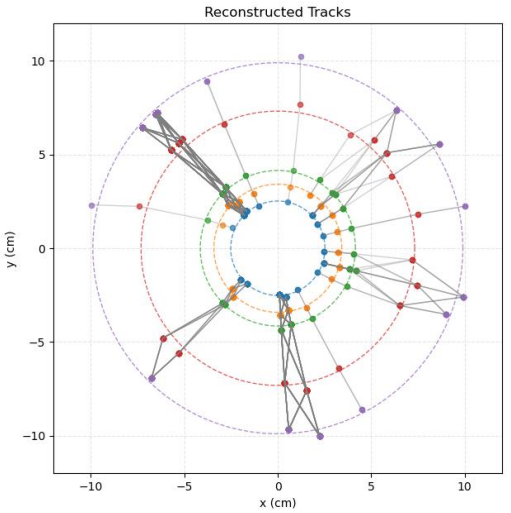
Track reconstruction



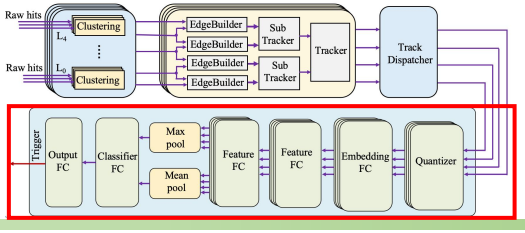
- Efficiency： 89%
- Purity： 72%
- Track Multiplicity Analysis
  - Peak: ~18 tracks/event
  - Maximum: ~60 tracks/event
  - Provides workload estimation for FPGA parallel design
  - φ-region overlap included for boundary track reconstruction
  - Reconstructed tracks show a broader high-multiplicity tail







b-hadron event identification



- Physics signatures
  - Large decay displacement
  - Hard decay products
- ML architecture
  - Lightweight MLP
  - Max + Mean pooling
- Output trigger rate
  - $\varepsilon_{\text{trig}} = 94.2\%$
  - $R_{\text{trig}} = 4$ kHz < 35 kHz

- Trigger rate estimation

$$R_{\text{trig}} = R_{\text{coll}}(f_{\text{b}}\varepsilon_{\text{trig}} + (1-f_{\text{b}})\varepsilon_{\text{bkg}})$$

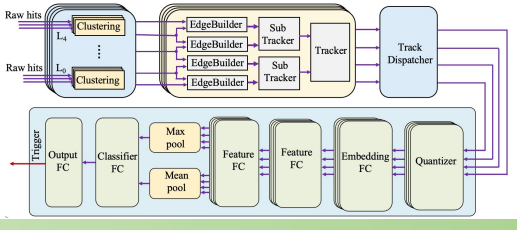
  - $R_{\text{coll}} = 2$ MHz
  - $f_{\text{b}} = 0.003\%$

Dim.	Param.	Signal efficiency $\varepsilon_{\text{trig}}(\%)$	Background acceptance $\varepsilon_{\text{bkg}}(\%)$	R_{trig} (kHz)
96	38,496	86.9	6.3	126
112	48,492	90.8	1.3	26
128	**76,712**	**94.2**	**0.2**	**4**

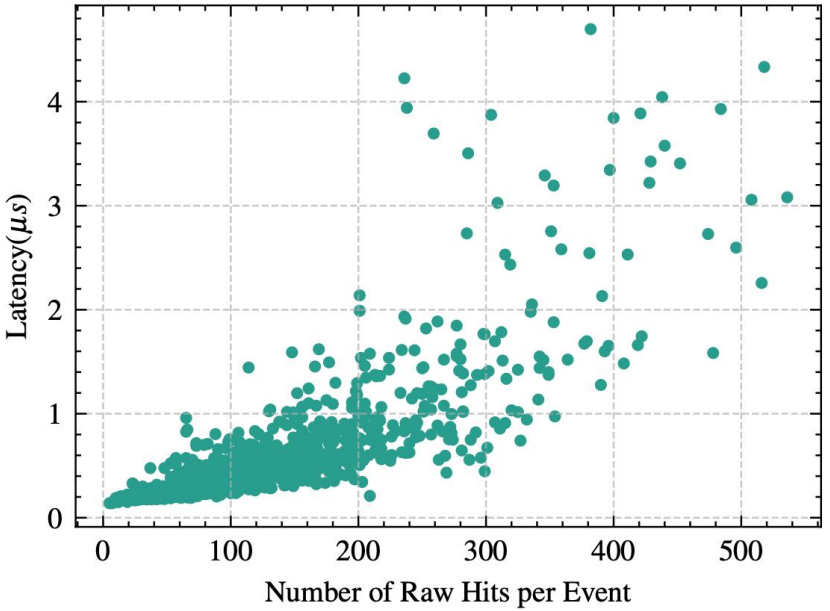
Outline

- Background & requirements
- Algorithms design @ CCNU
- Deployment on FPGA
- Summary

