Sharc Lab
Prof. Callie Hao

Graph Neural Network Accelerators for FPGAs

GenGNN and Future Research Directions

Stefan Abi-Karam and Rishov Sarkar, School of Electrical and Computer Engineering, Georgia Tech

Georgia Tech | Center for Research into Novel Computing Hierarchies

Georgia Tech | School of Electrical and Computer Engineering
College of Engineering

Ultra-fast Small Graph Accelerator

Motivation

- Small graphs: graphs that fit entirely within on-chip memory, e.g., tens of nodes and edges
- Example: molecular graphs

Goal

- Extend GenGNN to enable **high throughput** for thousands of small graphs

Key points

- Parallelization across nodes & edges
- Batched processing of graphs
- Quantization with minimal loss of accuracy

Scalable Large Graph Accelerator

Motivation

- Large graphs: graphs too large to fit into on-chip memory, e.g., millions of nodes and edges
- Example: social network graphs

Goal

- Extend GenGNN to optimize latency for large graphs, keeping **scalability** in mind

Key points

- Reordering nodes & edges on-the-fly
- Smart caching of nodes & edges
- Scalable to any embedding dimension
- Design-space exploration for different graph types

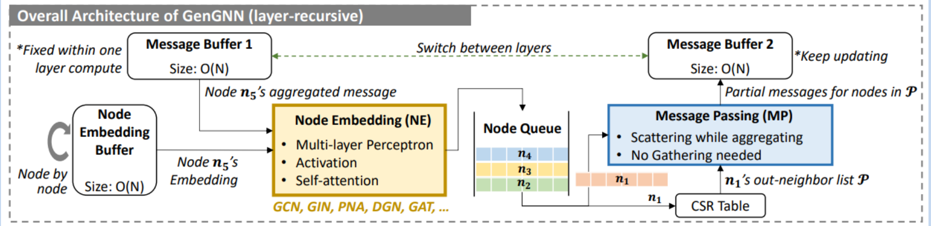
1

2

GenGNN: A Generic FPGA Framework for Graph Neural Network Acceleration

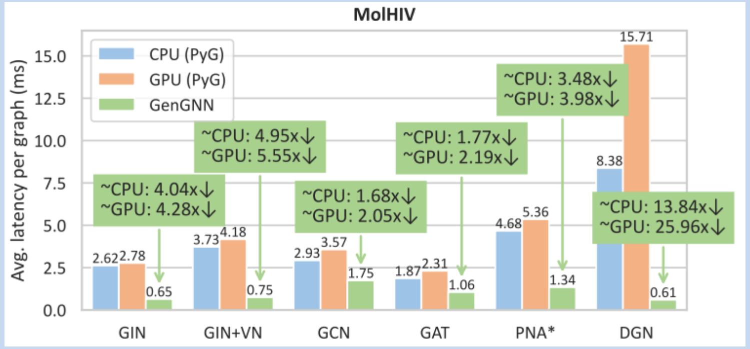
Stefan Abi-Karam*, Yuqi He*, Rishov Sarkar*, Lakshmi Sathidevi, Zihang Qiao, Cong Hao

(Submitted to FCCM 2022)



Two-fold goals

- Ultra-fast GNN inference with zero graph pre-processing for real-time requirements
- Support for diverse set of GNN models with extensibility to adapt to new models

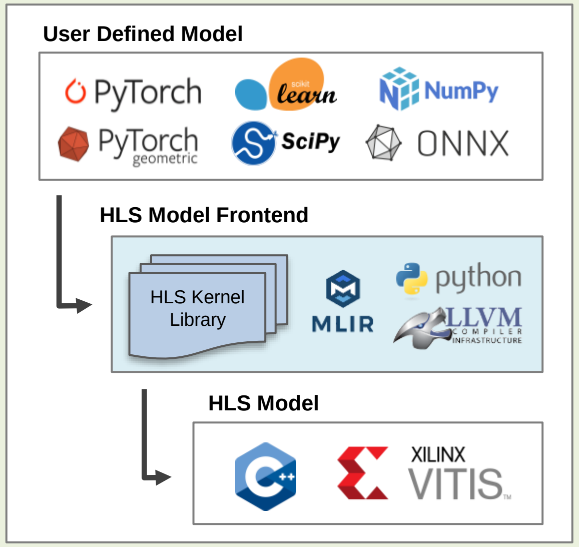


Implemented on Xilinx Alveo U50

- Up to 25x faster than CPU baseline
- Up to 13x faster than GPU baseline

3

Automated Accelerator Construction



Motivation

- GNN accelerators lack similar optimizations and repetition of same kernels across all models
- Programming effort high compared to design exploration

1 2 3

Ongoing directions based on GenGNN