A High-Performance Implementation of GNN-Based Trajectory Reconstruction on FPGA

Yun-Chen Yang, Hao-Chun Liang, and Bo-Cheng Lai

Institute of Electronics
National Yang Ming Chiao Tung University

Presenter: Hao-Chun, Liang
Supervisor: Bo-Cheng Lai

NATIONAL
YANG MING CHIAO TUNG
UNIVERSITY

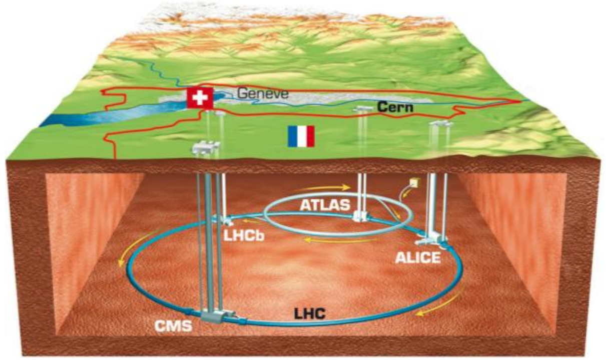
2025 VLSI Design / CAD Symposium; ID: 0026

Outline

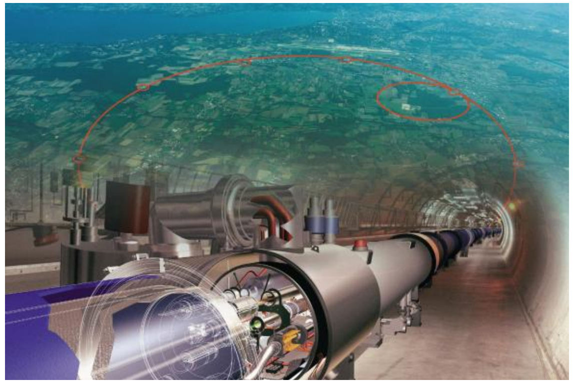
- **Background**
  - ***Challenges***
- **Related Work**
  - ***Limitations***
- **Contributions**
- **Experiment**
- **Conclusion**
- **Reference**

Background

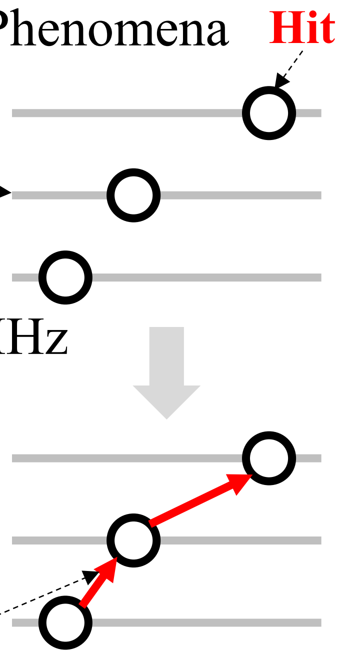
- **Collaborative Program with CERN**
  European Organization for Nuclear Research
  - ***Research Emphasis on High-Energy Physics – (HEP)***
  - ***Partnerships with Hundreds of International Universities***
    - University of Illinois Urbana-Champaign; University of Washington

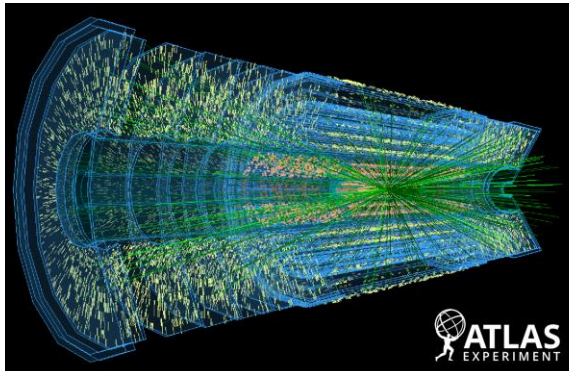


- **Large Hadron Collider (LHC) Infrastructure Overview[1]**
  - Collision-Event Analysis for Exploration of Novel Physical Phenomena



- **Methodology Framework for Collision Analysis**
  - ***Dual Proton-Beam Acceleration***
    - Near-Light Speed for High-Energy Collisions
    - High-Frequency Occurrence of Collision Events at 40 MHz
  - ***Generated Particles Pass through Detectors Yielding Hits***
  - ***Trajectory Reconstruction-Based Analysis of Hits***





[1] L. Evans, "The large hadron collider," New Journal of Physics, vol. 9, no. 9, p. 335, 2007.

