GenGNN: A Generic FPGA Acceleration Framework for Graph Neural Networks

Stefan Abi-Karam*, Yuqi He*, Rishov Sarkar*, Lakshmi Sathidevi, Zihang Qiao, **Cong Hao**

Sharc-lab @ Georgia Tech https://sharclab.ece.gatech.edu/

Acknowledgement

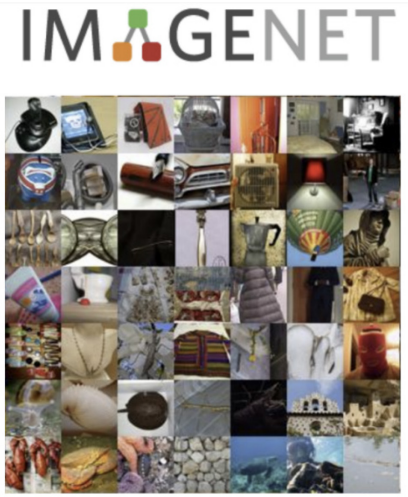
- All works are done at CRNCH Rogues Gallery
- Technical support by Dr. Jeffrey Young

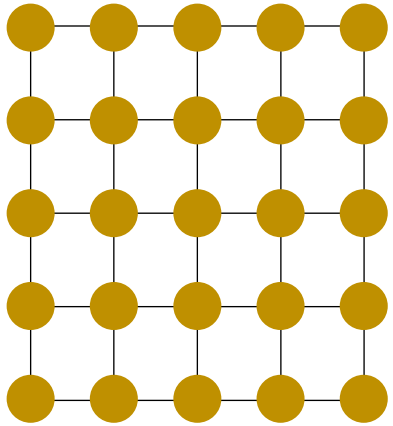
Outline

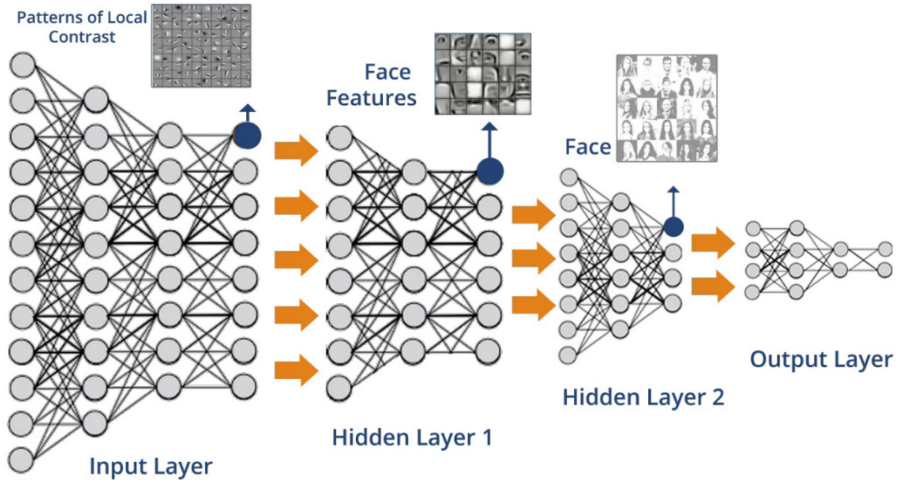
- **Background and Motivation**
  - What are Graph Neural Networks (GNNs)?
  - Why accelerating GNNs?
- **Existing Accelerator Limitations**
  - Not generic, not real-time
- **Ours: Generic GNN Accelerator – GenGNN**
  - Generic message passing framework
  - Model-specific components
- **Evaluation: CPU/GPU**

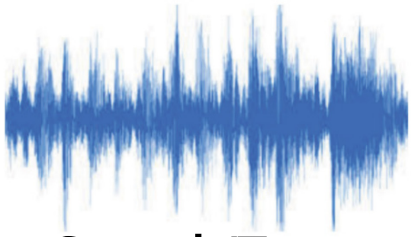
What is Graph Neural Network (GNN)

- **Traditional neural networks are designed for simple sequences & grids**

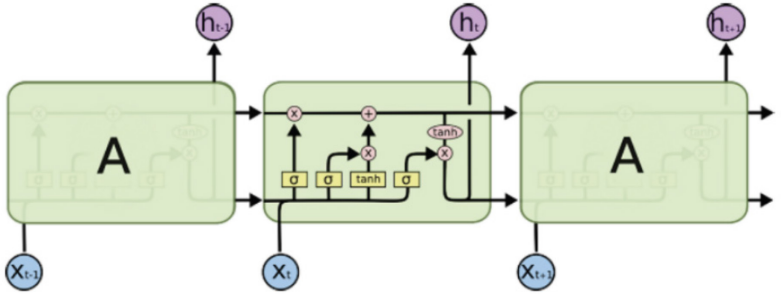








Speech/Text



[Slide credit: http://web.stanford.edu/class/cs224w]

What is Graph Neural Network (GNN)

- **Reality: A lot of real-world data does not "live" on grids**
  - Arbitrary size and complex topological structure
  - No fixed node ordering or reference point

