

HP-GNN: Generating High Throughput GNN Training Implementation on CPU-FPGA Heterogeneous Platform

Yi-Chien Lin*
yichienl@usc.edu

University of Southern California
Los Angeles, California, USA

Bingyi Zhang*
bingyizh@usc.edu

University of Southern California
Los Angeles, California, USA

Viktor Prasanna
prasanna@usc.edu

University of Southern California
Los Angeles, California, USA

ABSTRACT

Graph Neural Networks (GNNs) have shown great success in many applications such as recommendation systems, molecular property prediction, traffic prediction, etc. Recently, CPU-FPGA heterogeneous platforms have been used to accelerate many applications by exploiting customizable data path and abundant user-controllable on-chip memory resources of FPGAs. Yet, accelerating and deploying GNN training on such platforms requires not only expertise in hardware design but also substantial development efforts.

We propose HP-GNN, a novel framework that generates high throughput GNN training implementations on a given CPU-FPGA platform that can benefit both application developers and machine learning researchers. HP-GNN takes GNN training algorithms, GNN models as the inputs, and automatically performs hardware mapping onto the target CPU-FPGA platform. HP-GNN consists of: (1) data layout and internal representation that reduce the memory traffic and random memory accesses; (2) optimized hardware templates that support various GNN models; (3) a design space exploration engine for automatic hardware mapping; (4) high-level application programming interfaces (APIs) that allows users to specify GNN training with only a handful of lines of code. To evaluate HP-GNN, we experiment with two well-known sampling-based GNN training algorithms and two GNN models. For each training algorithm and model, HP-GNN generates implementation on a state-of-the-art CPU-FPGA platform. Compared with CPU-only and CPU-GPU platforms, experimental results show that the generated implementations achieve 55.67 \times and 2.17 \times speedup on the average, respectively. Compared with the state-of-the-art GNN training implementations, HP-GNN achieves up to 4.45 \times speedup.

CCS CONCEPTS

• **Hardware** \rightarrow **Reconfigurable logic applications**; *Hardware accelerators*.

KEYWORDS

Graph Neural Networks; FPGA Framework; Hardware Acceleration

ACM Reference Format:

Yi-Chien Lin, Bingyi Zhang, and Viktor Prasanna. 2022. HP-GNN: Generating High Throughput GNN Training Implementation on CPU-FPGA Heterogeneous Platform. In *Proceedings of the 2022 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA '22)*, February 27-March 1, 2022, Virtual Event, CA, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3490422.3502359>

1 INTRODUCTION

Recently, Graph Neural Networks (GNNs) have shown great success in many fields including recommendation systems [27, 33], molecular property prediction [14], traffic prediction [16], etc. Initially, GNN was trained using the full graph [17] as the input, directly. However, as graphs become larger, full-graph GNN training becomes inefficient because the model is only updated once for each epoch. Moreover, huge amount of intermediate results needs to be stored in full-graph GNN training, which leads to high memory footprint [20]. To overcome the above issues, many sampling-based GNN training algorithms [2, 8, 14, 29] have been proposed for training the GNN models. Sampling-based methods first sample the full graph to produce mini-batches, and then take these mini-batches as input for training. The sampling-based mini-batch training methods demonstrate great advantages compared with full-graph training in terms of accuracy, generalization, and scalability for large-scale graphs [2, 8, 14, 29]. Therefore, state-of-the-art GNN frameworks [11, 23] adopt sampling-based mini-batch training algorithms.

Compared with the CPU-only platforms or CPU-GPU platforms, CPU-FPGA platforms are promising platforms to deploy GNN training since this can support customized memory access pattern [32] and data layout to reduce the substantial memory traffic and random memory access in GNN training. However, deploying GNN training on CPU-FPGA heterogeneous platform is challenging due to notorious development efforts that requires hardware design expertise. Though there are many FPGA-based GNN accelerators [12, 19, 26, 30, 31], previous works either focus on full-graph GNN inference, or a specific GNN model, or a specific GNN training algorithm, and no general framework has been proposed.

Motivated by the challenges, we propose HP-GNN, a framework for mapping GNN training on CPU-FPGA heterogeneous platform. We first formulate an high-level abstraction to describe the computation of sampling-based mini-batch GNN training. Then we develop optimized hardware templates based on the GNN abstraction. In order to reduce development effort and eliminate the need for hardware expertise, HP-GNN provides easy-to-use software programming APIs that allow fast development without the need for hardware programming. To achieve high throughput and automate the accelerator generation process, we develop a general design

*Both authors contributed equally to this research.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

FPGA '22, February 27-March 1, 2022, Virtual Event, CA, USA

© 2022 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-9149-8/22/02.

<https://doi.org/10.1145/3490422.3502359>

space exploration engine that optimizes the hardware configuration based on the selected GNN training algorithm. In addition, we propose a data layout and internal representation that reduces the memory traffic and random memory access. We summarize our contributions as follow:

- We propose a general framework for mapping various sampling based mini-batch GNN training onto a CPU-FPGA platform. We demonstrate the applicability of HP-GNN to various sampling algorithms and GNN models.
- We provide high-level and easy-to-use software programming interface that abstracts the hardware implementation details.
- To enable hardware mapping and high throughput GNN training, the proposed framework consists of the following optimizations:
 - Data layout and internal representation that reduce the memory traffic and random memory access caused by the irregular computation pattern of GNN.
 - Optimized hardware templates that support various widely-used GNN models.
 - General design space exploration algorithm that generates hardware design configuration to optimize the training throughput for various sampling methods and algorithmic parameters.
- We evaluate our framework using two state-of-the-art GNN models: GraphSAGE [14], GCN [17], along with two commonly used sampling methods: neighbor sampling [14], and subgraph sampling [29]. Running on a Xilinx Alveo U250 board hosted by a 64-core AMD processor, the experimental results show that the accelerators produced by our framework can achieve 2.17 \times throughput on the average compared with a state-of-the-art GNN framework on CPU-GPU platform.

2 BACKGROUND AND RELATED WORK

2.1 GNN Models

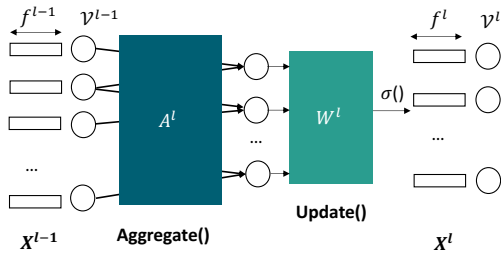


Figure 1: Computation abstraction of a GNN layer

Given an input graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{X})$, a GNN model is specified by:

- L : number of layers.
- \mathcal{V}^l : a set of target vertices to be inferred.
- f^l : hidden dimension in layer l ($1 \leq l \leq L$).
- A mechanism of how to construct:
 - \mathcal{V}^l : the set of vertices in layer l ($0 \leq l \leq L$). $|\mathcal{V}^l|$ denotes the number of vertices in layer l . Moreover, $\mathcal{V}^L = \mathcal{V}^T$.

- $\mathbf{A}^l \in \mathbb{R}^{|\mathcal{V}^{l-1}| \times |\mathcal{V}^l|}$: adjacency matrix for feature aggregation in layer l ($1 \leq l \leq L$). \mathbf{A}^l defines the inter-layer connectivity between \mathcal{V}^{l-1} and \mathcal{V}^l .