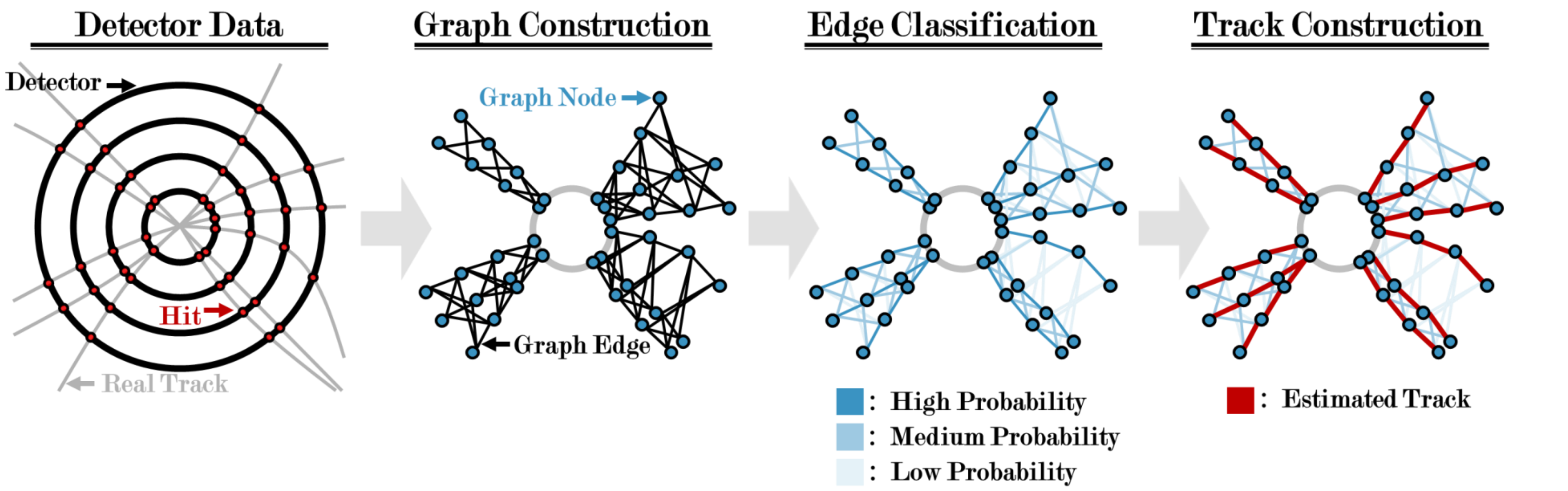
Challenge

- **Archiving High-Volume Collision Hits**
  Offline Analysis Capacity Considerations
  - ***Real-Time Data Reduction via Level-1 Trigger (L1T)[2] System***
    - Selective Acquisition of Critical Data for Offline Processing
    - Trajectory Reconstruction in Trigger Decision Making
- **Stringent Latency and Throughput Constraints in L1T System**
  - ***HL-LHC Upgrades[3] and CMS[4] Detector Enhancement Strategies***
    HL-LHC – High-Luminosity Large Hadron Collider
    - Latency Budget Allocation of 4 μs for Data Selection
    - Event-Rate Processing of 2.22 MHz through Time-Multiplexing
      Time-Multiplexed Distribution of 40 MHz Collisions to 18 FPGAs
- **Insufficiency of Current LHC Tracking Algorithm[5]**
  - Inadequate Performance under Stringent Constraints

[2] "The Phase-2 Upgrade of the CMS Level-1 Trigger," CERN, Geneva, Tech. Rep., 2020, final version.
[3] O. Aberle, C. Adorisio, A. Adraktas, M. Ady, J. Albertone, L. Alberty, M. Alcaide Leon, A. Alekou, D. Alesini, B. Almeida Ferreira et al., "High-luminosity large hadron collider (hl-lhc): Technical design report," 2020.
[4] "The Phase-2 Upgrade of the CMS Tracker," CERN, Geneva, Tech. Rep., 2017.
[5] R. Frühwirth, "Application of kalman filtering to track and vertex fitting," Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 262, no. 2-3, pp. 444–450, 1987.

Related Work

- **Interaction Network (IN)[6] Framework**
  - Specialized Graph Neural Network for Object-Object Interaction Modeling
- **GNN-Based Trajectory Reconstruction Framework**
  - **Graph Construction Stage –** Mapping Hits and Segment Candidates to Nodes and Directed Edges
  - **Edge Classification Stage –** Probabilistic Assessment of Edge Validity
  - **Track Construction Stage –** Integration of Edge Probabilities in Trajectory Reconstruction



[6] P. Battaglia, R. Pascanu, M. Lai, D. Jimenez Rezende et al., "Interaction networks for learning about objects, relations and physics," Advances in neural information processing systems, vol. 29, 2016.

Related Work – Limitations

- **Software-Driven Framework Leveraging CPUs or GPUs**
  - ***Accuracy-Centric Methodology with Execution-Speed Agnosticism*** [7,8]
    - CPU-Based Constraints Impeding Task-Specific Optimization Potential
    - GPU-Based Inefficiency in Single-Event and Latency-Critical Scenarios
    - Substantial Millisecond-Scale Deviation from L1T Microsecond Requirements
- **FPGA-Accelerated Framework**
  - ***Throughput-Driven Processing Limited to Minor Graph Subregions***[11]
    - Impact of Small Subgraph on Accuracy Degradation
  - ***Scope Constrained to GNN Edge-Classification Stage***[9,10]
    - Imposition of Host-to-Device Data Transfers Resulting in FPGA Underutilization

[7] G. DeZoort, S. Thais, J. Duarte, V. Razavimaleki, M. Atkinson, I. Ojalvo, M. Neubauer, and P. Elmer, "Charged particle tracking via edge-classifying interaction networks," Comput. Softw. Big Sci., vol. 5, no. 1, pp. 1–13, 2021.

[8] X. Ju, D. Murnane, P. Calafiura, N. Choma, S. Conlon, S. Farrell, Y. Xu, M. Spiropulu, J.-R. Vlimant, A. Aurisano et al., "Performance of a geometric deep learning pipeline for hl-lhc particle tracking," The European Physical Journal C, vol. 81, pp. 1–14, 2021.

[9] A. Elabd, V. Razavimaleki, S.-Y. Huang, J. Duarte, M. Atkinson, G. DeZoort, P. Elmer, S. Hauck, J.-X. Hu, S.-C. Hsu et al., "Graph neural networks for charged particle trac king on fpgas," Frontiers in big Data, vol. 5, p. 828666, 2022.

[10] S. Huang, Y. Yang, Y. Su, B. Lai, J. Duarte, S. Hauck, S. Hsu, J. Hu, and M. S. Neubauer, "Low latency edge classification gnn for particle trajectory tracking on fpgas," in 2023 33rd International Conference on Field-Programmable Logic and Applications (FPL). Los Alamitos, CA, USA: IEEE Computer Society, sep 2023, pp. 294–298.

