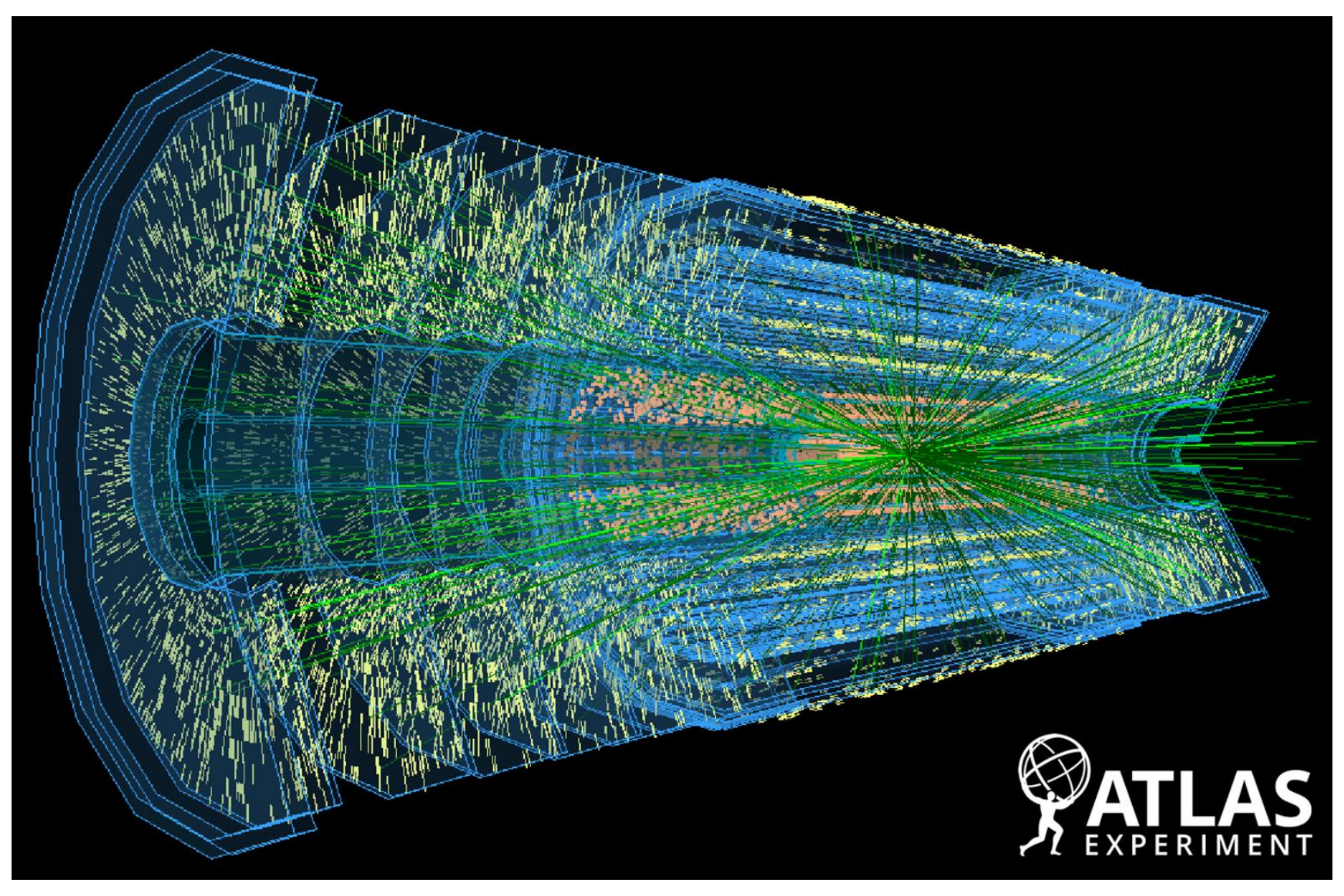
TJCAS 2025
Taiwan and Japan Conference on Circuits and Systems

An Integrated FPGA Implementation of Complete GNN-Based Trajectory Reconstruction