- $\mathbf{W}^l \in \mathbb{R}^{f^{l-1} \times f^l}$: weight matrix of layer l ($1 \leq l \leq L$) that is used in update function to perform linear transformation of vertex features.
- $\mathbf{X}^l \in \mathbb{R}^{|\mathcal{V}^l| \times f^l}$: feature matrix of layer l ($1 \leq l \leq L$)
- $\text{Aggregate}()$ function that is used by each vertex to aggregate information from its neighbors.
- $\text{Update}()$ function including an one-layer multi-layer perceptron (MLP) and an activation function $\sigma()$ that is used by each vertex to perform feature transformation.

Figure 1 depicts the computation abstraction of a GNN layer. The computations using a GNN model can also be expressed using the aggregate-update paradigm [26], as shown in Algorithm 1. $\mathbf{h}_v^l \in \mathbb{R}^{f^l}$ is the feature vector of $v \in \mathcal{B}^l$ in layer l , and $\mathbf{a}_v^l \in \mathbb{R}^{f^l}$ is the intermediate result of $v \in \mathcal{B}^l$. There are several widely used GNN

Algorithm 1 GNN Computation Abstraction

```

1: for  $l = 1 \dots L$  do
2:   for vertex  $v \in \mathcal{V}^l$  do
3:      $\mathbf{a}_v^l = \text{Aggregate}(\mathbf{h}_u^{l-1} : u \in \mathcal{N}(v) \text{ and } u \in \mathcal{V}^{l-1})$ 
4:      $\mathbf{h}_v^l = \text{Update}(\mathbf{a}_v^l, \mathbf{W}^l, \sigma())$ 
5:   end for
6: end for

```

models proposed in the literature:

GCN: GCN is proposed in [17]. Given the input graph $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{X})$, the GCN model is specified by:

- $\mathcal{V}^1 = \mathcal{V}^2 = \dots = \mathcal{V}^L = \mathcal{V}^T = \mathcal{V}$
- $\mathbf{A}^1 = \mathbf{A}^2 = \dots = \mathbf{A}^L = \mathbf{D}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-\frac{1}{2}}$, where \mathbf{A} and \mathbf{D} are the adjacency matrix and the Laplacian matrix of the input graph. \mathbf{I} is the identity matrix.
- L : number of layers; f^l : feature size in layer l ($1 \leq l \leq L$);

The $\text{Aggregate}()$ function and $\text{Update}()$ function of GCN are expressed as:

$$\begin{aligned} \mathbf{a}_v^l &= \text{Sum}\left(\frac{1}{\sqrt{D(v) \cdot D(u)}} \cdot \mathbf{h}_u^{l-1} : u \in \mathcal{N}(v) \cup \{v\}\right) \\ \mathbf{h}_v^l &= \text{ReLU}\left(\mathbf{a}_v^l \mathbf{W}^l + \mathbf{b}^l\right) \end{aligned} \quad (1)$$

where $\mathcal{N}(v)$ denotes the neighbor set of v in \mathcal{V}^{l-1} , $D(v)$ denotes the degree of vertex v , and \mathbf{b}^l denotes the bias of the update function.

GraphSAGE: GraphSAGE is proposed in [14] for inductive representation learning on graphs. Starting from a set of target vertex \mathcal{V}^T , GraphSAGE neighbor sampler recursively samples the neighbors to build $\mathcal{V}^1, \mathcal{V}^2, \dots, \mathcal{V}^L$. The adjacency matrix \mathbf{A}^l defines the edge connections between \mathcal{V}^{l-1} and \mathcal{V}^l , and each edge has weight equal to one. The $\text{Aggregate}()$ function and $\text{Update}()$ function of GraphSAGE are expressed as:

$$\begin{aligned} \mathbf{a}_v^l &= \mathbf{h}_v^{l-1} || \text{Mean}\left(\mathbf{h}_u^{l-1} : u \in \mathcal{N}(v) \cup \{v\}\right) \\ \mathbf{h}_v^l &= \text{ReLU}\left(\mathbf{a}_v^l \mathbf{W}^l + \mathbf{b}^l\right) \end{aligned} \quad (2)$$

Note: In the rest of the paper, we use GCN and GraphSAGE to refer to their GNN-layer operators $\text{Aggregate}()$, $\text{Update}()$.

2.2 GNN training

To train a GNN model, the GNN training process consists of five stages [2, 8, 14, 29]: sampling, forward propagation, loss calculation, back propagation and weight update. In the sampling stage, a set of vertices and adjacency matrices are sampled from $\{\mathcal{V}^l : 0 \leq l \leq L\}$ and $\{A^l : 1 \leq l \leq L\}$ to form a mini-batch. We use \mathcal{B}^l to denote the vertices sampled from \mathcal{V}^l in layer l . A_s^l denotes the sampled adjacency matrix, which describes inter-layer connections (edges) between \mathcal{B}^l and \mathcal{B}^{l-1} within the mini-batch. A mini-batch consists of target vertices \mathcal{B}^L , sampled vertices for each layer $\{\mathcal{B}^l : 0 \leq l \leq L-1\}$, and sampled adjacency matrices $\{A_s^l : 1 \leq l \leq L-1\}$. In the forward propagation stage, the mini-batch is processed layer by layer. The output embeddings in the last layer $\{h_i^L : v_i \in \mathcal{B}^L\}$ are compared with the ground truth for loss calculation. The calculated loss will be used as input for back propagation, which performs a similar computation as forward propagation but in the reverse direction. Finally, the gradients of W^l in each layer are derived and be used to update the weight. We show the steps of GNN training in Algorithm 2. In Algorithm 2, $N_s(v)$ denotes neighbors of v in \mathcal{B}^{l-1} that are specified in A_s^l . A GNN training algorithm is specified by a sampling algorithm (see Section 2.3) to construct the mini-batch that consists of $\{\mathcal{B}^l : 0 \leq l \leq L-1\}$ and $\{A_s^l : 1 \leq l \leq L-1\}$.

In HP-GNN, we exploit data parallelism within each mini-batch by aggregating and updating multiple vertices concurrently (shown in Line 5 of Algorithm 2). The computation order within the same mini-batch does not affect the final results. Thus, training in our parallel framework leads to the same result and accuracy as training in serial fashion.

Algorithm 2 Sampling-based GNN Training Algorithm

```

1: for each iteration do
2:   Sampling() // Derive mini-batches
3:   for  $l = 1 \dots L$  do // Forward Propagation
4:     for vertex  $v \in \mathcal{B}^l$  do
5:        $a_v^l = \text{Aggregate}(h_u^{l-1} : u \in N_s(v) \text{ and } u \in \mathcal{B}^{l-1})$ 
6:        $h_v^l = \text{Update}(a_v^l, W^l, \sigma())$ 
7:     end for
8:   end for
9:   CalculateLoss( $\{h_i^L : v_i \in \mathcal{B}^L\}$ )
10:  BackPropagation() // Derive gradient of  $W^l$ 
11:  WeightUpdate()
12: end for

```

2.3 Sampling Algorithm

An algorithm to sample a mini-batch is specified by:

- A method to sample the vertices \mathcal{B}^l ($0 \leq l \leq L$) from \mathcal{V}^l .
- A method to construct the adjacency matrix A_s^l ($1 \leq l \leq L$) from A^l .