Image credit: Medium

Social Networks

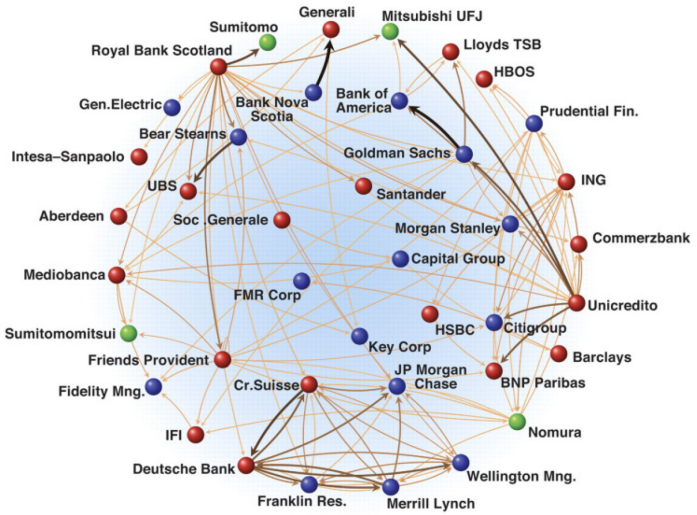


Image credit: Science

Economic Networks

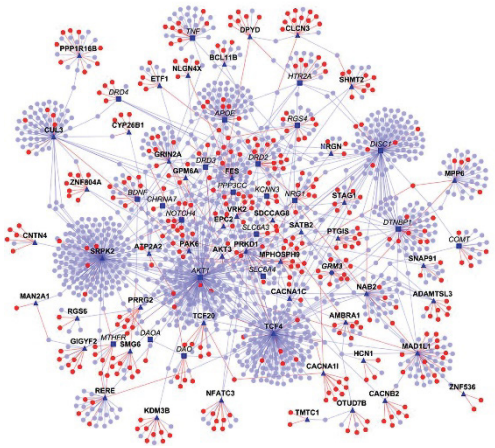
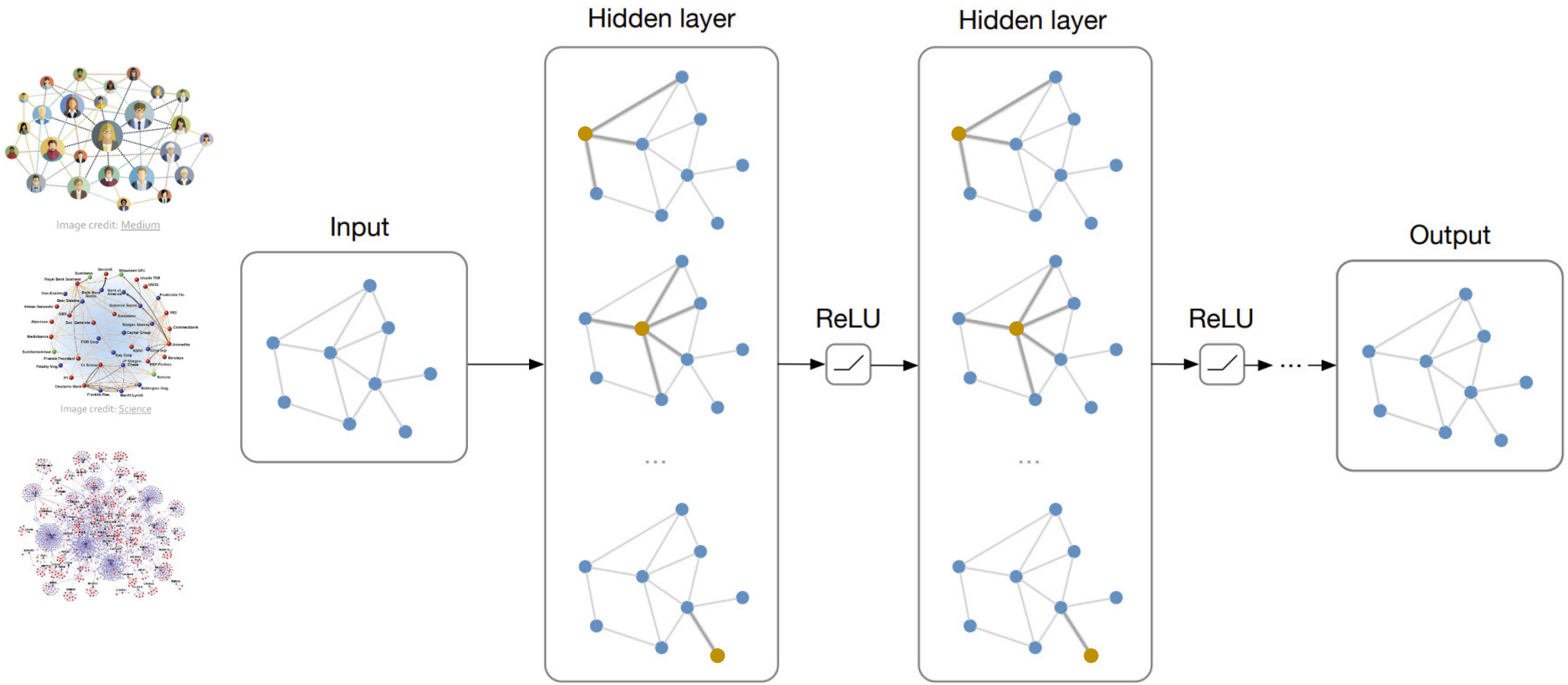


Image credit: Madhavicmu / Wikimedia Commons/CC-BY-SA-4.0

Protein Interaction Networks

What is Graph Neural Network (GNN)

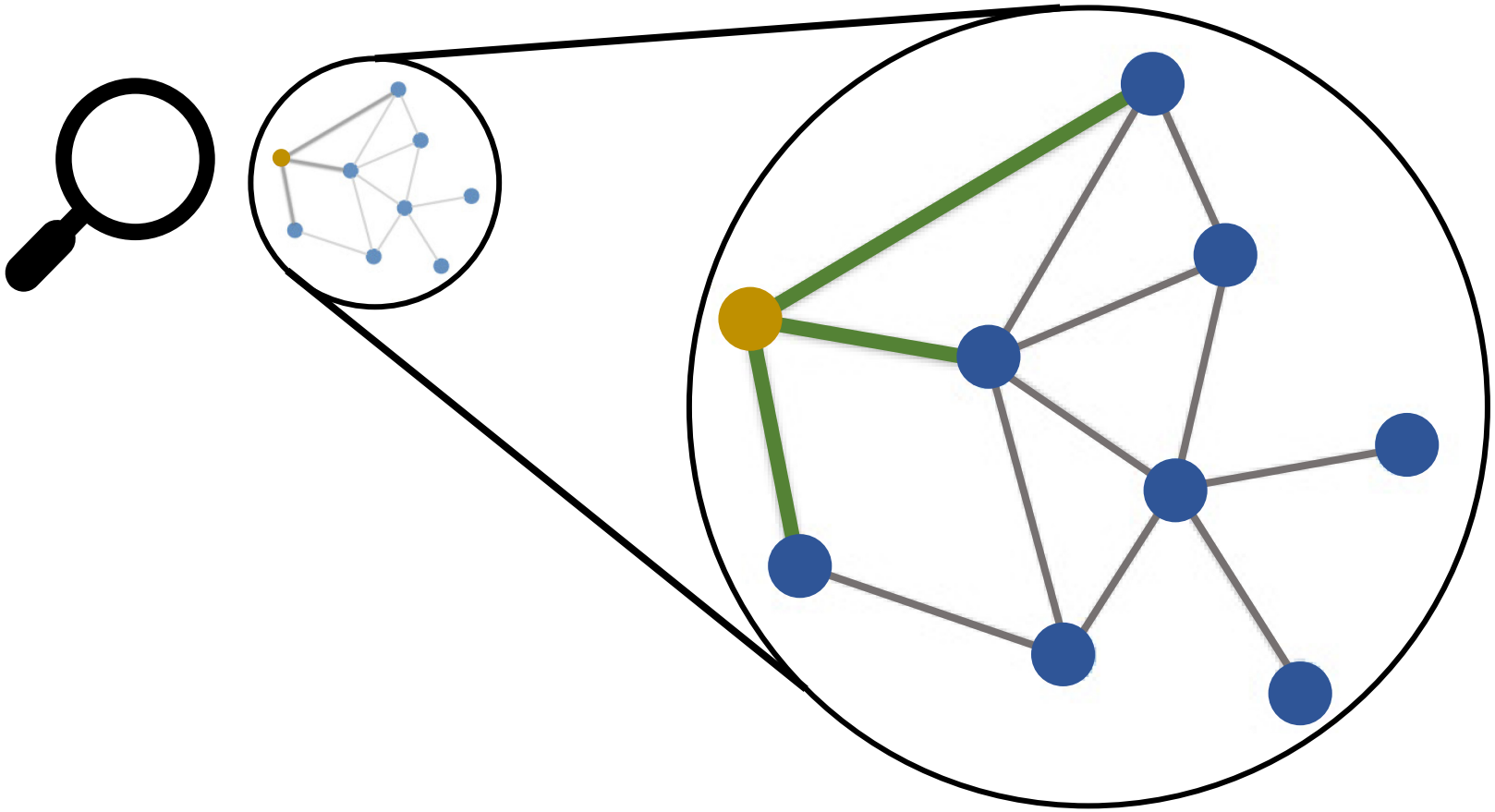
- **Mainstream:** Pass **messages** between pairs of nodes, **aggregate**, and **transform**



[Slide credit: Structured deep models: Deep learning on graphs and beyond]

What is Graph Neural Network (GNN)