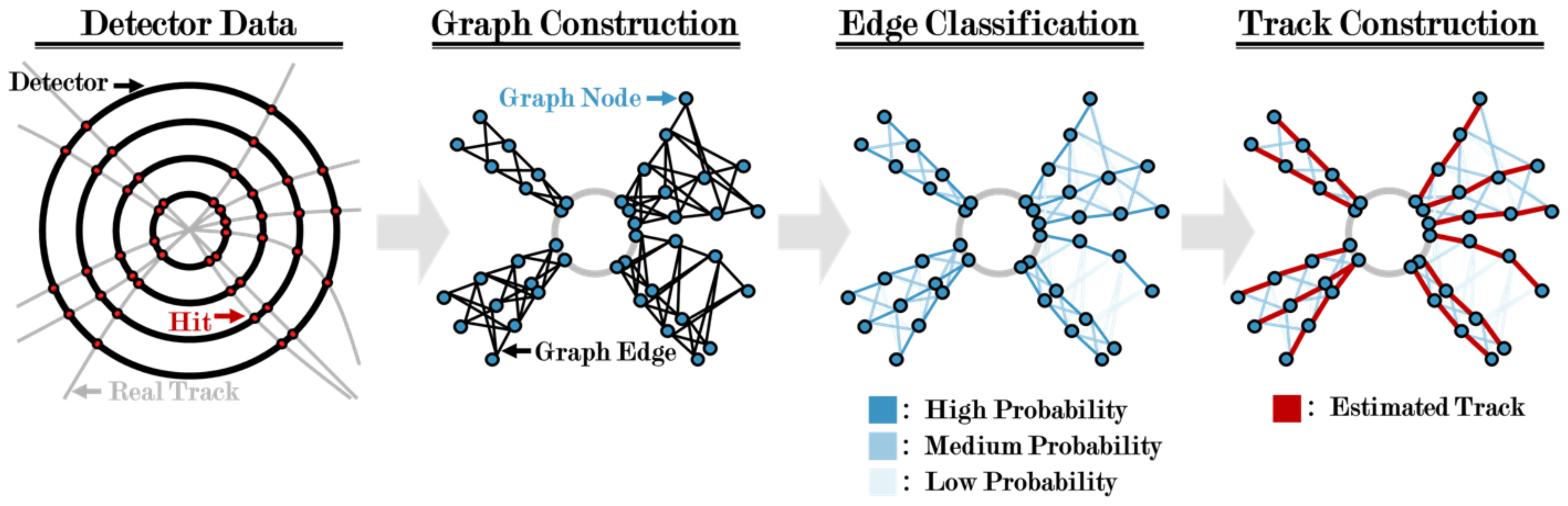
Yun-Chen Yang*, Hao-Chun Liang*, Hsuan-Wei Yu*,
Bo-Cheng Lai*, Shih-Chieh Hsu†, Mark Neubauer‡, Santosh Parajuli‡
National Yang Ming Chiao Tung University, Hsinchu, Taiwan*, University of Washington, USA† University of Illinois Urbana-Champaign, USA‡