By sampling the vertices \mathcal{B}^l ($0 \leq l \leq L$) and the adjacency matrix A^l ($1 \leq l \leq L$), we construct the mini-batch as the input for each training iteration. Various sampling methods [2, 10, 14, 27, 29] are proposed to form a mini-batch from input graphs. These sampling methods [20] falls into three categories: neighbor sampling, layer-wise sampling and subgraph sampling. We only introduce neighbor sampling and subgraph sampling, since layer-wise sampling [2] has the similar computation pattern with subgraph sampling [29].

Neighbor Sampling: Neighbor sampling [3, 14, 27] is a type of sampling strategy that recursively samples the neighbors from the target vertices. To perform neighbor sampling, users specify the size of target vertices $|\mathcal{V}^l|$ and the neighbor sample size NS^l for each vertex in layer l . The sampler first chooses a set of vertices as target vertices \mathcal{V}^l . Then, the sampler samples NS^L neighbors for each vertex in \mathcal{V}^l based on a specific probability distribution (e.g., uniform distribution [14]). After the first iteration of neighbor sampling, the set of 1-hop sampled neighbors \mathcal{B}^{L-1} is obtained, where $|\mathcal{B}^{L-1}| = |\mathcal{V}^L| \times NS^L$. Similarly, we obtain the 2-hop neighbors \mathcal{B}^{L-2} by sampling the neighbors of the 1-hop neighbors \mathcal{B}^{L-1} , where $|\mathcal{B}^{L-2}| = |\mathcal{V}^L| \times NS^L \times NS^{L-1}$. By performing neighbor sampling recursively, we obtain L -hop neighbors of the target vertices. After obtaining $\{\mathcal{B}^l : 0 \leq l \leq L-1\}$, A_s^l ($1 \leq l \leq L$) is constructed by:

$$A_s^l(u, v) = \begin{cases} A^l(u, v) & u \in \mathcal{B}^{l-1} \text{ and } v \in \mathcal{B}^l \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Subgraph Sampling: Subgraph sampling [8, 29] is a strategy to sample a subgraph from the input graph and perform GNN information propagation within the subgraph. In the subgraph sampling-based method, users specify sampling budget SB which denotes how many vertices to be sampled for the subgraph. Then, the sampler samples SB of vertices or edges based on specific probability, and induce a subgraph based on the sampled vertices or edges. For subgraph sampling, the sampled vertices are identical for each layer, i.e. $\mathcal{B}^0 = \mathcal{B}^1 = \dots = \mathcal{B}^L$. The construction of A_s^l ($1 \leq l \leq L$) is the same as neighbor sampling.

2.4 Related work

2.4.1 Software GNN Frameworks. Several software GNN frameworks [11, 23, 24, 33] have been proposed in the literature. PyTorch Geometric (PyG) [11] is one of the most commonly-used frameworks for GNN deployment that uses PyTorch [22] as the backend. PyG users can describe various GNN models using PyG's message passing API. Deep Graph Library (DGL) [23] adopts several parallelization strategies to achieve high performance and memory efficiency for GNN computations. Moreover, DGL offers several major frameworks [1, 4, 22] as the backend, allowing users to port their model across frameworks. Aligraph [33] supports training on heterogeneous attributed graph (HAG), i.e. graphs that contain different types of vertices and edges. These software frameworks share similar features: They are built on existing frameworks [1, 4, 22], and abstract away the detailed implementations by providing graph-oriented APIs so that users can describe GNN models easily.

2.4.2 Hardware GNN Acceleration. GraphACT [28] performs subgraph sampling based GNN training on CPU-FPGA heterogeneous platform. GraphACT exploits both task-level parallelism and data parallelism, and adopts a redundancy reduction technique to reduce the number of on-chip memory accesses. However, GraphACT optimizes the hardware design for inductive GNN models [14] using subgraph sampling, and does not support transductive model, such as GCN. As shown in [15], different sampling algorithms perform well in different applications, so there is no one-size-fits-all sampling algorithms. Rubik [6] decomposes GNN computations

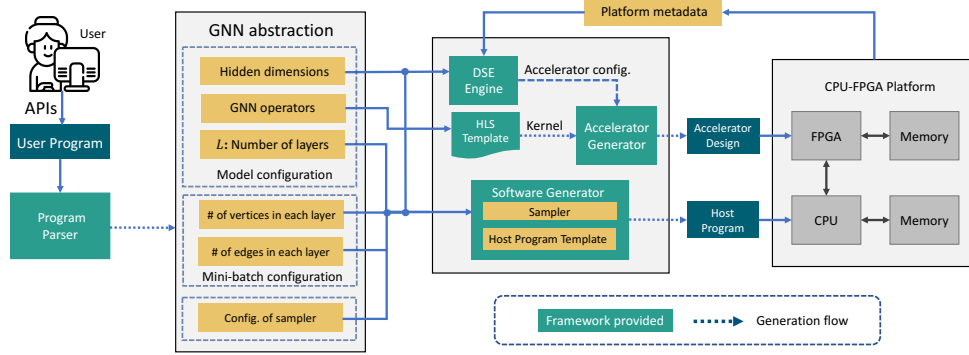


Figure 2: Framework overview

into graph-level and node-level computations, and proposes hierarchical task mapping strategies to exploit data reuse and parallelism in the two computation levels. However, Rubik is an ASIC accelerator that is hard to be optimized for various sampling algorithms. DeepBuring-GL [18] is a FPGA framework to accelerate GNN inference. DeepBuring-GL provides various templates to support different GNN computations and memory access patterns. DeepBuring-GL analyzes the GNN model with the input graph to identify performance bottleneck and choose appropriate hardware templates for kernel implementation to accelerate GNN inference. Though frameworks like DeepBuring-GL have been proposed, most of the hardware acceleration works still require significant hardware expertise to make use of them. Moreover, no framework has been proposed to support various mini-batch GNN training on CPU-FPGA platform, which motivates us to conduct this work. In this paper, we build a framework to accelerate GNN training on CPU-FPGA platform. Our framework provides infrastructures to support various GNN models and training algorithms.

We summarize the benefits of using CPU-FPGA platform: while CPU can support various sampling algorithms, FPGA platform allows customizable data path and memory access pattern that can be exploited to optimize the GNN training throughput.

3 FRAMEWORK

3.1 System Overview

Figure 3 depicts the mapping of graph data and various kernels onto the CPU-FPGA heterogeneous platform. Sampling is performed on the host CPU because CPU is flexible to support various sampling algorithms; GNN operations including feature aggregation and feature update are performed on the proposed FPGA accelerator. Based on this task assignment, the structural information of input graph (V, E) is stored in the host memory for the host CPU to perform sampling. After sampling is done, the structural information of the mini-batch is generated and transferred to the FPGA local memory. The vertex features X are stored in the FPGA local memory to be directly accessed by the FPGA accelerator, which can reduce the overhead of data movement. The state-of-the-art FPGA boards [13] have up to 260 GB memory; this can support medium size graph.

Regarding very large graphs, we store the vertex features in host memory and transfer the vertex features of the mini-batch to the FPGA accelerator after sampling.