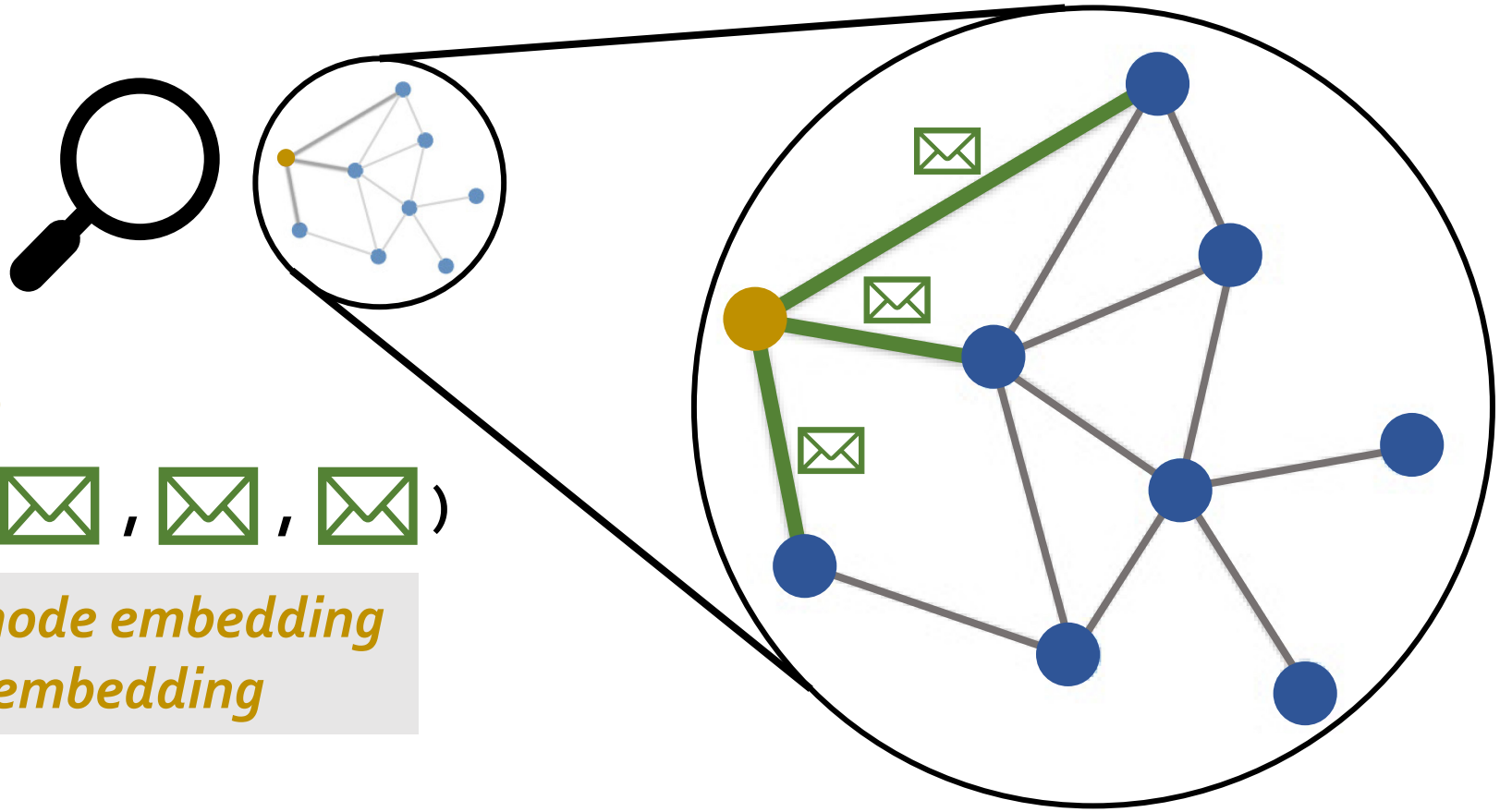
- Mainstream: Pass messages between pairs of nodes, aggregate, and transform



[Slide credit: Structured deep models: Deep learning on graphs and beyond]

What is Graph Neural Network (GNN)

- Mainstream: Pass messages between pairs of nodes, aggregate, and transform



✓ Aggregate

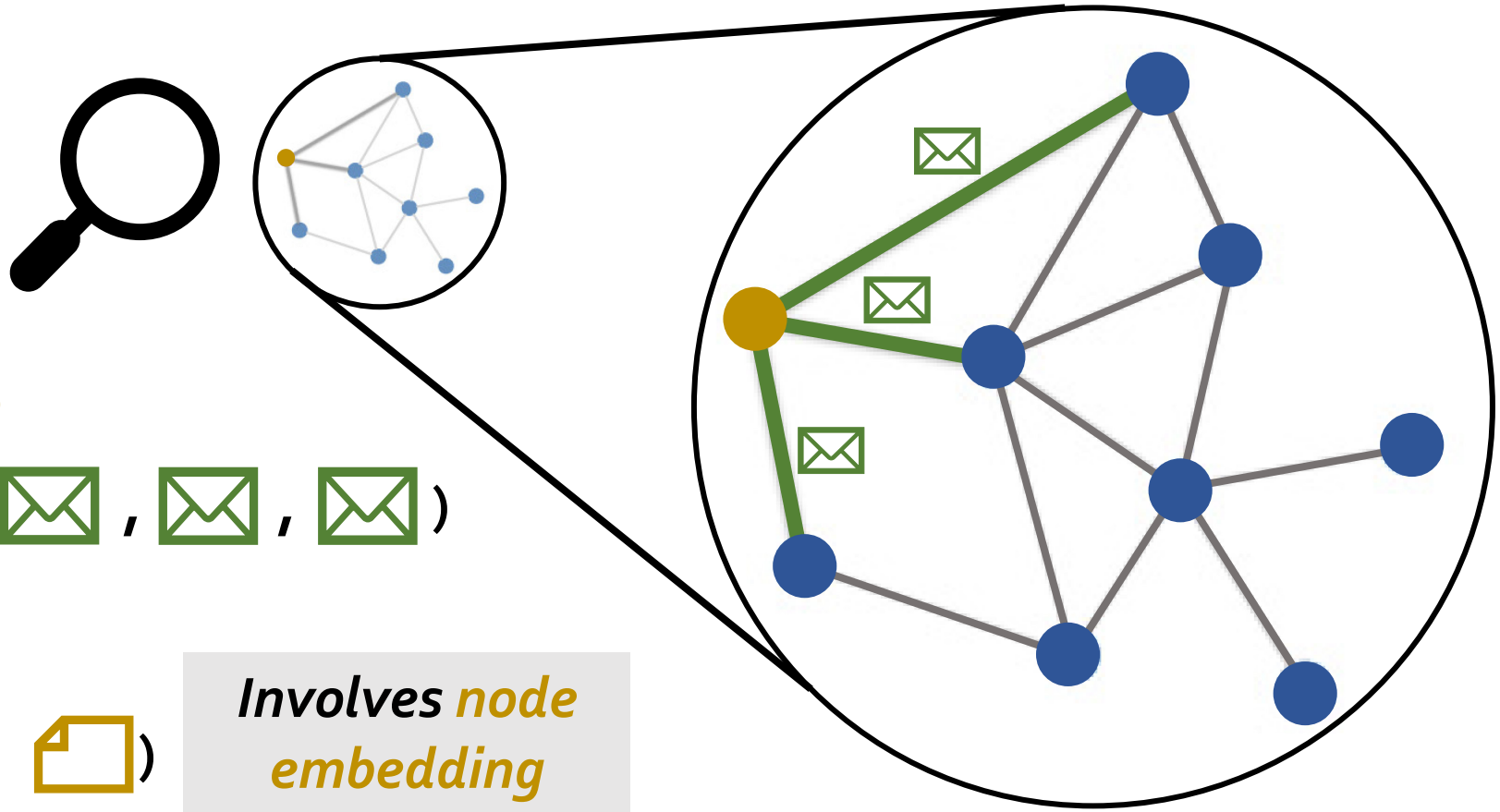
✉=Sum/Max(✉, ✉, ✉)

Involves both node embedding and edge embedding

[Slide credit: Structured deep models: Deep learning on graphs and beyond]

What is Graph Neural Network (GNN)

- Mainstream: Pass messages between pairs of nodes, aggregate, and transform



✓ Aggregate

=Sum/Max(, ,)