Motivation & Challenge

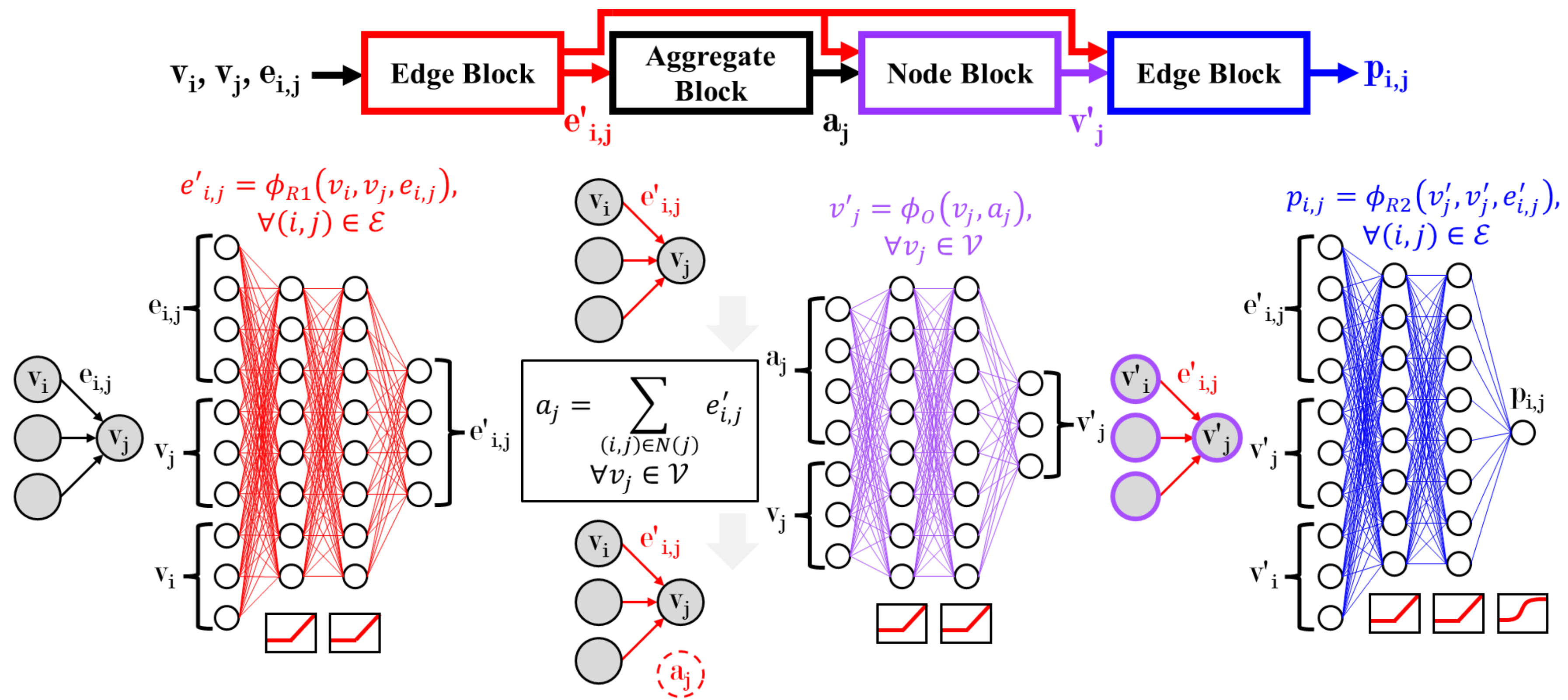


- Collaborate with CERN
  - ➢ Proton-Beam Collisions
  - ➢ Detector Produce Hits
- High-Volume Collision Hits
  - ➢ L1T System Filtering
  - ➢ via Track Reconstruction
- HL-LHC Requirement
  - ➢ Latency: 4 μs
  - ➢ Throughput: 2.22MHz
- Accuracy-Focused, Execution-Speed-Agnostic Approach [7, 8]
  - ➢ CPU-Based Limits Hindering Task-Specific Tuning
  - ➢ GPU-Based Inefficiency in Latency-Critical Scenarios
    - ➢ Missed Microsecond Targets By Milliseconds
- FPGA-Accelerated Framework
  - ➢ Scope Constrained to GNN Edge-Classification Stage [9,10]
    - ➢ Host-to-Device Data Transfers Latency
    - ➢ FPGA Resources Underutilization
  - ➢ Throughput-Driven Processing In Minor Graph Subregions[11]
    - ➢ Small Subgraph Lowers Accuracy

