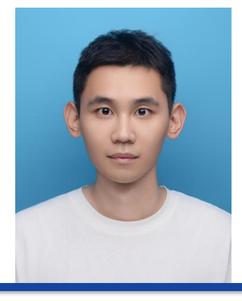
# A Heterogeneous Software Framework Design for the Full Software Trigger at CEPC

## Subtitle: GPU Track-Based Trigger Algorithms for CEPC Studies

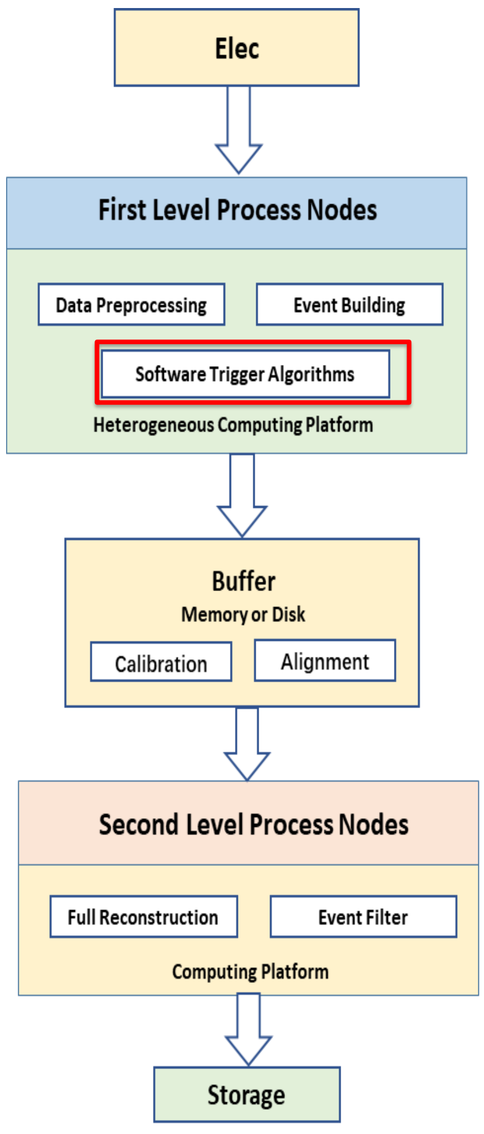
*Speaker: Xu Zhang*

***(1)University of Chinese Academy of Sciences, Beijing 100049, China***
***(2)State Key Laboratory of Particle Detection and Electronics, Institute of High Energy Physics, CAS, Beijing 100049, China***



## 1 Motivation

- **CEPC is a future Higgs / Z / W factory.**
- **In the 50 MW Higgs mode, the bunch crossing rate is 1.34 MHz.**
- **The full software trigger is proposed as an upgrade to the baseline hardware system.**
- **GPU acceleration is investigated as a practical path for real-time trigger computing.**




## 2 Methods

### A. GPU Hough Transform

- **Model equation:**

$$\frac{r\kappa}{2} \approx \phi - \phi_0, \quad z \approx \tan\lambda \cdot r.$$

- **Tracks are identified via hit voting in the helix parameter space.**
- **Parameter map: (κ, $\phi_0$ ,λ) :100×100×100**
- **GPU-accelerated: one thread per hit**

**Time complexity: O(kN)**

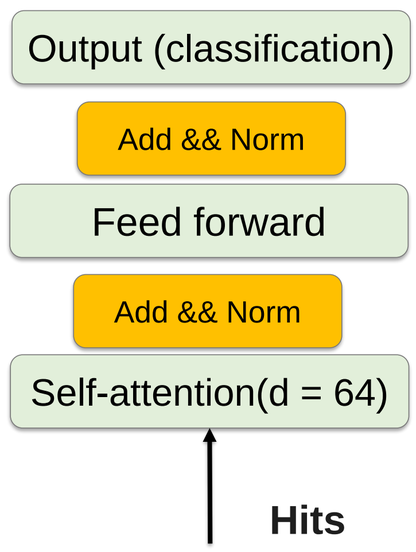
K is 100(scan range per dimension)
N is the number of hits in one event.

### B. Transformer

- **Encoder-only:**

Self-attention layers extract relational features between each hit and all others in the event to reconstruct track candidates

$$\text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V$$




### C. Linear Transformer

- **New attention:**

$$\frac{\phi(Q)(\phi(K)^T V)}{\phi(Q)(\phi(K)^T \mathbf{1}_N)}$$

**Parameters:**
**N: Number of hits per event.**
**d: Feature embedding dimension (64).**

- Standard Transformer **Time: O(N^2d)** computationally expensive for large N.
- Linear Transformer **Time: O(Nd^2)** faster than standard transformer

## 3 Dataset

- **The model is trained on 10,000 Monte Carlo collision events** (50% background,
  12.5% Z(νν)H(bb),
  12.5% Z(νν)H(μμ),
  12.5% $q\bar{q}$,
  12.5% $\mu^+\mu^-$).
- **The input features consist solely of hit positions in cylindrical coordinates (r,$\phi$, z), associated with the ITK, OTK, and out layers of VTX systems.**
- **Background hits are overlaid onto signal events to emulate realistic detector conditions.**
- **The dataset is split 80% / 20% for training and testing, respectively.**

## 4 Key Result

### A. Efficiency( Veto Efficiency for background)

| Process /Algorithm | Hough | Transformer | Linear Transformer |
|---|---|---|---|
| Z(νν)H(bb) | 100.0 | 100.0 | 100.0 |
| Z(νν)H(μμ) | 93.2 | 99.6 | 99.6 |
| $q\bar{q}$ | 98.88 | 98.4 | 99.2 |
| $\mu^+\mu^-$ | 83.36 | 97.6 | 98.4 |
| Beam background | 96.68 | 99.7 | 99.7 |

### B. Process Time

Evaluated at GPU peak throughput

| Configuration | Time (μs/event) |
|---|---|
| Intel Xeon Gold 6442Y (single core) | 1142.22 |
| Hough 4090 | 44.81 |
| Hough A800 | 26.97 |
| Transformer 4090 | 86.06 |
| Transformer A800 | 107.03 |
| Linear Transformer 4090 | 33.57 |
| Linear Transformer A800 | 57.55 |

## 5 Resource estimation

**Minimum number of GPUs required for each algorithm**

| Algorithm | Latency (μs/evt) | Target Rate (MHz) | RTX 4090 GPUs |
|---|---|---|---|
| Hough Transform | 44.81 | 1.34 | 60 |
| Linear Transformer | 33.57 | 1.34 | 45 |

## 6 Conclusion

- **The Hough Transform, as a classic and general-purpose track reconstruction algorithm, retains significant research value and serves as an important benchmark for future studies.**
- **Transformer-based models, particularly Linear Transformers, exhibit potential for track-based trigger at CEPC.**
- **Further validation requires more extensive testing and additional datasets to fully evaluate model capabilities, alongside future efforts to optimize inference speed.**