Clustering and tracking



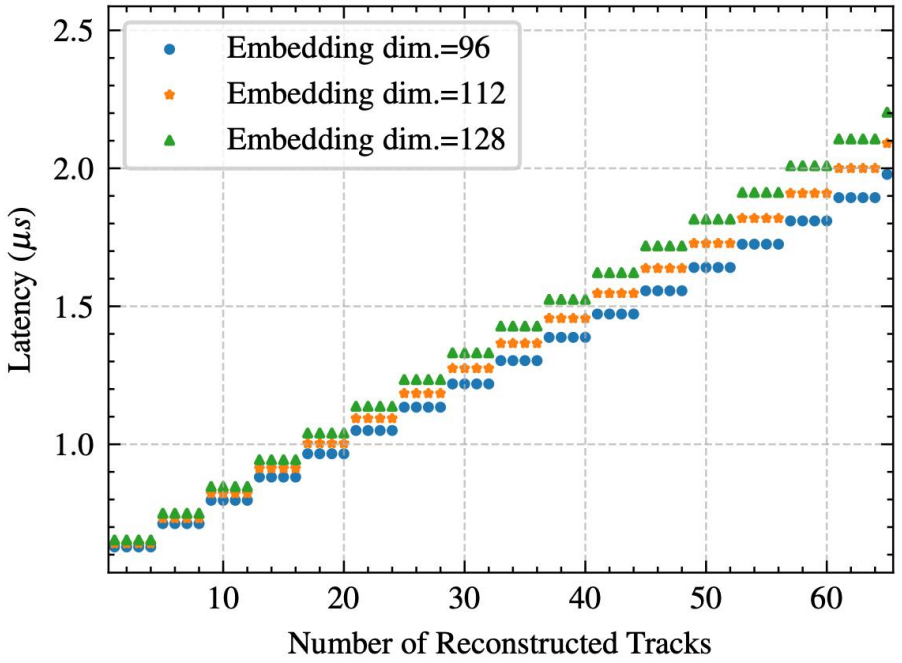
- Clock frequency： 320 MHz
- Latency： 2 μs (98% event coverage)
- Throughput: 21 MHz (10× higher than requirement)
- Platform: Xilinx VCU128 FPGA
  - Sufficient resources for implementation
  - Full trigger algorithm on a single FPGA



	BRAM	DSP	FF (%)	LUT (%)
Clustering and tracking	0	192 (2.1)	329k (12.6)	152k (11.7)
Xilinx VCU128	2,016	9,024	2,607k	1,303k

b-hadron event identification

- FPGA optimization
  - Quantization-Aware Training
  - weight fusion with DSP reuse
  - division elimination
- Clock frequency: 320 MHz
- Latency: 2 μs (up to 60 tracks)
- Throughput: 2.8 MHz
- Resource:
  - DSP-dominated utilization



Dim.	BRAM	DSP	FF (%)	LUT (%)
96	224 (11.1)	2802 (31.1)	133k (5.1)	270k (20.7)
112	224 (11.1)	3267 (36.2)	156k (6.0)	308k (23.6)
128	**224 (11.1)**	**3713 (41.1)**	**176k (6.8)**	**345k (26.4)**

Summary & Outlook

- Summary
  - Developed an FPGA-based low-latency *b*-quark trigger framework for sPHENIX
  - Integrated clustering, tracking, and lightweight ML inference
  - Optimized with QAT, pipelining, parallelism, and DSP reuse
  - Achieved 94.2% efficiency, 4 μs latency, and 2.86 MHz throughput

- Future Work
  - Extend to more physics channels
  - Validate with experimental data
  - Improve scalability with multi-FPGA systems

Thank you for listening