Trajectory Reconstruction

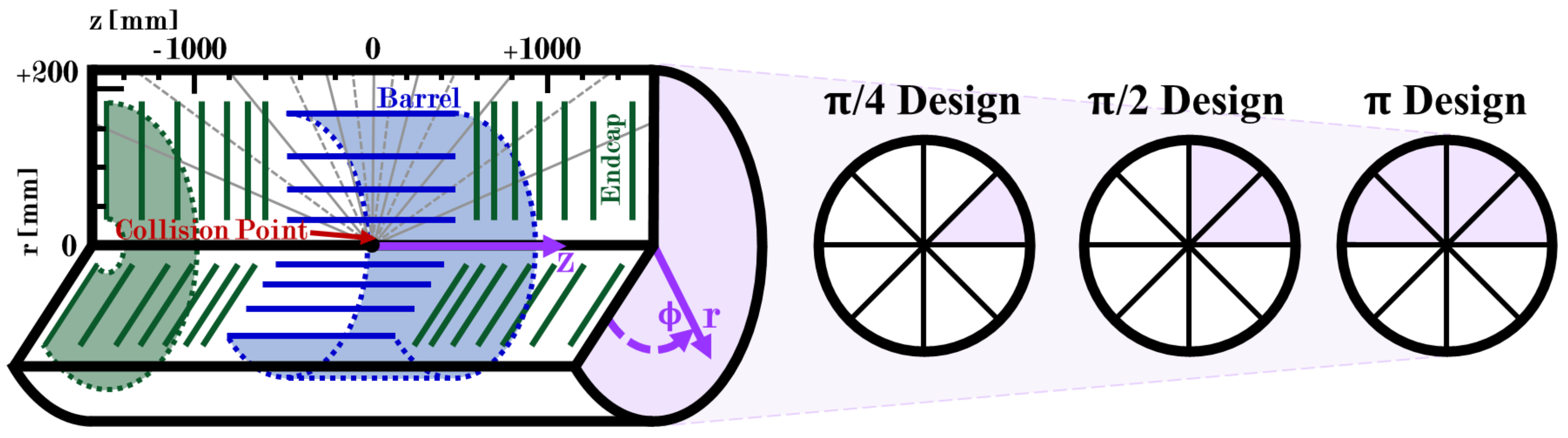


- Graph Construction – Map Hits to Nodes and Filter Directed Edges
- Edge Classification – Assign Probabilities to Edges
- Track Construction – Integrate Edges with High Probabilities


v_i, v_j, $e_{i,j}$
Edge Block
Aggregate Block
Node Block
Edge Block
$p_{i,j}$
$e'_{i,j}$
a_j
v'_j
$e'_{i,j} = \phi_{R1}(v_i, v_j, e_{i,j}),\ \forall (i,j) \in \mathcal{E}$
$a_j = \sum_{(i,j)\in N(j)} e'_{i,j}\ \ \forall v_j \in \mathcal{V}$
$v'_j = \phi_O(v_j, a_j),\ \forall v_j \in \mathcal{V}$
$p_{i,j} = \phi_{R2}(v'_j, v'_j, e'_{i,j}),\ \forall (i,j) \in \mathcal{E}$
v_i
v_j
$e_{i,j}$
v'_i
v'_j
a_j

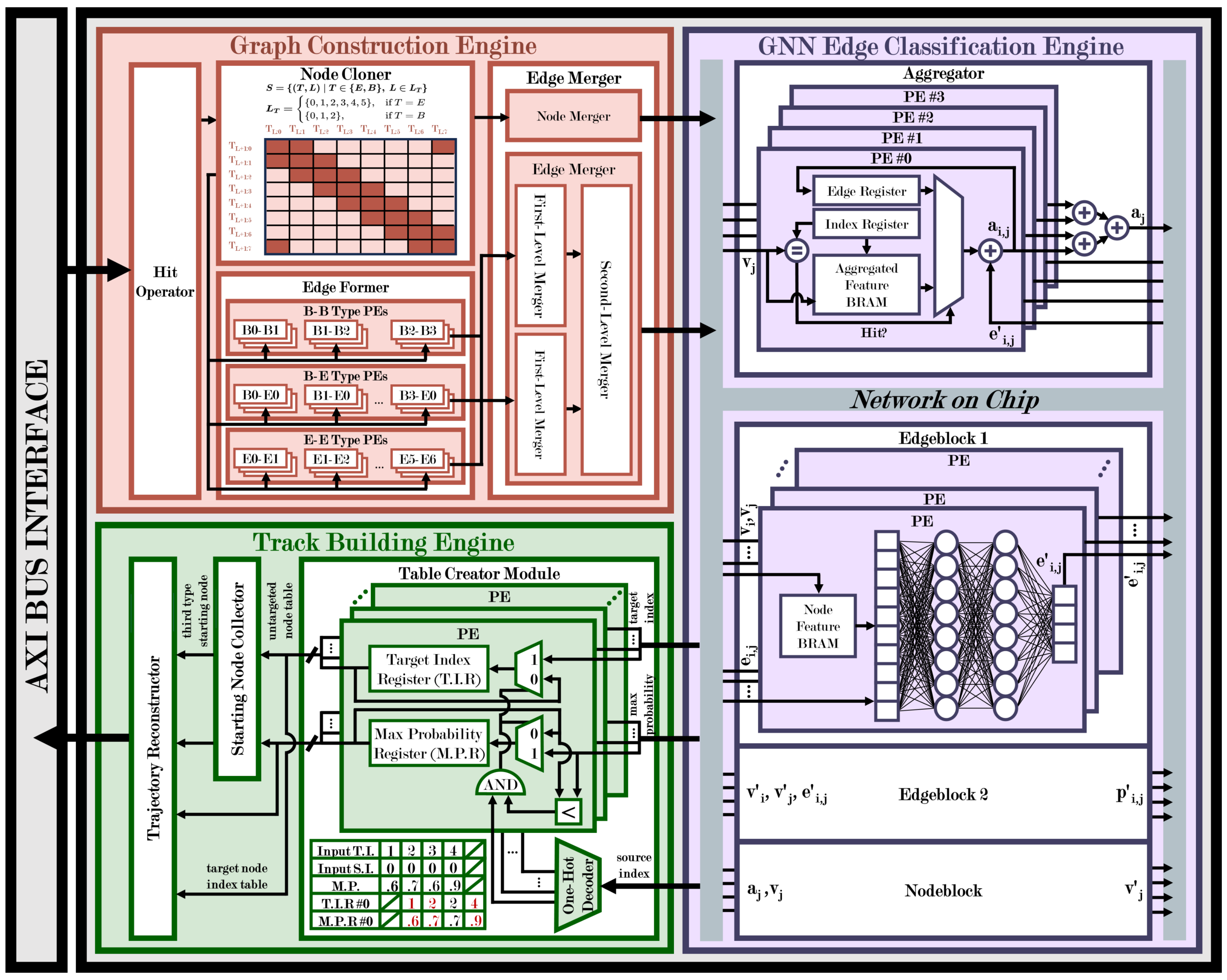
- Interaction Network (IN) Framework[6]
  - ➢ Graph Neural Network for Object-Object Interaction Modeling