✓ Transform

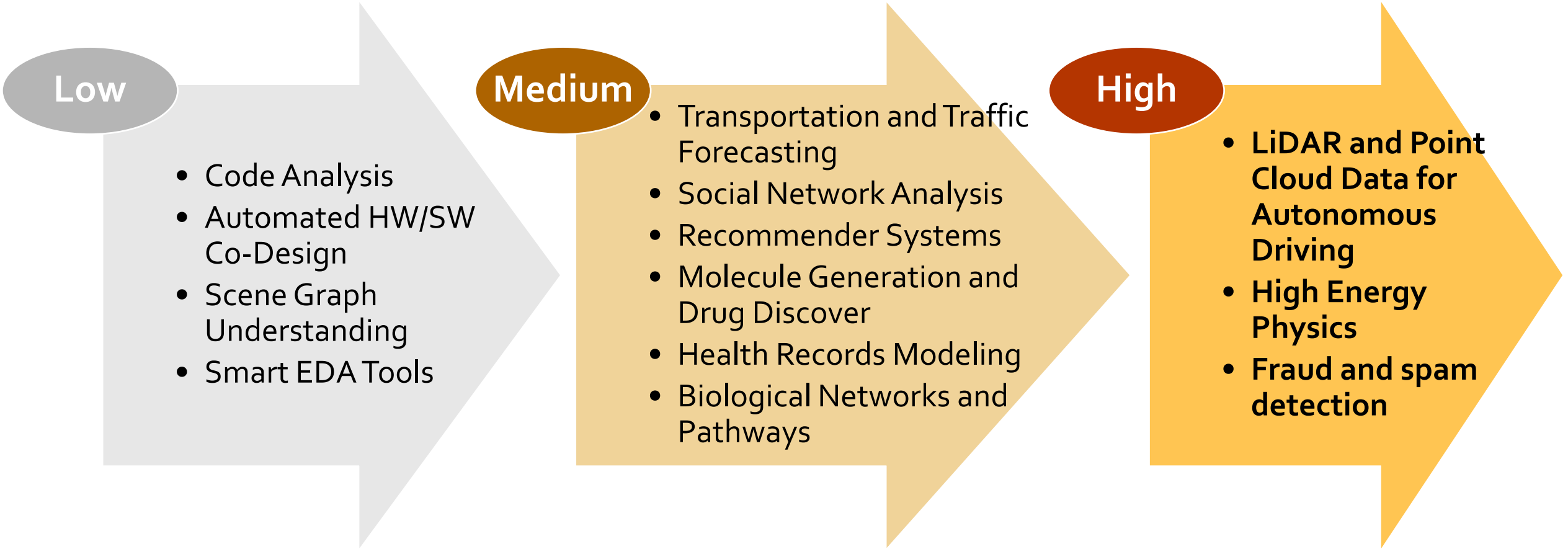
=MLP(,)

Involves node embedding

[Slide credit: Structured deep models: Deep learning on graphs and beyond]

Why Accelerating GNN Matters?

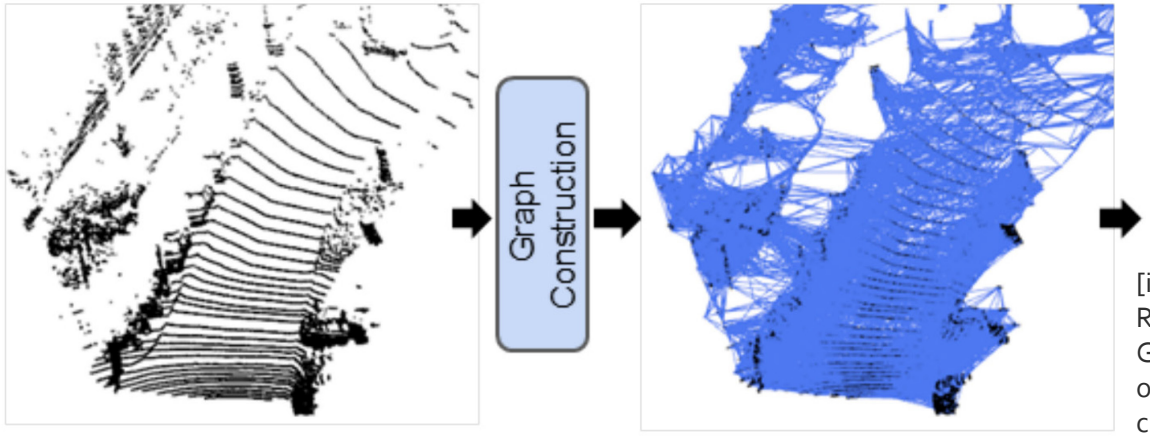
- **Numerous applications; many require real-time processing**



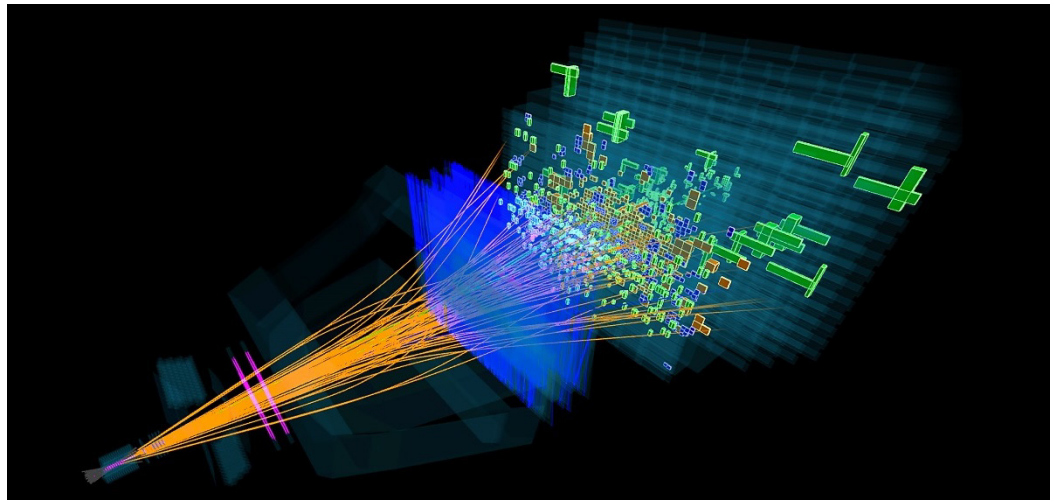
Zhou, Jie, et al. "Graph neural networks: A review of methods and applications." *AI Open*, 2020

Why Accelerating GNN Matters?

- **Numerous applications; many require real-time processing**



[image source] Shi, Weijing, and Raj Rajkumar. "Point-gnn: Graph neural network for 3d object detection in a point cloud." CVPR 2020



[image source] https://www.quantamagazine.org/growing-anomalies-at-the-large-hadron-collider-hint-at-new-particles-20200526/

High

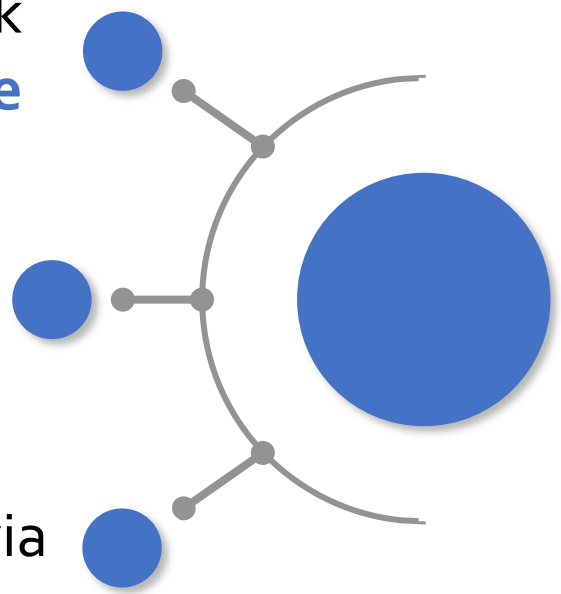
- **LiDAR and Point Cloud Data for Autonomous Driving**
- **High Energy Physics**
- **Fraud and spam detection**