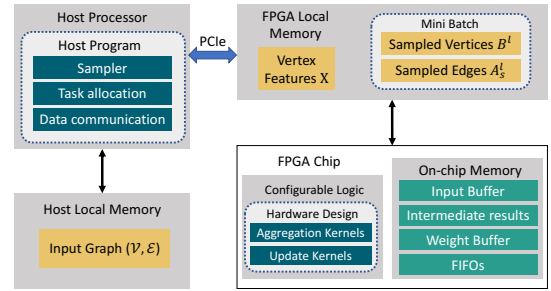


Figure 3: System overview

3.2 Framework Overview

Figure 2 demonstrates the framework overview. The generated design by the framework consists of two major components: (1) a host program that manages CPU-FPGA communication, kernel task scheduling and mini-batch sampling; (2) an accelerator design that runs on the FPGA. To generate the mini-batch GNN training implementation on CPU-FPGA platform, our framework takes the user program as the input and generates a high-level abstraction for mini-batch GNN training. In the input program, user specifies the following parameters:

- GNN parameters: number of layers L ; hidden dimension of each layer: f^l , $(0 \leq l \leq L)$. The hidden dimensions also define the dimension of weight matrix $W^l \in \mathbb{R}^{f^l \times f^{l+1}}$.
- Specify an off-the-shelf GNN model, or provide user-defined functions (UDFs) for `scatter()`, `gather()` and `update()` to build custom GNN computation layer.
- Sampling algorithm and its parameters. For example, a neighbor sampler for a 2-layer GNN model can be defined as `Sampler('NeighborSampler', L=2, budgets=[10, 25])` through our high-level API described in Section 3.3. We provide several off-the-shelf samplers for users to choose from.

The program parser extracts a GNN abstraction from user program, which serves as the intermediate representation for the software generator and hardware generator to generate the implementations on CPU-FPGA platform. The GNN abstraction consists of GNN model configuration (hidden dimensions f^l , GNN operators, number of layers L) and mini-batch configuration (number of vertices in each layer $|B^l|$, $(0 \leq l \leq L)$ and number of

| API Functions | Description |
|----------------------|--|
| Init() | Initialization the platform with FPGA bitstream |
| GNN_Parameters() | Number of layer L , feature length f^l , W^l and X |
| GNN_Computation() | The layer operators in GNN model Specify an off-the-shelf GNN model or "customized" |
| Scatter() | UDF, required if customized layer operator is specified |
| Gather() | UDF, required if customized layer operator is specified |
| Update() | UDF, required if customized layer operator is specified |
| GNN_Model() | Build GNN model using GNN parameters and computation |
| PlatformParameters() | FPGA Memory bandwidth, Number of DSPs, LUTs, etc. |
| LoadInputGraph() | Specify the input graph. |
| Sampler() | Sampling method with algorithmic parameters |
| DistributeData() | Distribute graph into host memory and FPGA local memory |
| GenerateDesign() | Generate hardware design and software design |
| Start_training() | Run GNN training |
| Save_model() | Save trained GNN model |
| PrepareEdges() | Prepare graph edge values that is used for training |

Table 1: Programming interfaces

edges in each layer $|\mathcal{E}^l|$, ($1 \leq l \leq L$). The mini-batch configuration is deduced from the sampling algorithm that implies number of vertices $|\mathcal{B}^l|$, ($0 \leq l \leq L$) in each layer and number of edges $|\mathcal{E}^l|$, ($1 \leq l \leq L$) in each layer.

DSE Engine: the DSE engine takes the GNN abstraction and the platform metadata as input and generates the accelerator configuration that optimizes the GNN training throughput (Section 5).

HLS Template/Hardware Template: In the framework, we provide optimized hardware templates written in high-level synthesis (HLS). The key computation operators of the templates (e.g. scatter(), gather()) are obtained from the GNN abstraction. We describe the details of the HLS template design in Section 4.

Accelerator generator: Given the generated accelerator configuration and hardware templates, the accelerator generator generates the hardware accelerators for the target FPGA board. The accelerator generator uses the available synthesis tools such as Xilinx Vitis as the backend.

Software generator: Given the input program, the software generator produces a runtime system that runs on the host processor. The runtime system performs the mini-batch sampling, CPU-FPGA communication management, and task scheduling.

3.3 High-Level APIs

Table 1 summarizes our provided high-level APIs for user to program the mini-batch training in Python. Listing 1 is an example for developing the GNN training using our proposed framework.

```

1 # design generation phase
2 Samp = Sampler('NeighborSampler', L=2, budgets=[10,25])
3
4 GNN_comp = GNN_Computation('SAGE')
5 GNN_para = GNN_Parameters(L=2, hidden=[256], feat)
6 Model = GNN_Model(GNN_comp, GNN_para)
7 # specify the resoruces of a single super logic region,
  using Xilinx-U250 as an example
8 ParaFPGA = PlatformParameters(board= , SLR = 4, DSP=3072,
  LUT=423000, URAM=320, BRAM= , BW=19.25)

```

```

9 bitstream, runtime = GenerateDesign(Sampler, Model,
  ParaFPGA) # generate hardware design and software
  design, return the pointers
10
11 # runtime phase
12 Graph = LoadInputGraph('Reddit', Path='')
13 Graph = PrepareEdges(Graph, Model)
14 Init(bitstream) # initialize the hardware platform
15 edgepointers, featurepointers = DistributeData(Graph) #
  distribute graph structural information and vertex
  features into host memory and fpga memory
16
17 start_training(runtime, edgepointers, featurepointers,
  epochs=10) # run GNN training
18 save_model() # save the trained model

```

Listing 1: An example user program for GNN training

In the design phase, user specifies the mini-batch sampler, GNN model and parameters of platform. The framework automatically generates the optimized accelerator design and software design. In the runtime phase, user starts the GNN training on the target CPU-FPGA platform. Using our high-level APIs, a GNN training program only requires a few dozen lines of code.

Use cases: Our framework can serve application developers, who utilize existing GNN models to build GNN applications. Our framework provides off-the-shelf GNN training implementations for some of the commonly-used GNN models (GCN [17], GraphSAGE [14], GIN [25]) which can be directly deployed on the CPU-FPGA platform. For machine learning (ML) researchers who develop novel GNN models, our framework allows them to customize their own GNN models. For both cases, our framework accelerates GNN training on CPU-FPGA platform without hardware expertise.

4 ACCELERATOR DESIGN

We design hardware templates based on the computation abstraction of the GNN layer described in Section 2.1. The hardware templates describe a general architecture of the GNN layer as in Figure 5 and Figure 6. Then, the accelerator generator takes user-defined functions as input and integrates them into the hardware templates to generate the accelerator design.

4.1 Data Layout and Internal Representation

Aggregating feature vector from neighbors incurs irregular memory access and large memory traffic. Figure 4 presents the data layout and internal representation used in our framework to reduce the memory traffic as well as the number of random memory accesses. The data layout is produced by the sampler through *renaming* and *sorting*.