Configuration

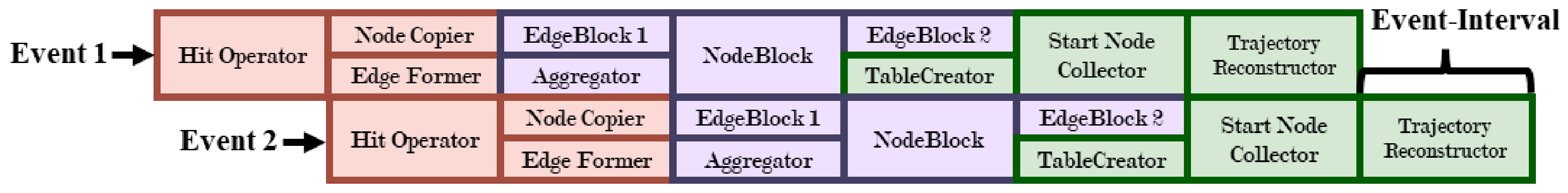


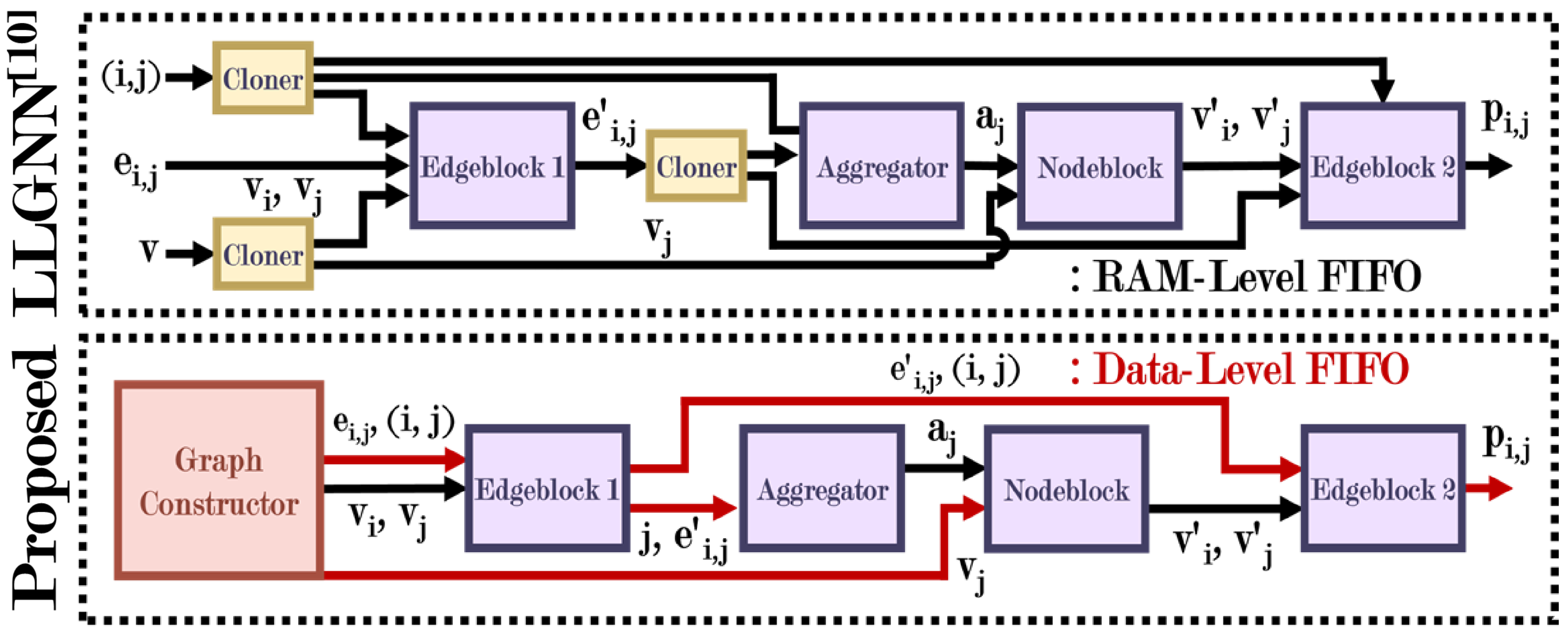
- Cross-Sectional View of Cylindrical Collider Detector Architecture
  - ➢ 4× Cylindrical Barrels and 14× Planar Endcaps
- Hit Distribution across Segment
  - ➢ Longitudinal Segmentation along the Z Axis
  - ➢ Azimuthal Segmentation along the Φ-Axis into 2/4/8 Sectors
    - ➢ Adapted to Support Three Distinct Design Variants
  - ➢ Spatial-Multiplexed FPGA for Parallel Processing