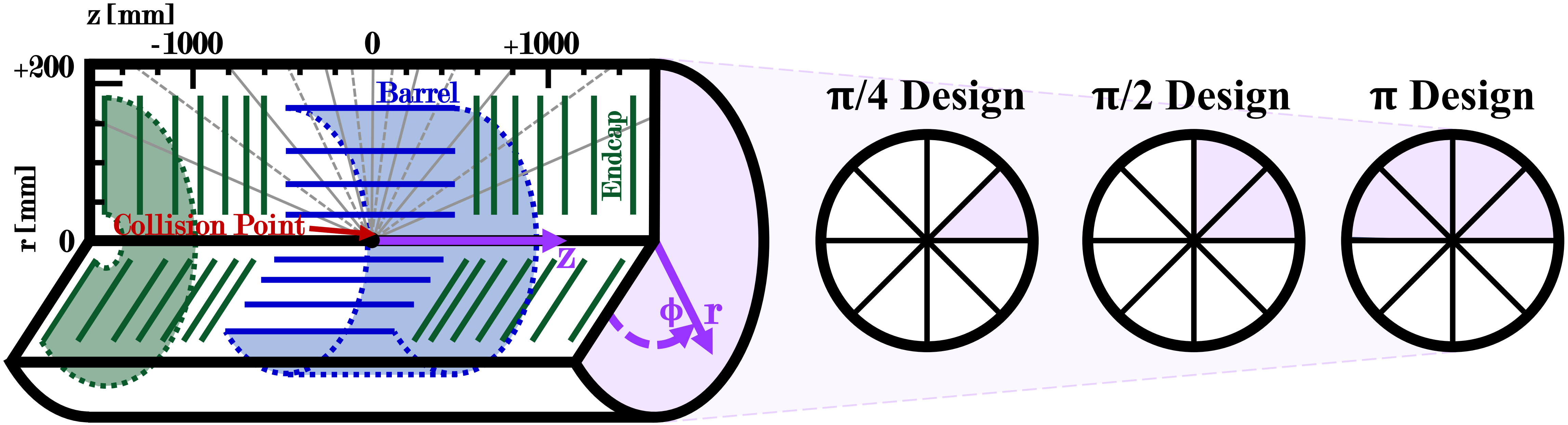
[11] Aneesh Heintz, Vesal Razavimaleki, Javier Duarte, Gage DeZoort, Isobel Ojalvo,Savannah Thais, Markus Atkinson, Mark Neubauer, Lindsey Gray, Sergo Jindari-ani, et al. 2020. Accelerated charged particle tracking with graph neural networkson FPGAs. arXiv preprint arXiv:2012.01563 (2020).

Contribution

- **End-to-End GNN-Based FPGA Accelerator for Trajectory Reconstruction**
  - ***Algorithmic Optimizations Derived from Reference[7]***
    - Geometry-Aware Edge Pruning on Graph Construction
      - Computational Load Minimization
    - Probability-Based Sequential Building on Track Construction
      - FPGA Deployability Enhancement
  - ***Data Streaming–Oriented High-Throughput Optimization Inspired by Reference[10]***
    - Batch Processing of Edge-Classification into Data-Streaming Paradigm
      - Achieved Latency Reduction of 52.3%
  - ***Consolidated Three-Stage Pipeline within High-Performance FPGA Accelerator***
    - **65024× Acceleration** over Software-Based Approach with **Enhanced Accuracy Metrics**[7,9]
    - First End-to-End GNN-Based FPGA Accelerator for Trajectory Reconstruction
    - Event-Throughput Rate of 2.35 MHz with 2.36 μs Latency Meeting L1T Criteria

Configuration

- **Cross-Sectional View of Cylindrical Collider Detector Architecture**
  - Geometrical Configuration of 4× Cylindrical Barrels and 14× Planar Endcaps
- **Hit Distribution across Detector-Segment via Spatially Multiplexed FPGA**
  - Longitudinal Segmentation along the Z Axis
  - Azimuthal Segmentation along the Φ-Axis into 2/4/8 Sectors for Scalability
- **Specifically Adapted to Support Three Distinct Design Variants**

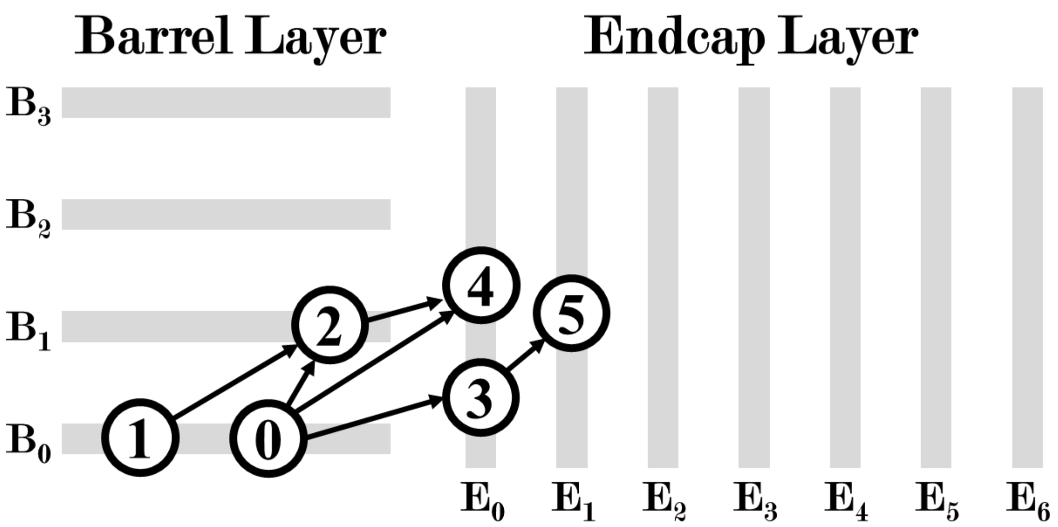


Graph Construction Algorithm – Baseline

- **Graph Construction from High-Energy Collision Data[7]**
  - Node **Exhaustive Enumeration** in Adjacent Layers for Edge Feature Extraction
  - Selection of Track-Segment Candidates via Feature Thresholding
  - Mapping of Hits and Track-Segment Candidates onto Nodes with Directed Edges
  - Conversion to Sparse Coordinate List (COO) Format for Input of GNN Stage
    - Specification of Source and Target Node Indices with Edge Feature Vectors

Four-Dimensional Edge Feature: $(\Delta r_{ij}, \Delta\Phi_{ij}, \Delta z_{ij}, \Delta R_{ij})$

$$\Delta R_{ij} = \sqrt{(\Delta\phi_{ij})^2 + \left(\ln\left(\frac{\tan\left(\frac{1}{2}\mathrm{atan2}(r_i, z_i)\right)}{\tan\left(\frac{1}{2}\mathrm{atan2}(r_j, z_j)\right)}\right)\right)^2}$$



Source Node Index (columns) / Target Node Index (rows)

	0	1	2	3	4	5
0						
1						
2	$e_{0,2}$	$e_{1,2}$				
3	$e_{0,3}$					
4	$e_{0,4}$		$e_{2,4}$			
5				$e_{3,5}$		