Existing Work

Most focus on Graph Convolution Network (GCN): A **limited type**

Heavy **pre-processing**: Not suitable for real-time

Most on ASIC via **simulation**: not end-to-end, far from practical

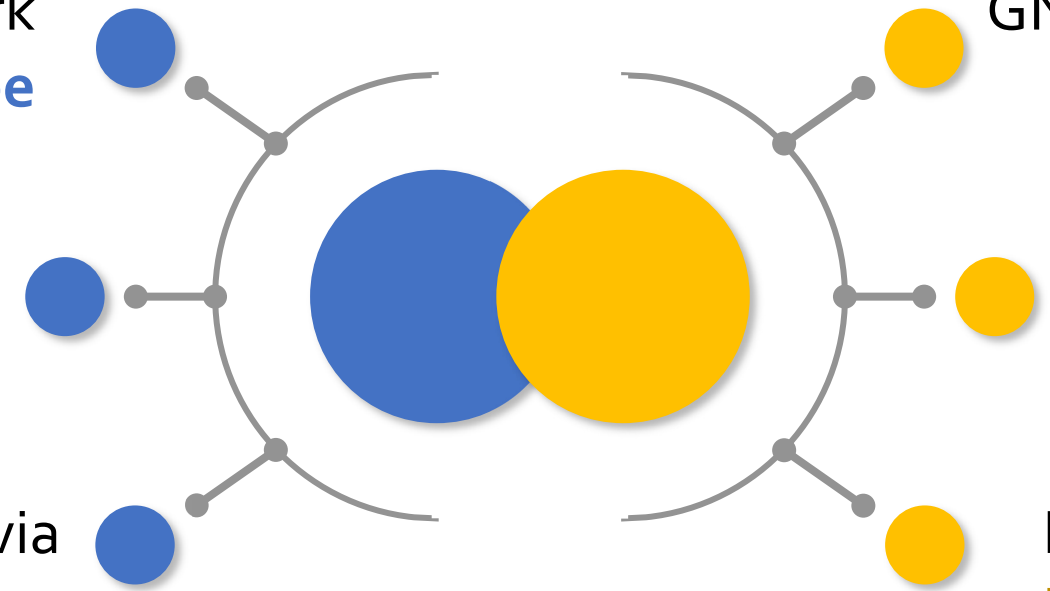


Existing Work v.s. Ours

Most focus on Graph Convolution Network (GCN): A **limited type**

Heavy **pre-processing**: Not suitable for real-time

Most on ASIC via **simulation**: not end-to-end, far from practical



Support a wide range of GNNs: **Generic**

No pre-processing: **Real-time** oriented

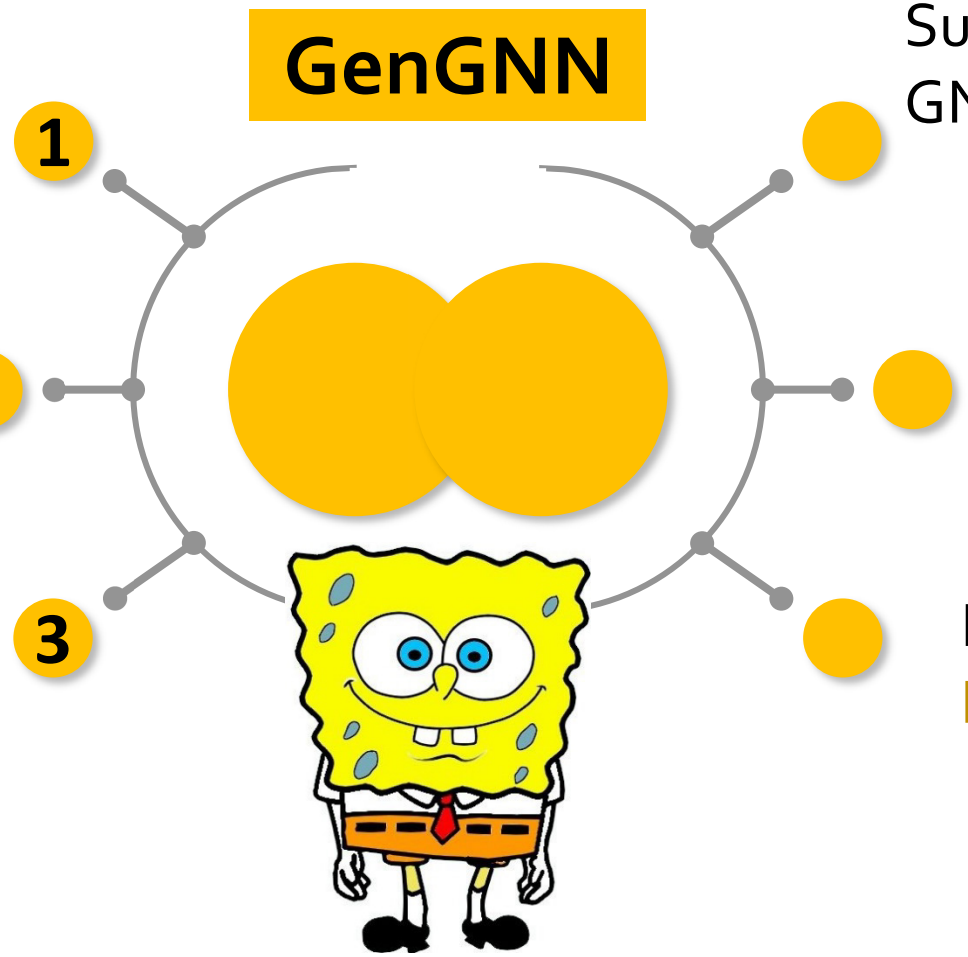
End-to-end **open-source FPGA** implementation

GenGNN Features

A general framework for **message passing** (a common mechanism)

A library for **model-specific** components

Fully **verified functionality**, written in High-Level Synthesis (**HLS**)