Reducing Memory Traffic (RMT): During aggregation stage, the feature vector of the source vertex is sent to its destination for aggregation. For the first layer of aggregation, the input feature matrix X is stored in the memory. Thus, loading feature vectors incurs a large number of random memory accesses. Since the edges are represented in coordinate (COO) format sorted by source vertices in our framework, edges that share the same source vertex can reuse the feature vector that has been loaded, and thus reduce memory traffic. The total number of memory traffic can be reduced from $O(|\mathcal{E}_1|f^0)$ to $O(|\mathcal{B}_0|f^0)$, where $|\mathcal{E}_1|$ is usually larger than $|\mathcal{B}_0|$. Fig 4 depicts

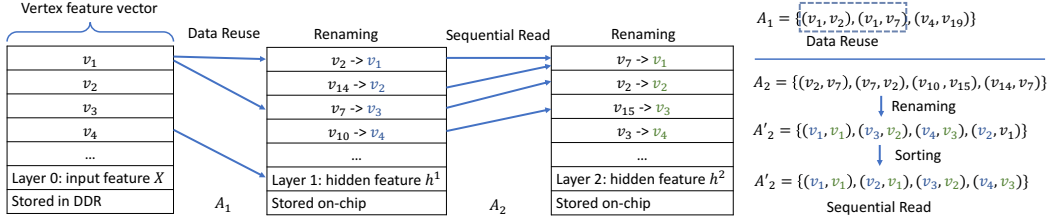


Figure 4: Data layout and Internal Representation

Algorithm 3 Aggregation by Scatter-Gather Paradigm

```

while not done do
  scatter phase:
  for each edge  $e$  do
    if  $\text{edge.src} = \text{feat.src}$  then
      Produce update  $u \leftarrow \text{Scatter\_func}(\text{feat.val}, \text{edge.val})$ 
    end if
  end for
  gather phase:
  for each update  $u$  do
    target vertex  $v = v_{u.dst}$ 
    Update vertex  $v \leftarrow \text{Gather\_func}(u.val, v.val)$ 
  end for
end while

```

an example: $\langle v_1, v_2 \rangle$ loads the feature vector of v_1 from memory, and the loaded feature vector of v_1 can be reused by (v_1, v_7) .

Reducing Random Access (RRA): Since the edges are sorted by the source index in the first layer, the destination vertex index of the edges are in a random order; thus, the hidden features are stored randomly as shown in the layer 1 and layer 2 of Figure 4. To reduce random access, our framework performs vertex renaming, which labels the vertices based on the order they are stored, this step also renames the vertices in each edge. Next, we sort the renamed edges by source vertices, and then accessing hidden features become sequential since the source vertex number follows the order it is stored.

4.2 Kernel Design

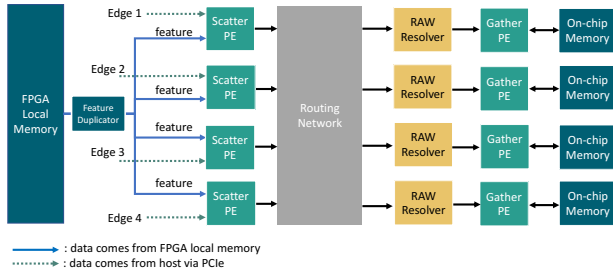


Figure 5: Architecture of aggregate kernel

Aggregate Kernel: The aggregate kernel adopts the scatter-gather paradigm [5] as illustrated in Algorithm 3. Figure 5 depicts the detailed architecture of the aggregate kernel. Multiple processing

elements (PEs) process multiple edges concurrently in each clock cycle. The vertex feature vectors are first streamed to a feature duplicator. The feature duplicator broadcasts the loaded feature vector to all the Scatter PEs. The feature vector is stored in the PEs' registers for data reuse. Then, Scatter PEs perform user-defined scatter() function, and stream the update u to its destination via the routing network. The routing network is implemented as a butterfly network [9]. After the Gather PEs receive the updates, Gather PEs perform user-defined gather function and obtain the intermediate results. The intermediate results are stored on-chip. Finally, when the aggregation is done, the results stored in the on-chip memory are written back to the FPGA local memory. Since the gather phase may incur reading and writing to the same destination vertices, read-after-write (RAW) data hazard may occur. The RAW Resolver addresses RAW data hazard by stalling.

Update Kernel: The update kernel is a systolic array based design that performs block matrix multiplication. The input buffer loads the aggregation results a^l (see Algorithm 2) from the FPGA local memory. a^l will then be streamed into the MAC array. Each MAC module is followed by an element-wise operator σ . Typically, weight of each layer W^l in GCN is small and frequently reused. Thus, W^l are stored on-chip in the Weight-Buffer. W^l will be broadcast to the multiply-accumulate (MAC) array during feature update. Finally, the result is stored into a result buffer before written back to the FPGA local memory.

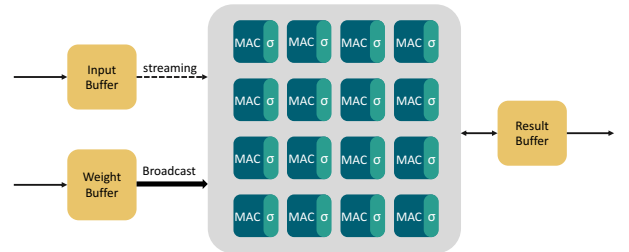


Figure 6: Architecture of update kernel

4.3 Parallel Computation Kernels

Many modern FPGAs consists of multiple dies and number of interconnection wires across the dies is limited. Therefore, we implement multiple copy of the kernels that is distributed into multiple dies as shown in Figure 7. Multiple dies and multiple DDR channels are connected through an all-to-all interconnection which

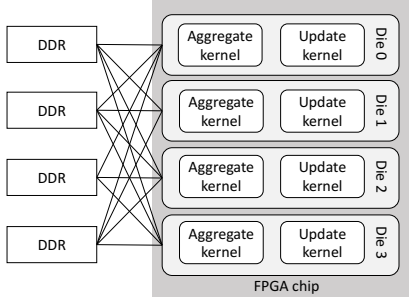


Figure 7: Architecture of the FPGA accelerator.

is generated by vendor tool, such as Xilinx Vitis. The input feature matrix X is equally partitioned into DDR channels. To utilize the multiple computation kernels for a single mini-batch, we perform task partitioning for the mini-batch training. In the forward propagation phase of layer 1, to infer the vertices in $\mathcal{B}^1 = \{v_1, v_2, v_3, \dots, v_{|\mathcal{B}^1|}\}$. The workload for inferring \mathcal{B}^1 are equally partitioned into multiple kernels. Suppose there is 4 aggregate kernels and 4 corresponding update kernels. Aggregate kernel 1 aggregates $\{v_1, v_2, \dots, v_{\lfloor \frac{|\mathcal{B}^1|}{4} \rfloor}\}$. Update kernel 1 updates $\{v_1, v_2, \dots, v_{\lfloor \frac{|\mathcal{B}^1|}{4} \rfloor}\}$ and write the results back to DDR. Similarly, aggregate kernel 2 aggregates $\{v_{\lfloor \frac{|\mathcal{B}^1|}{4} \rfloor + 1}, v_{\lfloor \frac{|\mathcal{B}^1|}{4} \rfloor + 2}, \dots, v_{\lfloor \frac{|\mathcal{B}^1|}{4} \rfloor}\}$, and so on and so forth. The same partitioning scheme is applied to each layer.