Source Node Index

0	1	0	0	2	3

In order

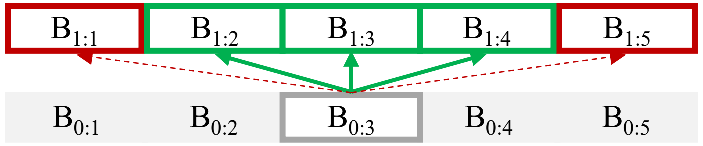
Target Node Index

2	2	3	4	4	5

Edge Features

$e_{0,2}$	$e_{1,2}$	$e_{0,3}$	$e_{0,4}$	$e_{2,4}$	$e_{3,5}$

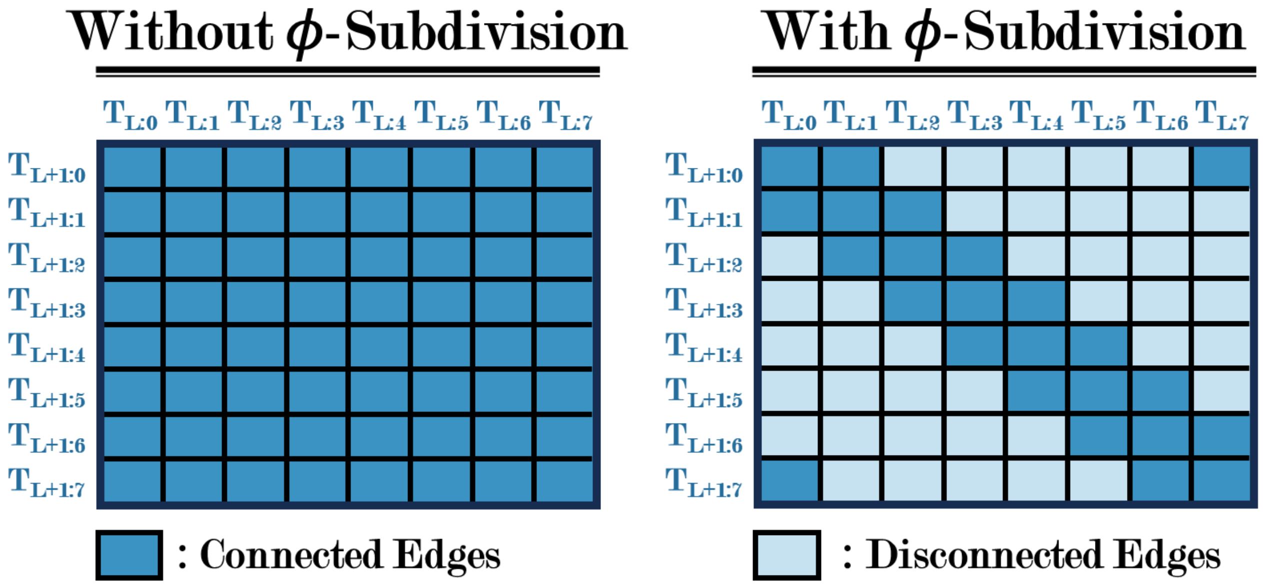
Graph Construction Algorithm – Proposed



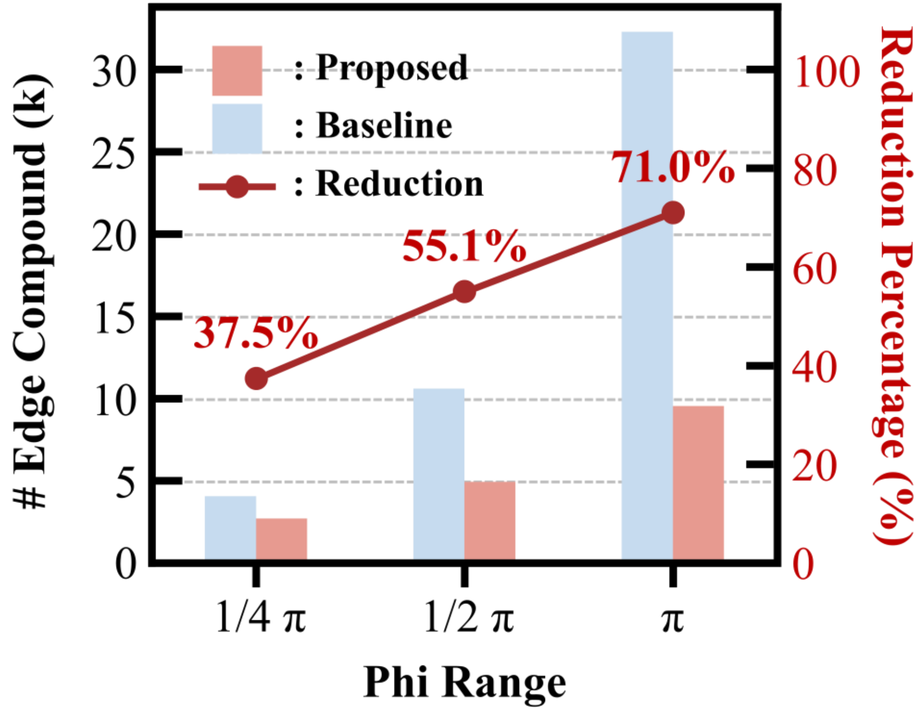
- **Exhaustive Connectivity Enumeration: Computational Infeasibility**
  - Edge Candidates Formation Confined to ΔΦ-Span ≪ Φ-Range
- **Neighborhood-Constrained Optimization Strategy**
  - Adjacency-Based Candidate Edge Restriction
  - **Computational Load Reduction and Scalability Enhancement**

