Paper introduction

Xu Zhang
TDAQ group

A Feasibility Study of Transformer-Based Track Trigger Algorithm for CEPC

Xu Zhang, Boping Chen, Liulong Gao, Fei Li, Xiaolu Ji, Minhao Gu, Tingxuan Zeng, Kejun Zhu

***Abstract*—The Circular Electron Positron Collider (CEPC) is envisaged as a multi-purpose facility operating as a factory for Higgs, Z, W bosons, and top-quark pairs ($t\bar{t}$), aimed at achieving high-precision measurements that will rigorously test the Standard Model and probe for new physics. As a future high-luminosity collider, the CEPC will produce a vast number of events, which should be filtered by a trigger system to reduce the data for storage.**

Charged particles traversing the detector leave sequences of energy deposits, known as hits, which form distinctive trajectories, or tracks. Identifying these tracks and utilizing them for trigger decisions is a conventional and fundamental strategy in high-energy physics experiments. Originally the cornerstone of modern Large Language Models, the Transformer utilizes a self-attention mechanism to efficiently capture global correlations within sequential data. While this capability has found success in high-energy physics for complex tasks like jet classification, the targeted application of Transformer-based algorithms to track-based trigger remains less widespread. Therefore, this work focuses exclusively on evaluating the feasibility, performance, and computational efficiency of deploying Transformer encoders for track triggering at the CEPC.

***Index Terms*—CEPC, track-based trigger, Hough Transform, Transformer**

I. Introduction

The discovery of the Higgs boson by the ATLAS and CMS collaborations at the LHC in 2012 provided strong confirmation of the Standard Model [1], [2]. Building on this achievement, scientists have proposed the Circular Electron Positron Collider (CEPC). Designed to serve as a multi-purpose factory, the CEPC will produce abundant samples of W, Z, and Higgs bosons, as well as top-quark pairs ($t\bar{t}$), enabling detailed studies of the fundamental particles to advance our understanding of the Standard Model. In addition, the CEPC provides a crucial experimental platform for studies of flavor physics and Quantum Chromodynamics (QCD).

The CEPC is designed to operate primarily at 240 GeV for Higgs studies within a 30 MW baseline synchrotron-radiation power, with additional operation at 91 GeV for Z boson studies (requiring only 12 MW) and at 160 GeV for W-boson studies focuses on the 50 MW Higgs-mode configuration as a physics-motivated benchmark scenario. Under the projected design parameters, this mode provides higher luminosity and event rates than the baseline Higgs configuration, thereby imposing increased requirements on online processing and computing resources. The 50 MW Higgs mode is therefore used as a demanding benchmark for evaluating the efficiency, latency, and scalability of the proposed trigger algorithms within the Higgs-program scenario.

To achieve precise charged-particle track reconstruction and accurate particle energy measurements, the CEPC reference detector adopts a multi-layer detector architecture consisting of the vertex detector (VTX), inner tracker (ITK), outer tracker (OTK), time projection chamber (TPC), electromagnetic calorimeter (ECAL), hadronic calorimeter (HCAL), and muon system (Muon), as in Fig. 1.

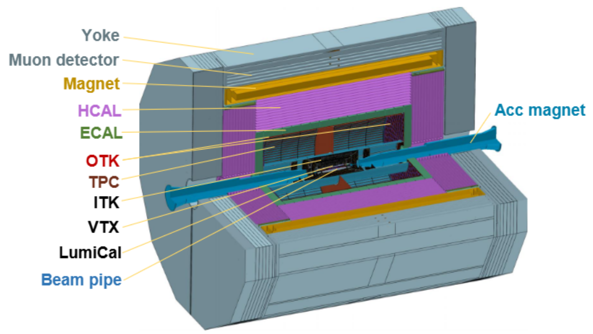


Fig. 1. CEPC detector overview [3].

After detector signals are recorded, the corresponding data volume is determined by the readout time window of each sub-detector. Under the 50 MW Higgs running scenario, the readout time windows, data volumes, and data rates for the individual sub-detectors are summarized in Table I.