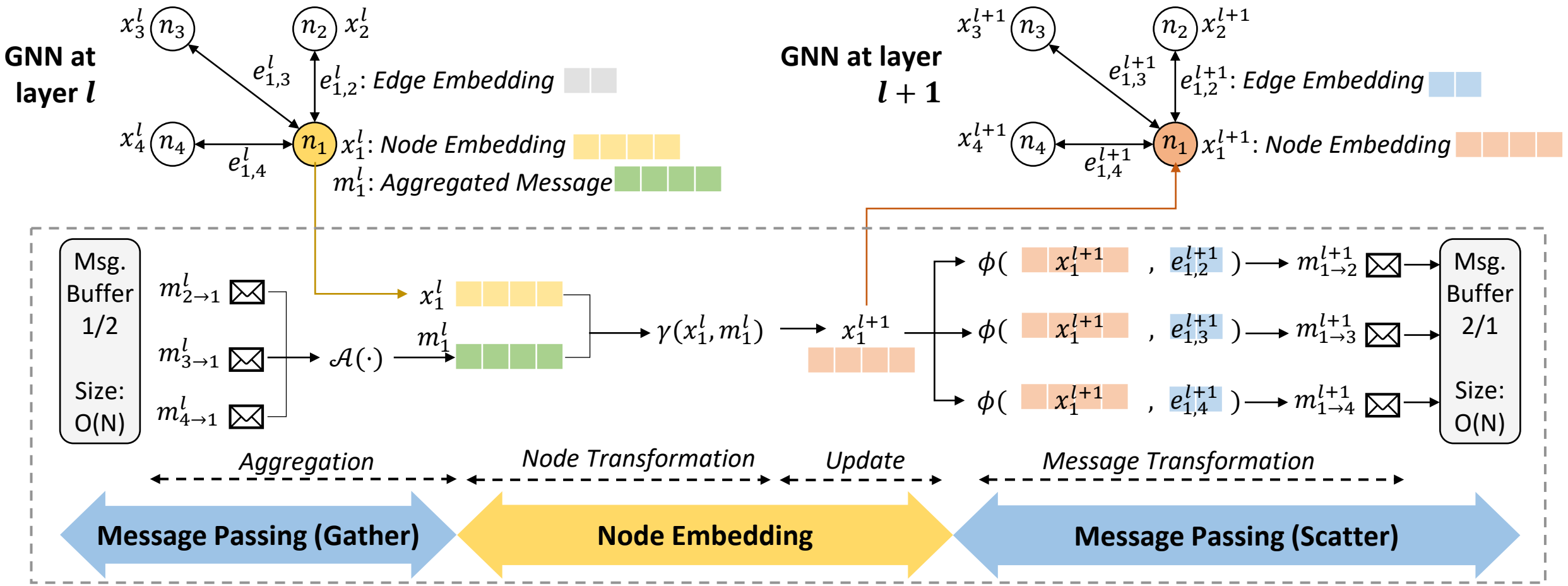


Support a wide range of GNNs: **Generic**

No pre-processing: **Real-time** oriented

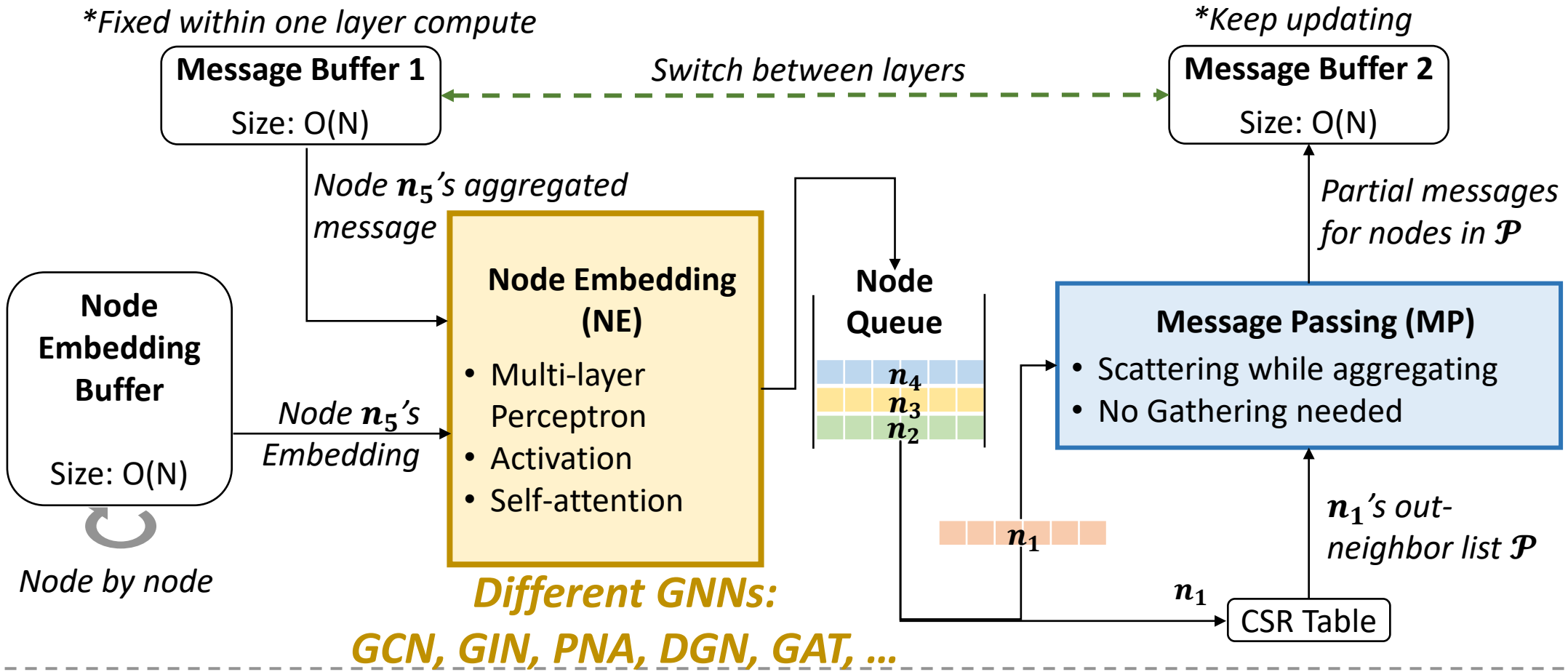
End-to-end **open-source FPGA** implementation