5 DESIGN SPACE EXPLORATION

Our framework provides a design space exploration (DSE) engine for optimizing the GNN training throughput, given the configuration of mini-batch ($\{|\mathcal{B}^l| : 0 \leq l \leq L\}$, $\{|\mathcal{E}^l| : 1 \leq l \leq L\}$), GNN hidden dimensions $\{f^l : 0 \leq l \leq L\}$, memory bandwidth α , and hardware resources per die (DSPs, BRAMs, URAMs). To drive the optimization, we develop a performance model (Section 5.1) that models the training throughput on the CPU-FPGA platform, and the resource utilization model (Section 5.2) that is used to specify the resource constraints. Then, our DSE engine (Algorithm 4) performs parameter sweep in the design space to identify the hardware design parameters that optimizes the training throughput.

5.1 Performance Model

We define the throughput of mini-batch GNN training as Number of Vertices Traversed Per Second (NVTPS):

$$\text{Throughput} = \frac{\sum_{l=0}^L |\mathcal{B}^l|}{t_{\text{execution}}} \quad (4)$$

The numerator indicates the total amount of vertices traversed in one mini-batch, and the denominator $t_{\text{execution}}$ is the average execution time of one training iteration (see Algorithm 2). The modeling of the average execution time is based on the our task scheduling on the CPU-FPGA platform.

We overlap sampling stage of the next batch with the execution of the current batch, so average execution time $t_{\text{execution}}$ is estimated

as:

$$t_{\text{execution}} = \max(t_{\text{sampling}}, t_{\text{GNN}}) \quad (5)$$

$$t_{\text{GNN}} = t_{\text{FP}} + t_{\text{LC}} + t_{\text{BP}} + t_{\text{WU}}$$

where t_{GNN} consists of the execution time of forward propagation t_{FP} , loss calculation t_{LC} , back propagation t_{BP} and weight update t_{WU} .

Modeling t_{GNN} : Loss calculation and weight update are executed on the host processor, which have optimized implementation in the software library. Forward propagation and backward propagation are executed on the FPGA platform, and their execution time depends on the hardware parameters and the mini-batch configuration ($\{|\mathcal{B}^l| : 0 \leq l \leq L\}$, $\{|\mathcal{E}^l| : 1 \leq l \leq L\}$). We drive the approximate execution time of two propagation stages as:

$$t_{\text{FP}} = \sum_{l=1}^L \max(t_{\text{aggregate}}^l, t_{\text{update}}^l) \quad (6)$$

$$t_{\text{BP}} = t_{\text{update}}^1 + \sum_{l=2}^L \max(t_{\text{aggregate}}^l, t_{\text{update}}^l)$$

The total propagation time t_{FP} or t_{BP} is the sum of the execution time of each layer, and the execution time of each layer is decided by the task that takes longer to complete since aggregation stage and update stage are pipelined.

The aggregation stage consists of two tasks: (1) loading vertex features or gradients, and (2) computation. Since the two tasks are pipelined, $t_{\text{aggregate}}$ can be modeled as:

$$t_{\text{aggregate}}^l = \max(t_{\text{load}}^l, t_{\text{compute}}^l) \quad (7)$$

$$t_{\text{load}}^l = \frac{|\mathcal{B}^{l-1}| \times f^l \times S_{\text{feat}}}{BW \times \alpha} \quad t_{\text{compute}}^l = \frac{|\mathcal{E}^l| \times f^l}{n \times 16 \times \text{freq}} \quad (8)$$

We model the vertex feature loading time t_{load}^l as $\frac{\text{data transferred}}{\text{effective bandwidth}}$. $|\mathcal{B}^l|$ indicates the number of vertices in each layer, f^l is the feature length, and S_{feat} is the data size of each feature. α is the effective bandwidth ratio. For the feature loading of the first layer of neighbor sampling method, α is estimated based on the memory burst transaction length S_{feat} [21] (for DDR4) because it incurs random memory access; for the rest of the layers, α is near to 1 [21] since the memory accesses are sequential from DDR memory or on-chip memory (Section 4.1). The value of α is obtained from the prior work [21] which performs a profiling for the characteristics of the FPGA DDR memory. We model the compute time as $(\# \text{ of operations}) / (\# \text{ of PEs} \times \text{kernel frequency})$. n denotes the there are n Scatter PEs and n Gather PEs instantiated in the aggregation kernel. $|\mathcal{E}^l|$ is the number of edges (i.e. non-zeros) in each layer. The size of $|\mathcal{E}^l|$ depends on the sampling method. We show the modeling of \mathcal{E}^l for various sampling methods in Table 2. For neighbor sampling, the number of edges $|\mathcal{E}^l|$ in each layer is decided by the size of target vertices $|\mathcal{V}^l|$ and sample size NS^l . For layer-wise and sub-graph sampling, we formulate the number of edges in layer $|\mathcal{E}^l|$ as $|\mathcal{B}^l| \times |\mathcal{B}^{l-1}| \times \kappa(|\mathcal{B}^l|)$ where $|\mathcal{B}^l| \times |\mathcal{B}^{l-1}|$ corresponds to the case all the sampled vertices in layer l and $l-1$ are connected, and $\kappa(|\mathcal{B}^l|)$ is a pre-trained function that estimates the graph sparsity based on sample size $|\mathcal{B}^l|$.

Table 2: # of vertices and # of edges in each layer of various sampling methods

| Method | # of Vertices $ \mathcal{B}^l $ | # of edges $ \mathcal{E}^l $ |
|-----------------------|---|---|
| Neighbor | $ \mathcal{V}^l \times \prod_{i=l+1}^L NS^i$ | $ \mathcal{V}^l \times \prod_{i=l}^L NS^i$ |
| Layer-wise | S^l | $S^l \times S^{l-1} \times \kappa(S^l)$ |
| Subgraph [‡] | SB | $SB \times \kappa(SB)$ |

[‡] Uses node sampler in [29] as an example

The feature update can be modelled as

$$t_{\text{update}} = \frac{|\mathcal{B}^l| \times f^l \times f^{l+1}}{m \times \text{freq}} \quad (9)$$

Similar to t_{compute} , we model the t_{update} as (# of operations)/(# of PEs \times kernel frequency). The numerator is the complexity of matrix multiplication, and m denotes how many parallel MACs are instantiated in the update kernel.

Modeling t_{sampling} : the mini-batch sampling is performed on the host processor, which can potentially be a performance bottleneck. We exploit multi-threading to sample multiple mini-batches concurrently. In the design phase, we estimate t_{sampling} under various number of threads and determine the minimum number of threads that satisfies $t_{\text{sampling}} < t_{\text{GNN}}$.

5.2 Resource Utilization Model

We set the hardware constraints to form the solution space for our DSE Engine. Among the various hardware resources on the FPGA platform, DSPs and LUTs are used the most as we increase the parallelism of the hardware modules. Thus, we model the usage of LUTs and DSPs as our constraints:

$$\lambda_1 \times m + \lambda_2 \times n \leq N_{\text{DSP}} \quad (10)$$

$$\rho_1 \times m + \rho_2 \times n + \rho_3 \times n \log(n) \leq N_{\text{LUT}} \quad (11)$$

The coefficients λ_i ($1 \leq i \leq 2$) and ρ_i ($1 \leq i \leq 3$) are constants that indicate the resource consumption for each PE. In the case of DSPs, the utilization grows linearly as we instantiate more PEs; In the case of LUTs, an additional $n \log(n)$ (see Section 4.2 for the definition of n) term is introduced. The $n \log(n)$ LUT overhead models the routing network in the aggregation kernel shown in Figure 5. N_{DSP} and N_{LUT} denote the available DSPs and LUTs on the FPGA platform.