In the current CEPC Trigger and DAQ design, the baseline

The goal is to develop a software-trigger algorithm that can reduce the data volume before the HLT stage by approximately two orders of magnitude.

The Hough transform, which is mentioned in the TDR, is used as a benchmark. The main focus is to study Transformer-based trigger algorithms in terms of both performance and computational efficiency.

We use a Transformer model to determine whether there is a track in an event. If a track is present, the event is classified as a good event; otherwise, it is classified as a bad event.

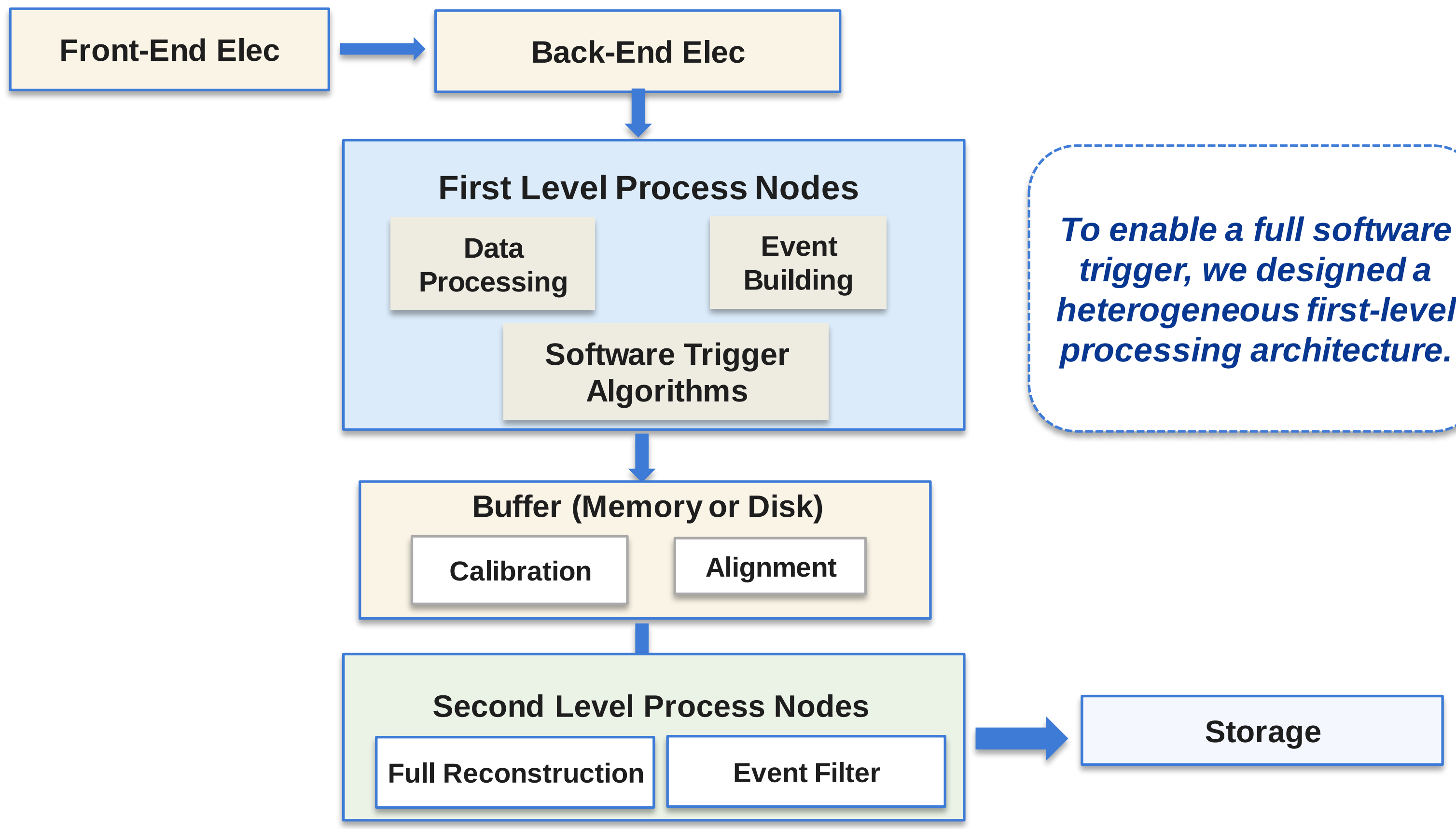
title: A FEASIBILITY STUDY OF TRANSFORMER-BASEDTRACK TRIGGER ALGORITHM FOR CEPC

Xu Zhang — IHEP, University of Chinese Academy of Sciences

1 Motivation

- CEPC is a future Higgs / Z / W factory with high-rate online event selection needs.
- In the 50 MW Higgs mode, the bunch crossing rate is 1.34 MHz.
- A full software trigger is proposed as an upgrade path beyond the hardware baseline.
- While a pure software trigger offers enhanced flexibility, it introduces significant computational overhead.
- GPU acceleration is a practical path for real-time trigger computing.

2 CEPC TDAQ Architecture