Message Passing in GNNs

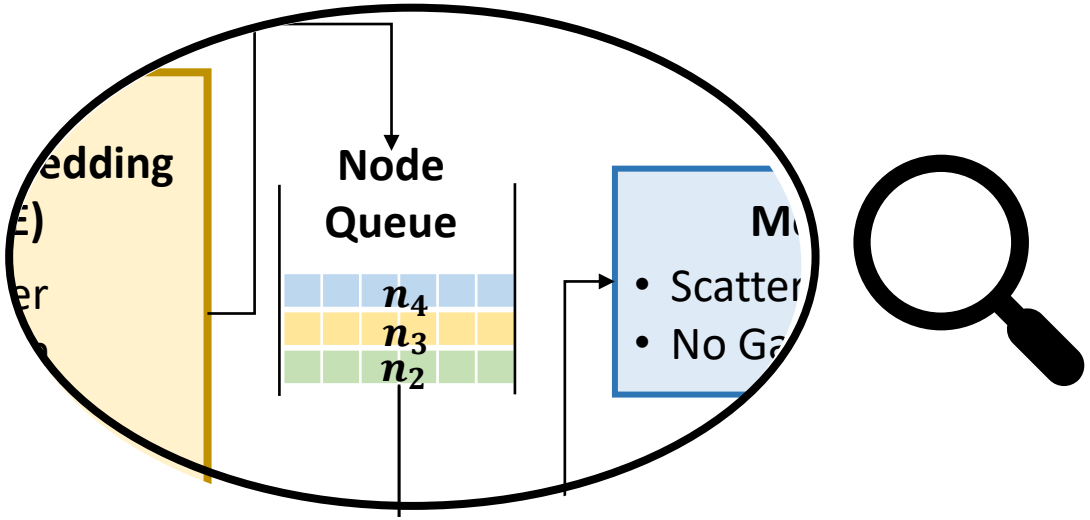


Architecture to Support Message Passing

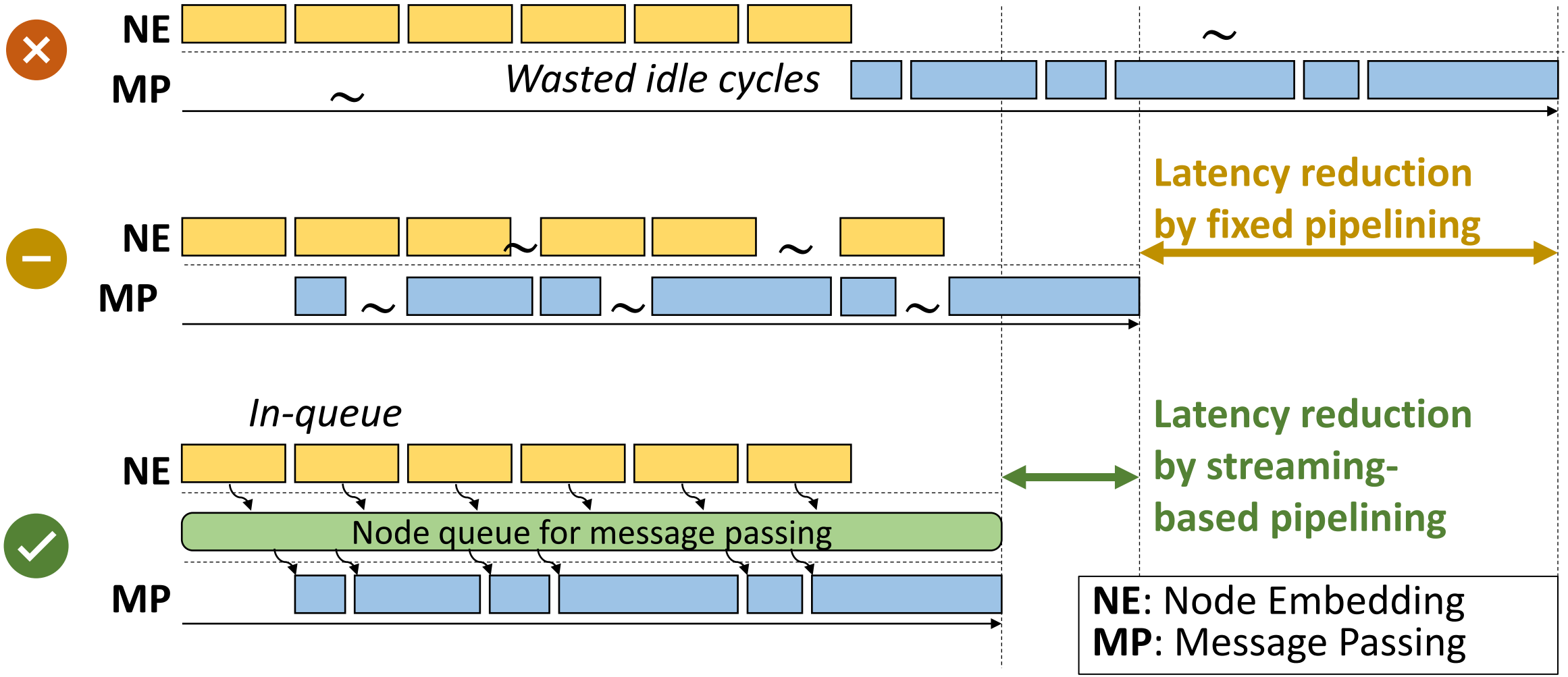
Overall Architecture of GenGNN (layer-recursive)



Streaming-based Pipelining



Streaming-based Pipelining



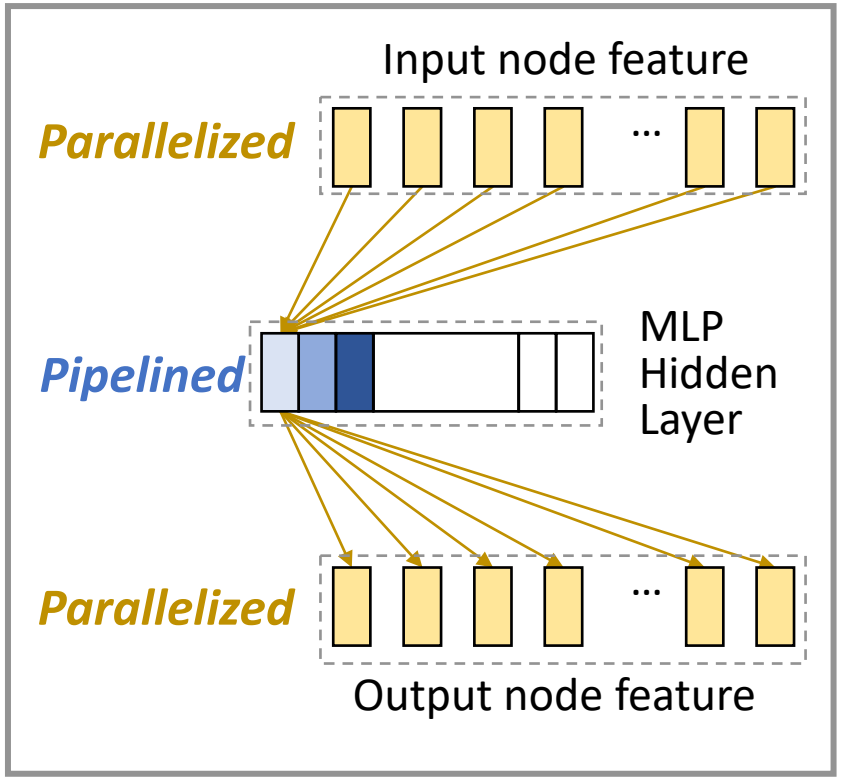
Supported GNNs (so far)

Represent GNN Family that…

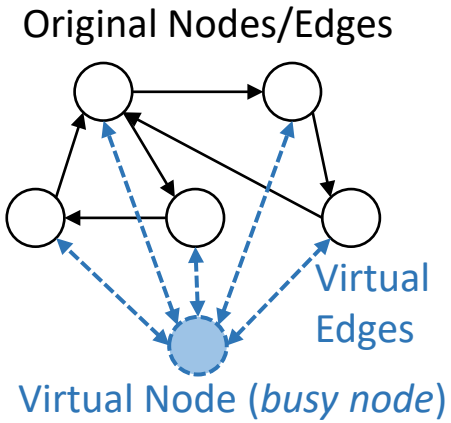
GCN	Can be represented as sparse matrix-matrix multiplication (SpMM)
GIN	With edge embedding and transformation where SpMM does not apply
GAT	With self-attention and possibly with edge embeddings
PNA	Arbitrarily uses multiple aggregation methods
DGN	With a directional flow at each node for guided aggregation
+VN	With virtual nodes connecting to all other nodes

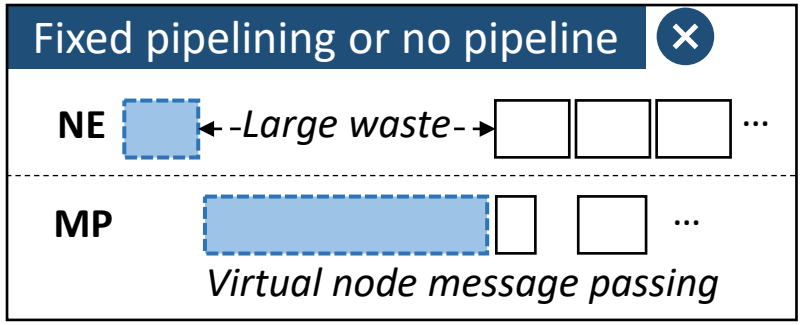
Examples of Model-Specific Optimization

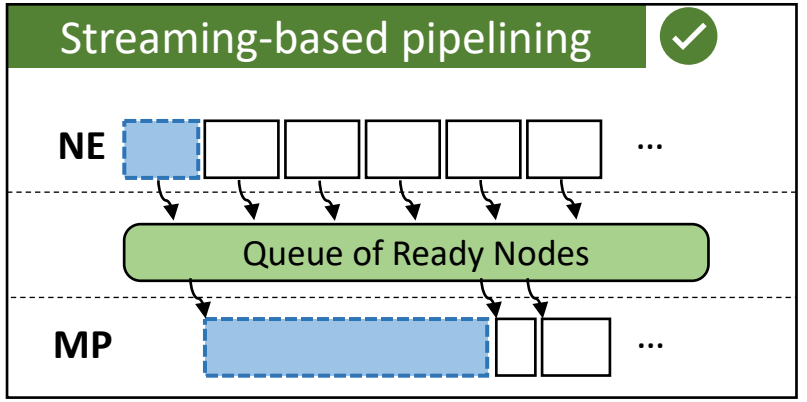
Customized Multi-layer perceptron (MLP) for GIN model