Φ-Range	π/4	π/2	π
Φ-Range/ΔΦ-Span	9.13	18.23	36.53

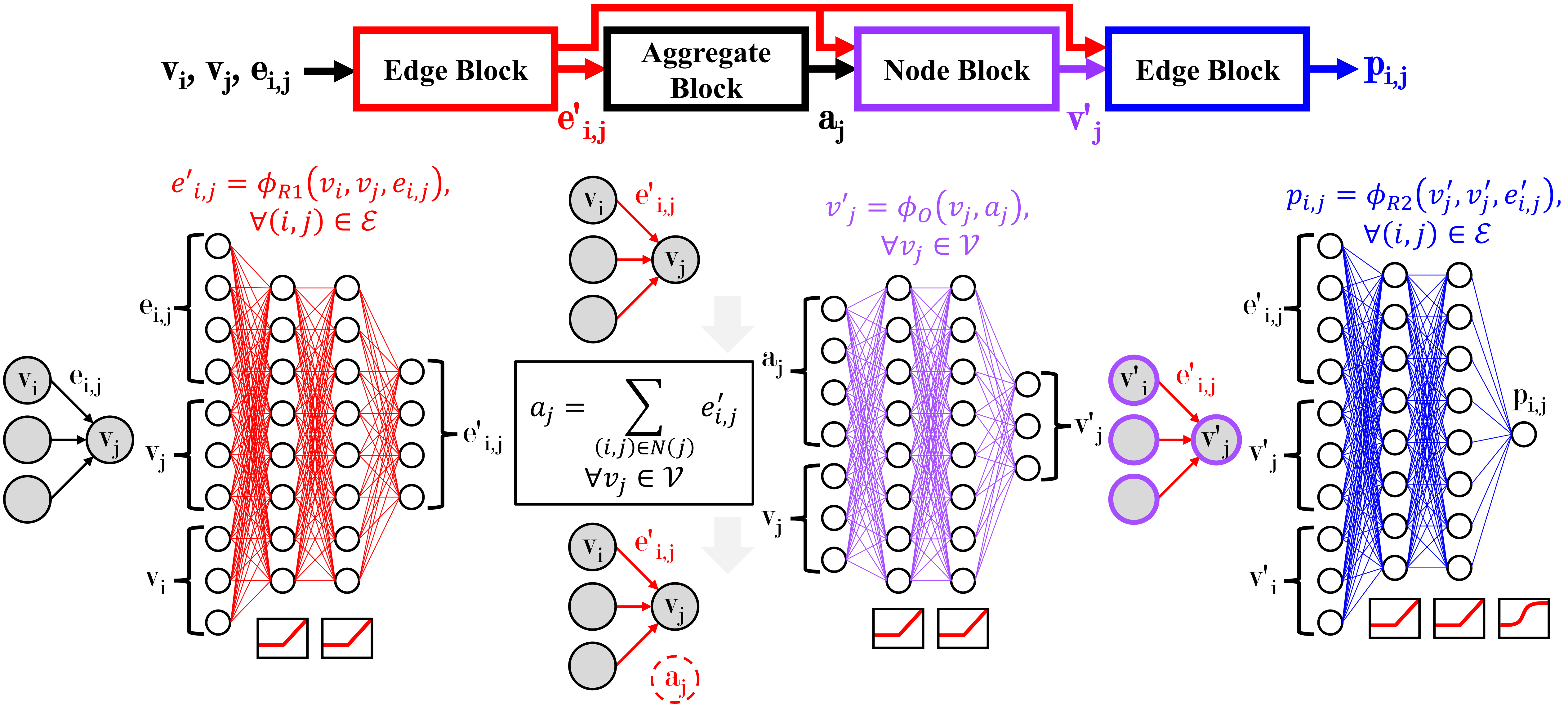
※ *Ratios of Full Φ-Range to Maximum ΔΦ for Potential Can-didate Edges Across Three Azimuthal Segmentation*



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



Edge Classification Algorithm



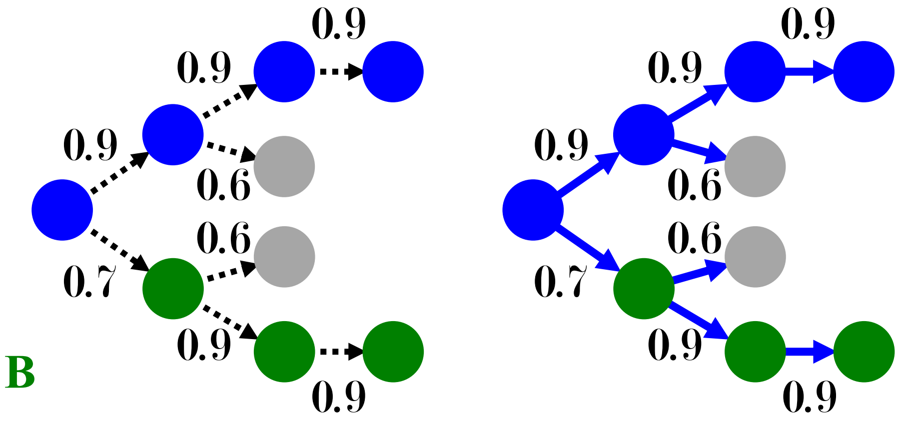
Track Building Algorithm – Baseline

Cluster Indices	0	1
Node Indices	0, 2, 3	1, 4

- **Retention of Edges Exceeding Probability Threshold**
  Distance Matrix ΔR_{ij} Generation for Candidate Edges
- **Density-Based Spatial Clustering of Applications with Noise**
  - Exhaustive Recursive Clustering for Localized Node Neighborhoods
    Number of Minimum Points = 2 in Neighborhood Range = 0.4
  - Treating Intra-Cluster Nodes as Individual Particle Paths

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

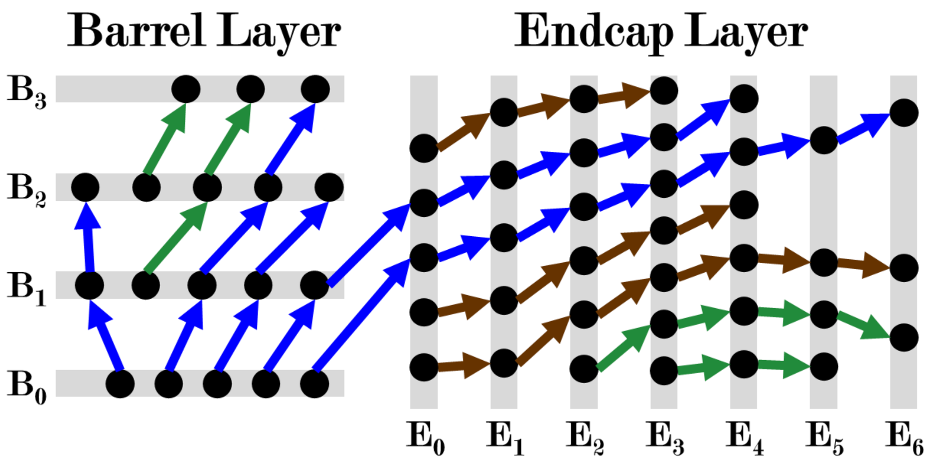
- **Inefficiency in Computation and Parallelism**
  - Dense Matrix Interpretation for Sparse Graph
    Time-Complexity $O(V^2)$ for Pair-Wise Distance Querying
  - Scheduling Requirements of Partitioned Intra-Cluster Node
- **Degradation of Accuracy**
  - Complete Omission of Probability Information
  - Suboptimal Differentiation through ΔR Metrics
  - Bifurcated Trajectory Paths