5.3 DSE Engine

Many modern FPGAs consists of multiple dies [7], and the available resources may vary across dies. Thus, we perform DSE for each die to explore the optimal hardware configuration. We assume that each die is connected to one DDR channel (e.g. Xilinx Alveo U250) for simplicity in Algorithm 4; The DSE engine first constructs the search space by deriving the maximum value of n and m separately based on Equations (10) and (11). Then, the engine performs an exhaustive search through all the possible configurations. For each configuration, the engine evaluates its throughput using Equation 4, and chooses the optimal design.

Algorithm 4 Design Space Exploration Engine

```

1: for each die do
2:   Construct_Search_Space( ) // Derive  $n_{\text{max}}, m_{\text{max}}$ 
3:   for  $n = 1 \dots n_{\text{max}}$  do // Exhaustive search
4:     for  $m = 1 \dots m_{\text{max}}$  do
5:       #Check resource availability based on Eq. (10), (11)
6:       Valid  $\leftarrow$  Check_resource_availability( $n, m$ )
7:       if Valid And Throughput( $n, m$ )  $>$   $\text{max\_val}$  then
8:          $\text{max\_val} \leftarrow$  Throughput( $n, m$ )
9:         Save_configuration( $n, m$ )
10:      end if
11:    end for
12:  end for
13: end for

```

6 EXPERIMENTS

6.1 Experimental Setup

We use our framework to generate GNN training implementations on a CPU-FPGA heterogeneous platform, and compare the training throughput with CPU-only platform and CPU-GPU platform. We list the information of each platform in Table 3. The CPU-only and CPU-GPU baseline are implemented using PyTorch-Geometric[11]¹ and GraphSAINT[29]².

Table 3: Specifications of the platforms

| Platforms | CPU | GPU | FPGA |
|------------------|-----------------|----------------|-------------------|
| | AMD Ryzen 3990x | Nvidia A100 | Xilinx Alveo U250 |
| Technology | TSMC 7 nm | TSMC 7 nm | TSMC 16 nm |
| Frequency | 2.90 GHz | 1410 MHz | 300 MHz |
| Peak Performance | 3.7 TFLOPS | 19.5 TFLOPS | 0.6 TFLOPS |
| On-chip Memory | 256 MB L3 cache | 40 MB L2 Cache | 54 MB |
| Memory Bandwidth | 107 GB/s | 1555 GB/s | 77 GB/s |

Samplers, Models and Datasets: We generate mini-batch for GNN training using two sampling algorithms: (1) GraphSAGE neighbor sampler [14] for neighbor sampling (NS) and (2) GraphSAINT node sampler [29] for subgraph sampling (SS). For GraphSAGE neighbor sampler, we set the size of target vertices $|\mathcal{V}^l|$ as 1024, neighbor sampling size NS as 25 and 10 for 1-hop neighbors and 2-hop neighbors; for GraphSAINT-node sampler, we set the sampling budget SB as 2750. We measure the GNN training throughput of two-layer GCN model and two-layer GraphSAGE model on four medium-scale graph datasets (Flickr [29], Reddit [14], Yelp [29] and AmazonProducts [29]) that fit in the FPGA local DDR memory. Details of the datasets and the GNN-layer dimensions are shown in Table 4.

6.2 Framework Implementation

We implement the program parser, DSE engine, software and hardware generator in our framework using Python, and the accelerator templates are implemented using Xilinx HLS. The host program template is programmed in OpenCL. Users interface with our framework using our APIs programmed in Python. To serve application

¹https://github.com/pyg-team/pytorch_geometric/blob/master/examples/reddit.py

²<https://github.com/GraphSAINT/GraphSAINT>

Table 4: Statistics of the Datasets and GNN-layer dimensions

| Dataset | #Nodes | #Edges | f_0 | f_1 | f_2 |
|---------------------|---------|-----------|-------|-------|-------|
| Flickr (FL) | 89250 | 899756 | 500 | 256 | 7 |
| Reddit (RD) | 232965 | 11606919 | 602 | 256 | 41 |
| Yelp (YP) | 716847 | 6977410 | 300 | 256 | 100 |
| AmazonProducts (AP) | 1598960 | 132169734 | 200 | 256 | 107 |

developers, our framework includes several commonly used GNN models that can be used off-the-shelf. For ML researchers, our APIs allow users to define new models. In Listing 2, we provide some examples of user inputs to specify platforms, mini-batch samplers and GNN models via our APIs. Based on the given inputs, our framework generates the host program and synthesizable accelerator design. For example, based on the GNN model specified by the user, user decides the parameters in the host program template, and what aggregation function should be filled in the HLS template to generate the accelerator design. Based on the platform parameters, our DSE engine fills in the hardware configurations such as the unroll factor in our HLS template. In Listing 3, we provide an example that shows part of generated host program and synthesizable accelerator design.

Table 5: Resource utilization and Parallelism

| Resources | NS-GCN | NS-GraphSAGE | SS-GCN | SS-GraphSAGE |
|-----------|---------|--------------|---------|--------------|
| LUTs | 50% | 54% | 44% | 76% |
| DSPs | 70% | 54% | 70% | 82% |
| URAM | 34% | 34% | 14% | 20% |
| BRAM | 28% | 28% | 30% | 34% |
| (m,n) | (256,4) | (256,4) | (256,4) | (256,8) |

In Table 5 we show the resource utilization of the implementations generated by our design. The number n which denotes the number of Scatter PEs and Gather PEs are restricted to power of 2, and number of MACs m is restricted to square of power of 2 due to the design of our accelerator.

```

1 '''Example of platform specification'''
2 Parameter = PlatformParameters(board='xilinx-U250')
3 Parameter = PlatformParameters(board= , SLR = 4, DSP
   =3072, LUT=423000, URAM=320, BRAM= , BW=19.25)
   #specify parameters for new boards
4
5 '''Example of different sampler'''
6 Samp = Sampler( 'NeighborSampler', L=2, budgets=[10,25])
7 Samp = Sampler( 'SubgraphSampler', L=2, budgets=[1500])
8
9 '''Example of different GNN models'''
10 GNN_para = GNN_Parameters(L=2, hidden=[256], v_feat)
11 GNN_comp = GNN_Computation('SAGE')
12 Model = GNN_Model(GNN_comp, GNN_para)
13
14 GNN_para = GNN_Parameters(L=2, hidden=[256], v_feat)
15 GNN_comp = GNN_Computation('customize')
16 Model = GNN_Model(GNN_comp, GNN_para)
17
18 '''User defined functions for custom layer operator'''
19 Scatter(edge, feat, msg){msg.dst = edge.dst;
20 msg.val = edge.val*feat[edge.src];}
21

```

```

22 Gather(msg, v_ft){v_ft[msg.dst] += msg.val;}
23
24 Update('ReLU')