Virtual Nodes benefit from streaming-based pipelining

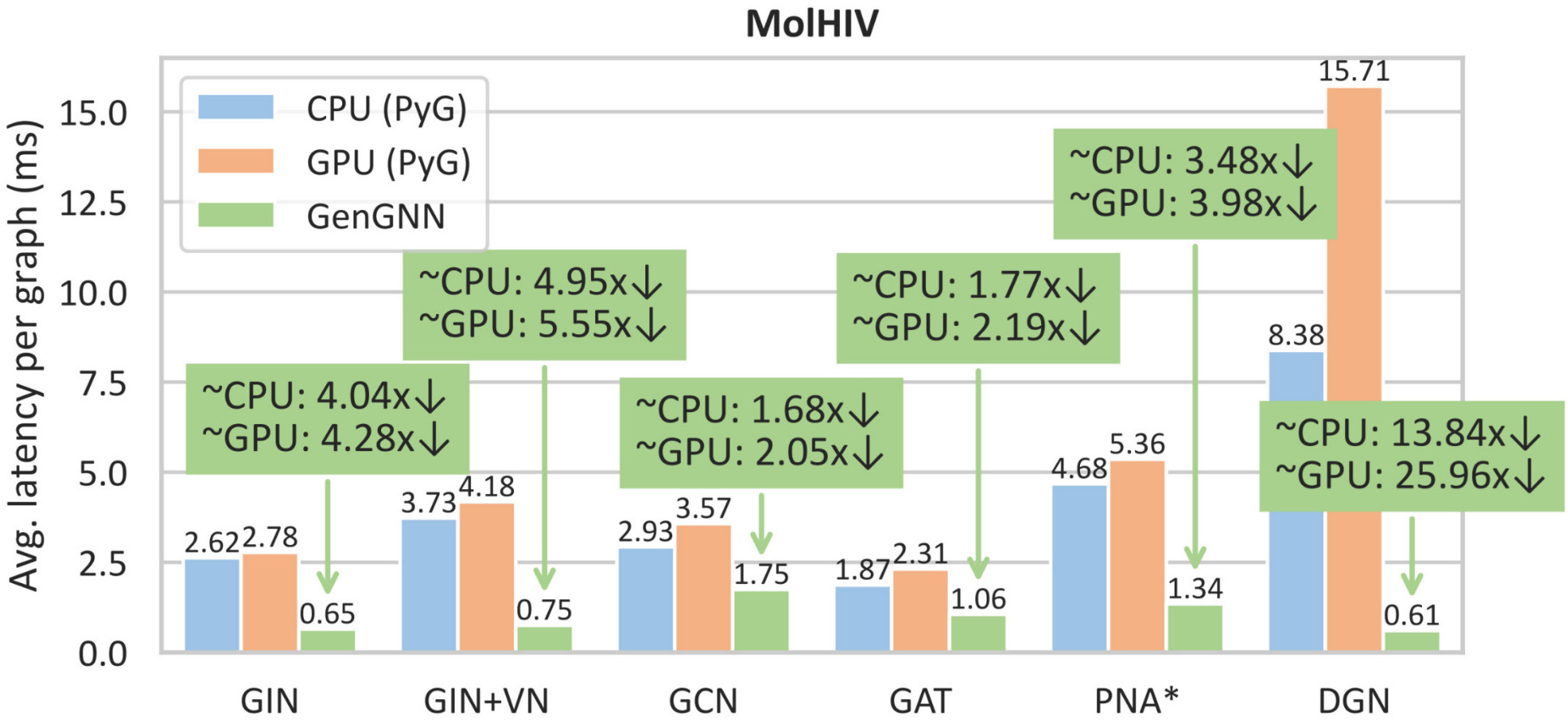






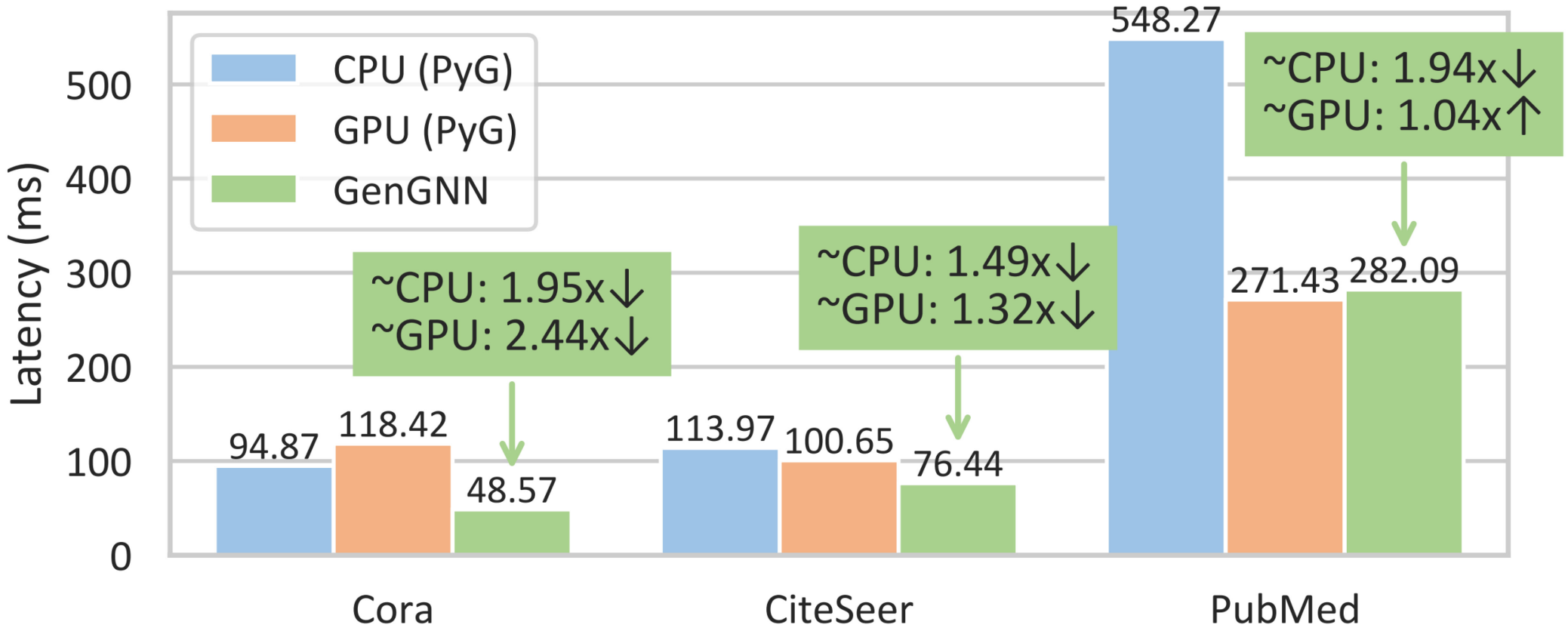
Evaluations on Small Graphs

- Baseline: CPU/GPU (batchsize = 1)
- Dataset: MolHIV and MolPCBA (molecule classification); 4k – 40k graphs



Evaluations on Large Graphs

- **Baseline: CPU/GPU (batchsize = 1)**
- **Dataset: Cora** (N=2,708)**, CiteSeer** (N=3,327)**, PubMed** (N=19,717)



Summary & Thanks!

- **Graph Neural Networks (GNNs) require acceleration and real-time processing**
  - Not all GNNs are sparse matrix multiplications (only few of them are…)
- **GenGNN: our GNN acceleration framework on FPGA**
  - Generic: supports a wide range of GNN models
  - Real-time: no pre-processing
  - Open-source: High-Level Synthesis (HLS) based
- **Beats GPU and CPU on both small and large graphs**
  - Future direction1: ultra-small graphs
  - Future direction2: ultra-large graphs
  - Future direction3: design automation and design space exploration