To enable a full software trigger, we designed a heterogeneous first-level processing architecture.

3 Methods

A — **A. GPU Hough Transform**

helix parameter-space voting.

B — **B. Transformer**

encoder-only event classification.

C — **C. Linear Transformer**

change attention equation for faster inference.

Results and Conclusions

Accuracy, latency, and resource estimation

4 Key Results: Trigger Accuracy (total 10000 events)

Process	Hough	Transformer	Linear Transformer
Z(vv)H(bb) (1250 events)	100.0	98.93	99.49
Z(vv)H(μμ) (1250 events)	93.2	99.60	99.23
$q\bar{q}$ (1250 events)	98.88	98.16	98.59
$\mu^+\mu^-$ (1250 events)	83.36	97.60	97.23
Beam background (5000 events)	96.68	99.36	99.27

Transformer-based methods improve classification performance.

5 Latency Benchmark

Configuration	Latency (μs/event)
Intel Xeon Gold 6442Y (single core)	1142.22
Hough 4090	44.81
Hough A800	**26.97**
Transformer 4090	84.87
Transformer A800	105.25
Linear Transformer 4090	32.18
Linear Transformer A800	57.27

Benchmarks were measured at peak GPU throughput

Fastest:

Hough A800

26.97 μs/event

6 Resource Estimation

Algorithm	Latency (μs/evt)	Target Rate (MHz)	RTX 4090 GPUs
Hough Transform	44.81	1.34	60
Linear Transformer	32.18	1.34	43