● : **Particle A** ● : **Particle B**

Track Building Algorithm – Proposed

- **Establishment of Node Index Tables**
  - ***Target / Untargeted Node Index Table***
    Outgoing Edge with Highest Probability above Threshold
  - ***Layer-Level Parallelism***
    Partial Data-Level Parallelism for Starting Node Collector
  - ***Sparse COO Format with $O(E) \approx O(V)$ Complexity***
- **Probability-Based Sequential Track Building**
  - ***LUT-Based Target-Node Mapping***
    Initiated from B0 (Blue), E0 (Brown), Others (Green)
  - ***Fine-Grained Node-Level Parallelism***
    Attaining Constant-Time Complexity O(1)



Edge Stream						
Source Indices	0	0	1	1	2	…
Target Indices	1	2	3	4	5	
Probability	0.9	0.8	0.9	0.2	0.1	

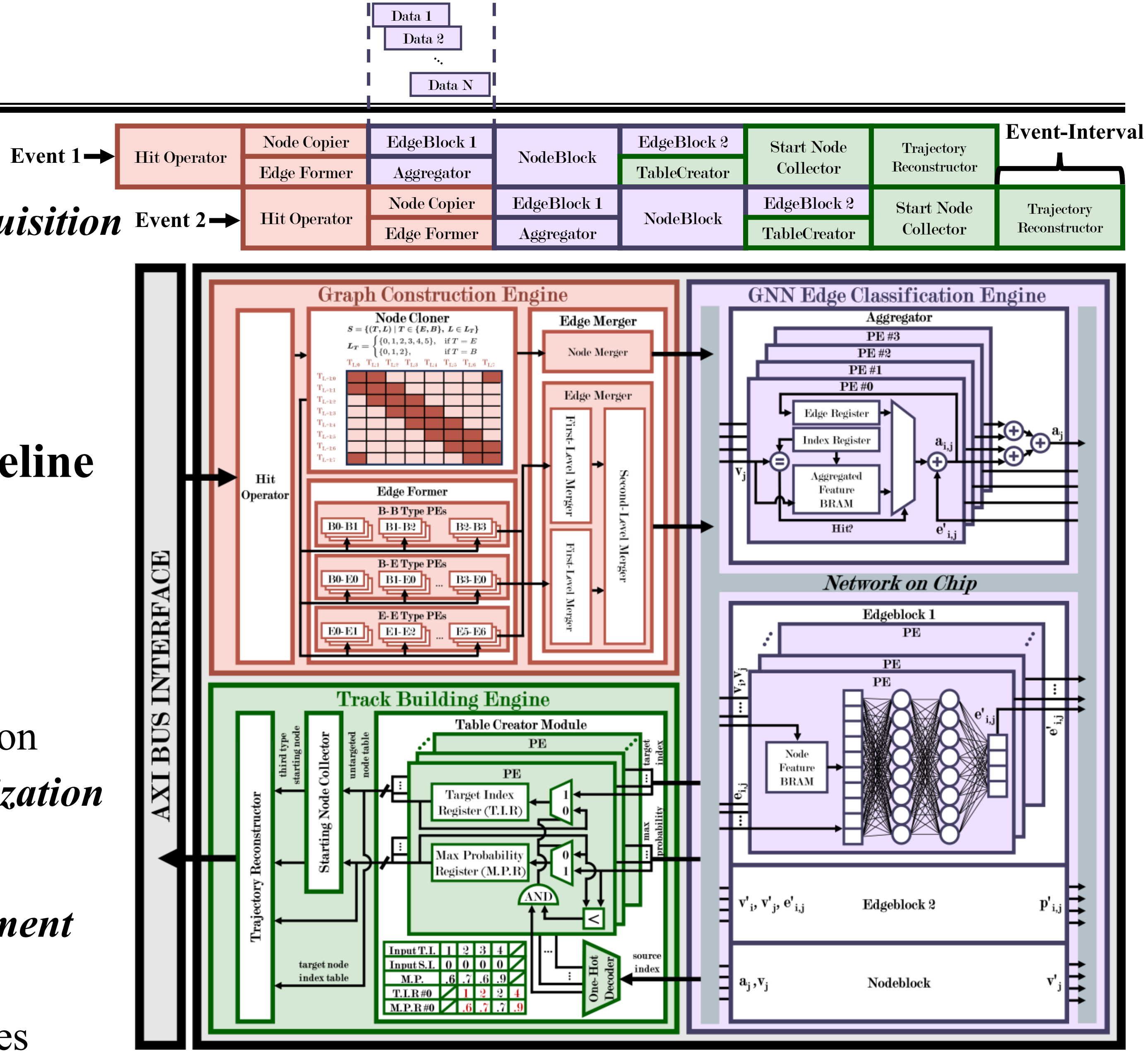
Target Node Index Table				
Source Indices	0	1	2	…
Target Indices	1	3	N/A	

Untargeted Node Index Table						
Node Indices	1	2	3	4	5	…
Value	F	T	F	T	T	

Graph Analysis				Cycle-Based Analysis			
Design	π/4	π/2	π	Design	π/4	π/2	π
# Nodes	113	201	378	Baseline	12,769	40,401	142,884
# Edges	196	334	596	Proposed	119	186	223
Ratio	1.73	1.65	1.57	**Speedup**	**107**	**217**	**641**