Architecture



Input T.I.	1	2	3	4	
Input S.I.	0	0	0	0	
M.P.	.6	.7	.6	.9	
T.I.R #0		1	2	2	4
M.P.R #0		.6	.7	.7	.9

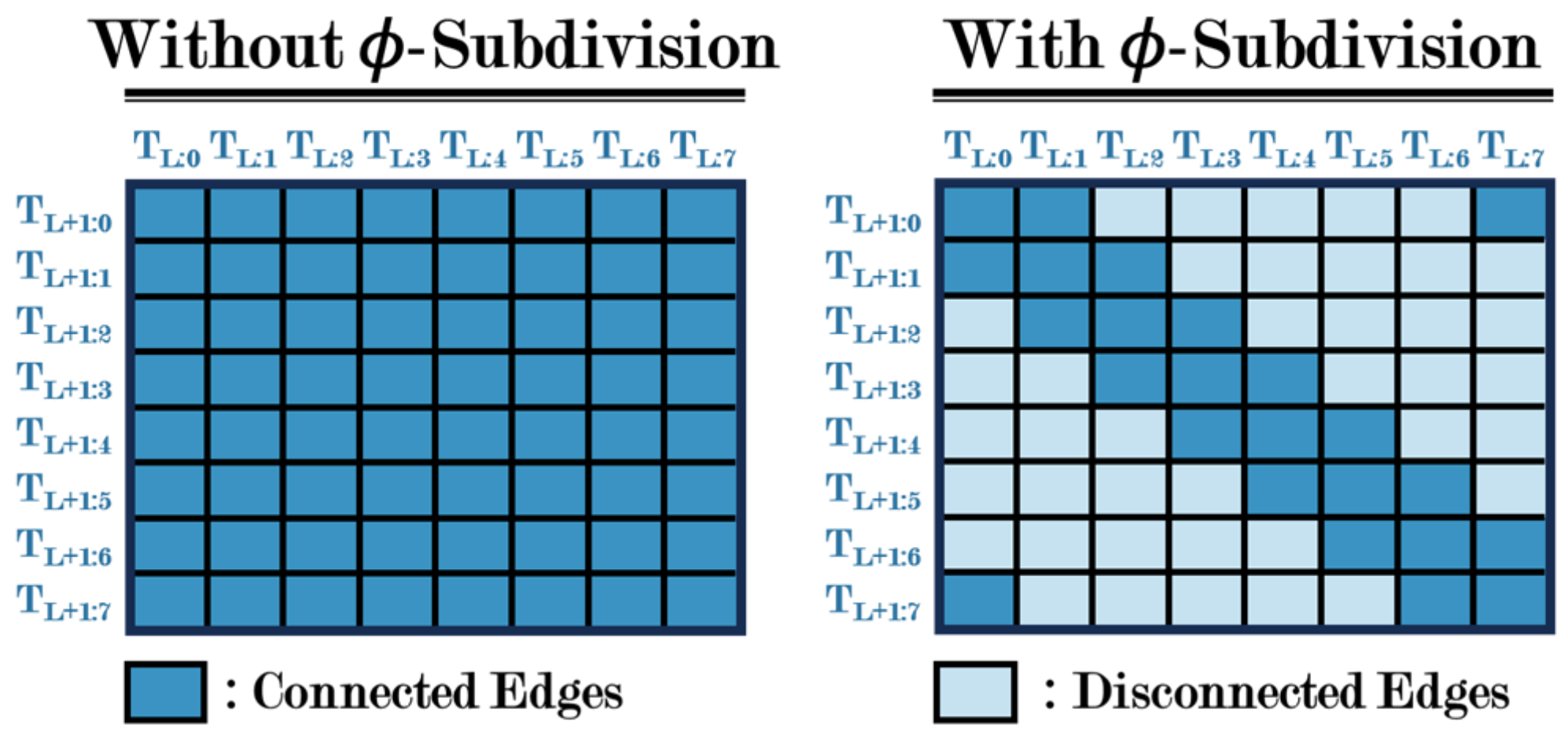