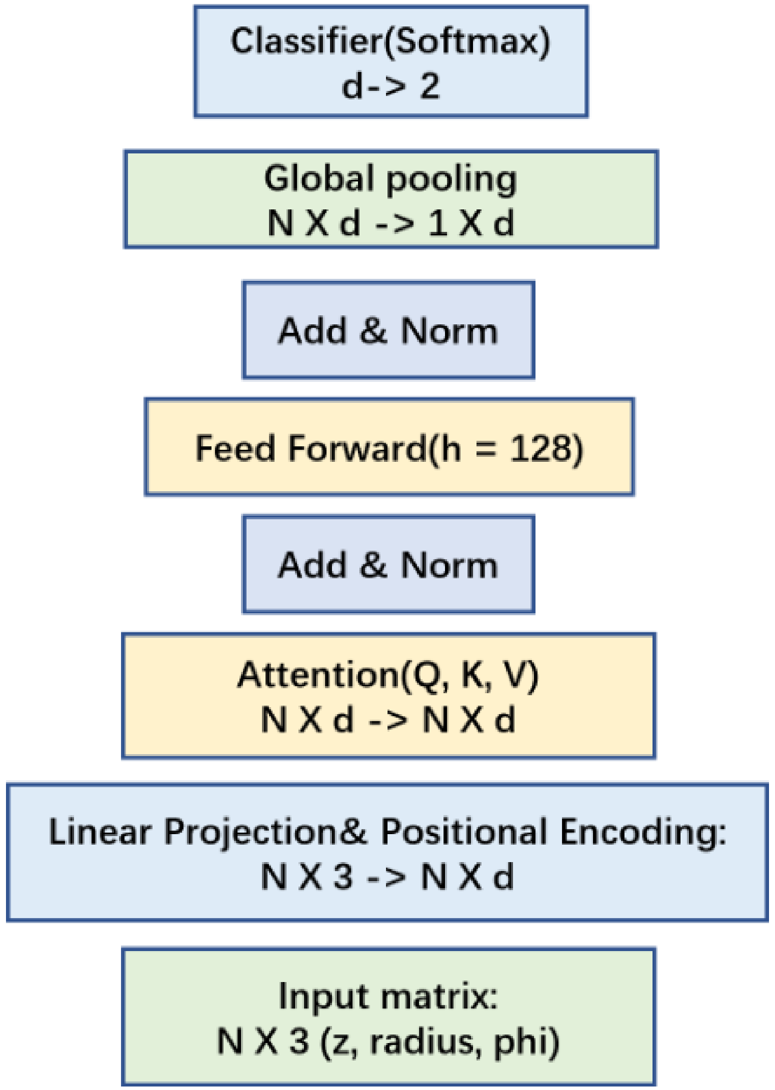
Minimum number of GPUs required for each algorithm.

7 Conclusions

- **a** **Hough: Classic, reliable baseline for tracking R&D.**
- **b** **Transformer: Strong potential for CEPC real-time triggers.**
- **c** **Future: Requires more data for validation and speed optimization.**

Backup

- Transformer-based network architecture design



- In a binary classification problem, the two classes are usually defined as follows: good examples are one class, and background examples are the other class.

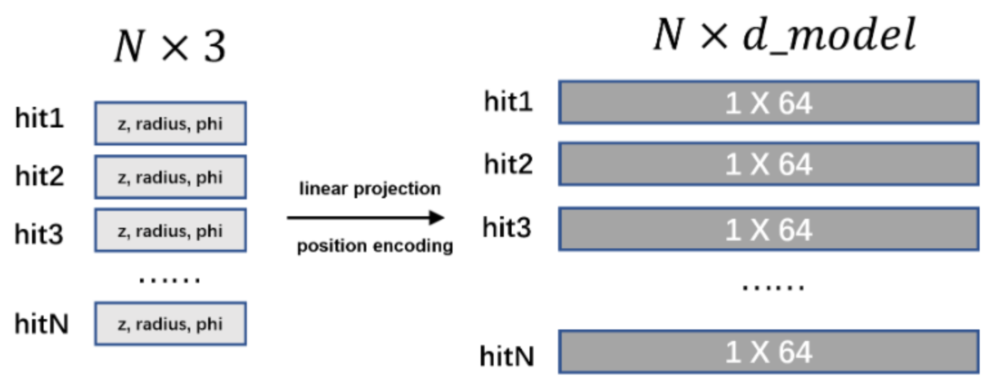
- The good-example class already contains a mixture of background examples. The safety factor is 2.

- 10000 events in ITK OTK VTX(outermost four layer)

Backup

- The actual computation of attention

Step 1: build matrix based on hits in one event



Here, N is the number of hits in one event O(1000)
d_model is the embedding dimension of each hit, which is 64.

Step 3: normal attention, time complexity is O(N^2 x d_model)

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_k}}\right)V,$$

Step 3: linear attention, time complexity is O(N d_model^2)

$$\mathrm{Attention}(Q, K, V) = \frac{\phi(Q)\left(\phi(K)^{\mathsf{T}} V\right)}{\phi(Q)\left(\phi(K)^{\mathsf{T}} \mathbf{1}_N\right)},$$

Step 2: build QKV matrix based on the previous matrix