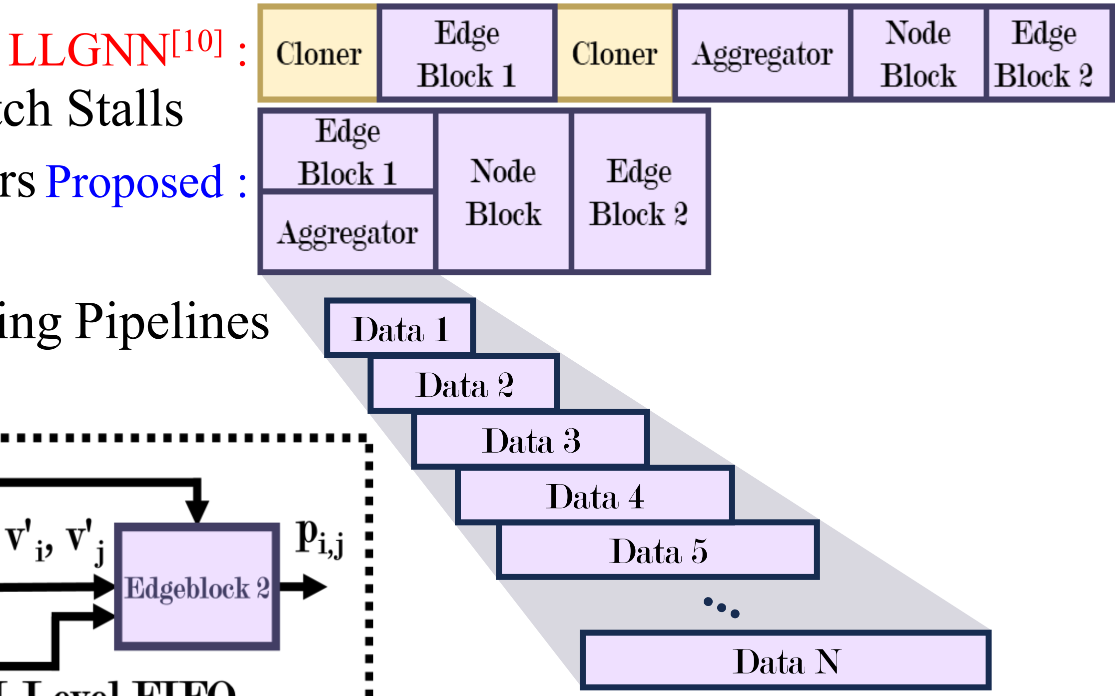
Overview – Architecture

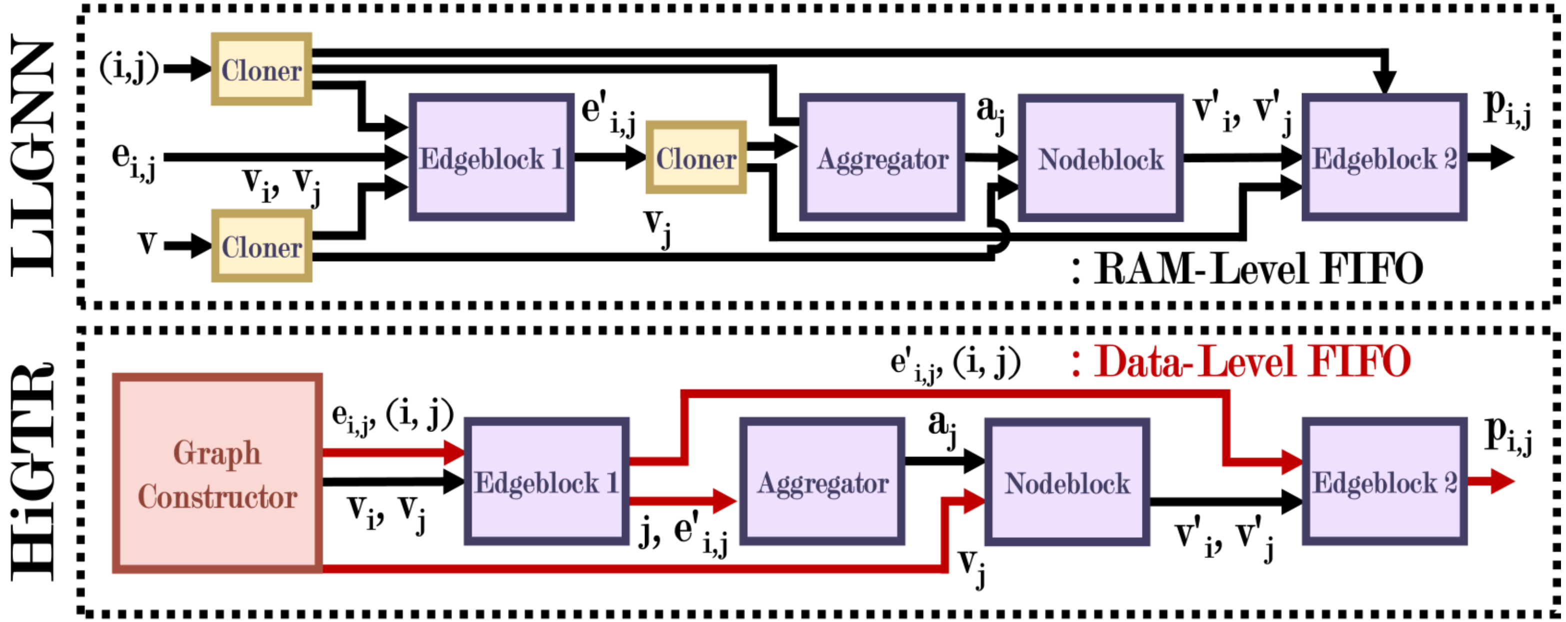
- **Consecutive-Collision Processing**
  - ***AXI4-Stream Multi-Point Event Acquisition***
    - Graph Construction Engine
    - GNN Edge Classification Engine
    - Track Building Engine
- **Throughput-Tuned Processing Pipeline**
  Support Multi-Level Granularity
  - ***Module-Level Execution Pipeline***
    - Event-Interval Balancing
  - ***Data-Level Processing Pipeline***
    - Cycle-Wise Intra-Module Ingestion
  - ***Maximized Hardware-Resource Utilization***
- **Latency-Tuned Data Streaming**
  - ***Compact FIFO with Interface Alignment***
    - Reduced Inter-Module Footprint
    - Overlap Execution across Modules



Edge-Classification Architecture – Proposed

- **Data Streaming–Oriented High-Throughput Paradigm**
  - ***Limitations of Batch Processing***
    - Latency Amplification by Downstream Batch Stalls
    - High RAM Demand in Batch Data Transfers
  - ***Advantages of Data Streaming***
    - Early Downstream Processing in Overlapping Pipelines
    - Minimal FIFO for Continuous Streams





	LLGNN	HiGTR	Gain
#LUT	161,308	145220	9.8%
#Flip-Flop	128826	125506	2.9%
#BRAM	24	18	25%
Latency	2.86 μs	1.365 μs	52.3%