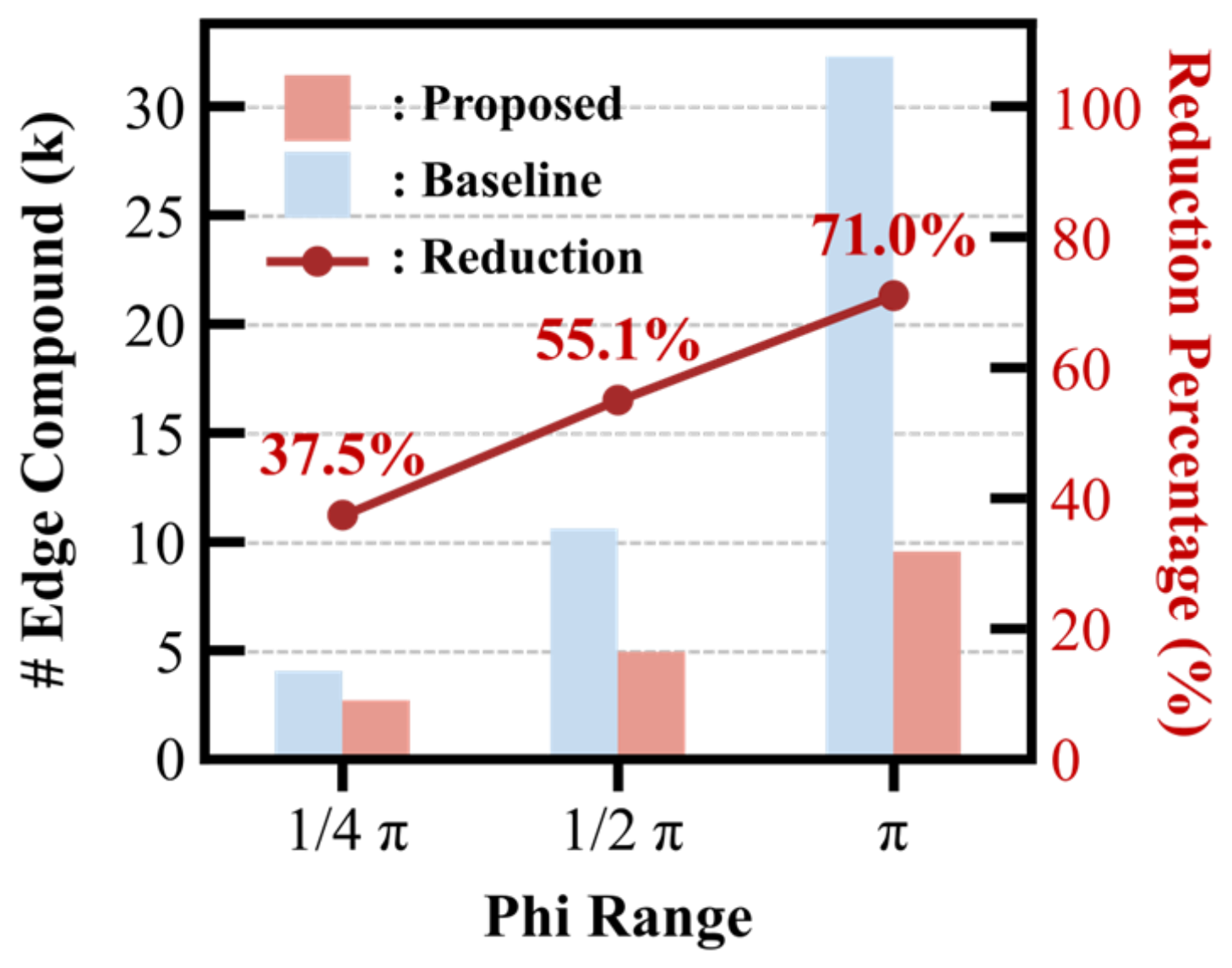
- Edge Classification from Batch Processing[10] to Data Streaming
  - ➢ 52.3% Latency Reduction from 2.86 μs to 1.365 μs