Experiment

- **FPGA Platform: AMD-Xilinx Virtex UltraScale+ VU9P**
  Designated FPGA Platform for L1T Subsystem in HL-LHC
  - ***Development Toolkit*** – Vitis HLS 2023.2
  - ***Resource Utilization*** – Vivado 2023.2 Post-Place-And-Route Metric
  - ***Operating Frequency*** – 200 MHz
  - ***Evaluation Dataset*** – 1,000 Collision Events from TrackML[12]

Final Performance Metrics		
Design	***Throughput***	***Latency***
π/4	2.35 MHz	2.36 μs
π/2	2.24 MHz	2.90 μs
π	1.53 MHz	3.80 μs
Target	2.22 MHz	4.00 μs

※ *Resource overutilization for π design*

Final Accuracy Metrics			
Design	***Baseline[7,9] (Software)***	***Proposed (Software)***	***Proposed (Hardware)***
π/4	87.31% / 93.04%	90.35% / 97.22%	88.38% / 95.89%
π/2	86.04% / 91.25%	91.64% / 97.63%	89.57% / 96.92%
π	84.75% / 89.76%	92.34% / 97.94%	90.81% / 96.12%

※ *Perfect match efficiency / double-majority efficiency*

Comparison Table for π/2 Design Latency				
	Graph Construction	***Edge Classification***	***Track Building***	***Entire Flow***
Software[7,9]	187 ms	0.58 ms	0.99 ms	188.57 ms
Proposed	1.47 μs	1.36 μs	0.93 μs	2.90 μs
Speedup	130,769	426	1,065	65,024

[12] M. Kiehn, S. Amrouche, P. Calafiura, V. Estrade, S. Farrell, C. Germain, V. Gligorov, T. Golling, H. Gray, I. Guyon et al., "The trackml high-energy physics tracking challenge on kaggle," in EPJ Web of Conferences, vol. 214. EDP Sciences, 2019, p. 06037

Conclusion

- **First End-to-End GNN-Driven FPGA Accelerator for Trajectory Reconstruction**
  - ***Geometry-Aware Edge Pruning in Graph Construction***
    - Edge Count Reduction Achieving **37.5%–71.0%** Pruning [7]
  - ***Linear-Time Probability-Driven Sequential Track Building***
    - Latency Reduction Ranging from **107× to 641×** [7]
    - Enhanced Tracking Accuracy Metrics
  - ***Data-Streaming-Oriented High-Throughput Paradigm***
    - Latency Reduction Attaining **52.3%** [10]
  - ***Consolidated Three-Stage Pipeline within High-Performance FPGA Accelerator***
    - **65024× Acceleration** over Conventional Software-Based Approach [7,9]
- **Performance Alignment Coupled with Accuracy Enhancements**
  - Substantial Potential for Practical Deployment within L1T System
  - Especially Significant for HL-LHC Upgrades

Reference

- [1] L. Evans, “The large hadron collider,” New Journal of Physics, vol. 9, no. 9, p. 335, 2007.
- [2] “The Phase-2 Upgrade of the CMS Level-1 Trigger,” CERN, Geneva, Tech. Rep., 2020, final version.
- [3] O. Aberle, C. Adorisio, A. Adraktas, M. Ady, J. Albertone, L. Alberty, M. Alcaide Leon, A. Alekou, D. Alesini, B. Almeida Ferreira et al., “High-luminosity large hadron collider (hl-lhc): Technical design report,” 2020.
- [4] “The Phase-2 Upgrade of the CMS Tracker,” CERN, Geneva, Tech. Rep., 2017.
- [5] R. Frühwirth, “Application of kalman filtering to track and vertex fitting,” Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 262, no. 2-3, pp. 444–450, 1987.
- [6] P. Battaglia, R. Pascanu, M. Lai, D. Jimenez Rezende et al., “Interaction networks for learning about objects, relations and physics,” Advances in neural information processing systems, vol. 29, 2016.
- [7] G. DeZoort, S. Thais, J. Duarte, V. Razavimaleki, M. Atkinson, I. Ojalvo, M. Neubauer, and P. Elmer, “Charged particle tracking via edge-classifying interaction networks,” Comput. Softw. Big Sci., vol. 5, no. 1, pp. 1–13, 2021.
- [8] X. Ju, D. Murnane, P. Calafiura, N. Choma, S. Conlon, S. Farrell, Y. Xu, M. Spiropulu, J.-R. Vlimant, A. Aurisano et al., “Performance of a geometric deep learning pipeline for hl-lhc particle tracking,” The European Physical Journal C, vol. 81, pp. 1–14, 2021.
- [9] A. Elabd, V. Razavimaleki, S.-Y. Huang, J. Duarte, M. Atkinson, G. DeZoort, P. Elmer, S. Hauck, J.-X. Hu, S.-C. Hsu et al., “Graph neural networks for charged particle tracking on fpgas,” Frontiers in big Data, vol. 5, p. 828666, 2022.
- [10] S. Huang, Y. Yang, Y. Su, B. Lai, J. Duarte, S. Hauck, S. Hsu, J. Hu, and M. S. Neubauer, “Low latency edge classification gnn for particle trajectory tracking on fpgas,” in 2023 33rd International Conference on Field-Programmable Logic and Applications (FPL). Los Alamitos, CA, USA: IEEE Computer Society, sep 2023, pp. 294–298.
- [11] Aneesh Heintz, Vesal Razavimaleki, Javier Duarte, Gage DeZoort, Isobel Ojalvo,Savannah Thais, Markus Atkinson, Mark Neubauer, Lindsey Gray, Sergo Jindari-ani, et al. 2020. Accelerated charged particle tracking with graph neural networkson FPGAs. arXiv preprint arXiv:2012.01563 (2020).
- [12] M. Kiehn, S. Amrouche, P. Calafiura, V. Estrade, S. Farrell, C. Germain, V. Gligorov, T. Golling, H. Gray, I. Guyon et al., “The trackml high-energy physics tracking challenge on kaggle,” in EPJ Web of Conferences, vol. 214. EDP Sciences, 2019, p. 06037

Question and Answer

Yun-Chen Yang, Hao-Chun Liang, and Bo-Cheng Lai

Institute of Electronics
National Yang Ming Chiao Tung University

Presenter: Hao-Chun, Liang
Supervisor: Bo-Cheng Lai

NATIONAL
YANG MING CHIAO TUNG
UNIVERSITY