Algorithm & Conclusion

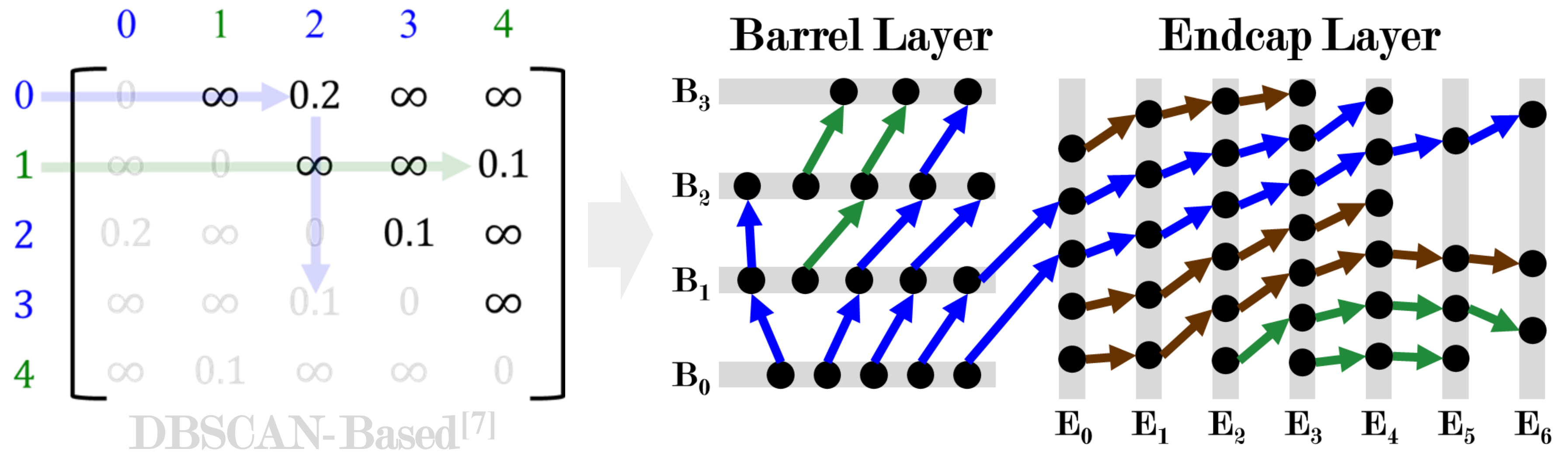
- Geometry-Aware Edge Pruning in Graph Construction
  - ➢ Edge Count Reduction Achieving 37.5%–71.0% Pruning[7]


Without ϕ-Subdivision
With ϕ-Subdivision
$T_{L:0}$ $T_{L:1}$ $T_{L:2}$ $T_{L:3}$ $T_{L:4}$ $T_{L:5}$ $T_{L:6}$ $T_{L:7}$
: Connected Edges
: Disconnected Edges

$$L(T) \in \begin{cases} \{0,1,2,3,4,5\}, & \text{if } T = E \\ \{0,1,2\}, & \text{if } T = B \end{cases}$$



- Probability-Based Sequential Building in Track Construction
  - ➢ LUT-Based Mapping from B_0 (Blue), E_0 (Brown), Other (Green)
  - ➢ Latency Reduction Ranging from 107× to 641×[7]



$$\begin{array}{c|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & \infty & 0.2 & \infty & \infty \\ 1 & \infty & 0 & \infty & \infty & 0.1 \\ 2 & 0.2 & \infty & 0 & 0.1 & \infty \\ 3 & \infty & \infty & 0.1 & 0 & \infty \\ 4 & \infty & 0.1 & \infty & \infty & 0 \end{array}$$

- First End-to-End FPGA of GNN-Based Trajectory Reconstruction
  - ➢ 65024× Acceleration with Enhanced Accuracy Metrics[7,9]
  - ➢ Throughput 2.35 MHz with Latency 2.36 μs Meets L1T Criteria