```

Listing 2: Examples of user inputs

```

1 //part of the generated host program
2 cl::Device devices = xcl::get_xil_devices();
3 for(int iter = 0; iter < Layer; iter++) {
4 // variable "Layer" is specified by user
5     buffer_out = cl::Buffer(CL_MEM_WRITE_ONLY);
6     aggregate_krnls.setArg(0, buffer_out);
7     ...//set arguments for kernel
8     q.enqueueTask(aggregate_krnls);
9     q.finish();
10    q.enqueueMigrateMemObjects(buffer_out);
11    //copy result back to host
12 //part of the generated accelerator design
13 read_from_stream(input, tmp_update);
14 if(tmpupdate.valid == 1){
15     index_type dst = tmp_update.dst - dst_offset;
16     reg_update = result_buffer[dst];
17
18     for (int j = 0; j < 16; j++){
19         #pragma HLS unroll = 8 //decide by DSE
20         regupdate.data[j] = regupdate.data[j] + tmpupdate.
           value.data[j];
21     }//user-specified aggregate function
22
23     resultbuffer[dst] = regupdate;
24 }

```

Listing 3: Example of generated code

6.3 Impact of Optimizations

We evaluate the two optimizations of our data layout and internal representation described in Section 4.1 on a two-layer neighbor sampling GCN. The two optimizations are: (1) reducing memory traffic (RMT) by reusing loaded vertex features in different edges that share the same source vertex; (2) reducing random access (RRA) by vertex renaming followed by edge sorting. We first measure the throughput of the baseline implementation with no optimizations, and then incrementally apply the two optimizations. As shown in Table 6, both optimizations increase the GNN training throughput and can deliver up to 57% improvement in total.

Table 6: Throughput improvement due to the optimizations

| Throughput (NVTPS) | FL | RD | YP | AP |
|--------------------|------------|------------|------------|------------|
| Baseline | 10.45 M | 12.98 M | 19.71 M | 23.17 M |
| RMT | 11.98 M | 16.48 M | 22.39 M | 27.22 M |
| RMT+RRA | 16.38 M | 18.50 M | 24.60 M | 29.27 M |
| Improvement | 57% | 43% | 25% | 26% |

6.4 Comparison with State-of-the-art

For evaluation, we use the throughput defined in Section 5 as metric, i.e. number of vertices traversed per second (NVTPS). To measure the throughput, we count the total number of vertices in each mini-batch, and measure the average execution time of one training iteration. Table 7 shows the throughput comparison of GNN training across the three platforms. All three implementations use single precision floating point as data type.

Table 7: Cross Platform Comparison (Throughput)

| | Data | CPU | CPU-GPU | CPU-FPGA |
|----------------|------|--------------------|-----------------------|------------------------|
| NS-GCN | FL | 265.5K (1×) | 2.69M (10.1×) | 16.38M (61.7×) |
| | RD | 85.65K (1×) | 7.15M (83.5×) | 18.50M (216×) |
| | YP | 275.6K (1×) | 9.36M (34.0×) | 24.61M (89.2×) |
| | AM | 480.6K (1×) | 13.0M (29.0×) | 29.26M (60.8×) |
| NS-SAGE | FL | 225.2K (1×) | 2.74M (12.2×) | 11.84M (52.6×) |
| | RD | 78.50K (1×) | 6.90M (88.0×) | 13.10M (166×) |
| | YP | 266.0K (1×) | 9.19M (34.5×) | 18.12M (68.1×) |
| | AM | 479.3K (1×) | 13.57M (28.3×) | 21.15M (44.1×) |
| SS-GCN | FL | 215.2K (1×) | 768.3K (3.59×) | 2.81M (13.0×) |
| | RD | 118.9K (1×) | 536.4K (4.51×) | 2.56M (21.5×) |
| | YP | 159.1K (1×) | 751.0K (4.71×) | 3.08M (19.4×) |
| | AM | 25.55K (1×) | OoM | 1.47M (57.5×) |
| SS-SAGE | FL | 179.9K (1×) | 626.7K (3.48×) | 2.71M (15.1×) |
| | RD | 94.72K (1×) | 505.2K (5.33×) | 2.43M (25.6×) |
| | YP | 126.7K (1×) | 709.7K (5.60×) | 2.78M (22.0×) |
| | AM | 17.40K (1×) | OoM | 1.45M (83.3×) |
| Average | | 193.4K (1×) | 4.96M (25.66×) | 10.77M (55.67×) |

Comparing with the CPU-only baseline, the CPU-GPU platform achieves 25.66× throughput on the average. This is because CPU-GPU platform provides massive data parallelism with 5.27× peak performance and 14.5× memory bandwidth compared with the CPU-only platform (Table 3). Comparing with the CPU-only baseline and CPU-GPU baseline, the CPU-FPGA implementation generated by our framework achieves 55.67× and 2.17× throughput on the average respectively. Though CPU-GPU platform has higher memory bandwidth and peak performance than CPU-FPGA platform, the throughput is limited by the memory access overhead during aggregation stage. While our accelerator can access the on-chip memory in one cycle (3.3 ns), CPU and GPU need to access the data in multi-level caches. Taking AMD Ryzen 3990 as an example, the L2 cache latency is 5 to 12 ns, and the L3 cache latency is around 32 ns. Moreover, as shown in Table 6, our data layout and internal representation also improves the training throughput by reducing the memory traffic and reducing random memory accesses.

We compare our results with two state-of-the-art GNN training implementations: GraphACT [28] and Rubik [6]. As shown in Table 8, our framework achieves up to 4.45× and 3.4× throughput respectively. Compared with GraphACT, the achieved speedup is due to (1) the vertex features are fetched directly from the FPGA local memory, (2) the proposed aggregate kernel has higher computation parallelism compared with Feature Aggregation Module in GraphACT. Compared with ASIC design Rubik, the obtained speedup is due to (1) larger on-chip memory of FPGA that can fully store the intermediate results under the setting of the experiments, (2) our proposed data layout optimizations that reduce the external memory traffic and random memory accesses.

7 DISCUSSION

We discuss the novelty of this work compared with previous work GraphACT [28] and the applicability of proposed optimizations to other platforms (e.g., CPU, GPU).

Table 8: Comparison with State-of-the-art

| | | GraphACT [28]‡ | Rubik [6] | This work |
|-------------------------|--------------|----------------|----------------|---------------|
| Platform | Device | Alveo U250 | ASIC | Alveo U250 |
| | Peak Perf. | 0.6 TFLOPS | 1 TFLOPS | 0.6 TFLOPS |
| | Bandwidth | 77 GB/s | 432 GB/s | 77 GB/s |
| | On-chip Mem. | 54 MB | 2 MB | 54 MB |
| SS-SAGE (Throughput) | RD | 546.8K (1×) | 717.0K (1.31×) | 2.43M (4.45×) |
| | YP | 769.8K (1×) | N/A | 2.78M (3.61×) |

‡ Scaled from U200 to U250 using the number of DSPs.

Comparison with GraphACT. In GraphACT [28], the redundancy reduction requires that all the edges have uniform weight value. Therefore, it can not support GCN [17]. In contrast, the proposed optimizations such as RMT and RRA do not have requirements on the weight, therefore, can support broader range of GNN models. Moreover, the Feature Aggregation Module in GraphACT has limited computation parallelism in feature-level that limits its performance for neighbor-sampling-based GNN training. In comparison, the proposed aggregate kernel adopts the scatter-gather paradigm with routing network, which can enable massive computation parallelism within feature aggregation.

Optimizations on CPU/GPU platforms. In this work, we proposed a highly optimized aggregate kernel that adopts the scatter-gather paradigm to accelerate feature aggregation. While the scatter phase is optimized by our proposed data layout optimization (Section 4.1), the performance of gather phase depends on the routing network of aggregate kernel (Section 4.2) that efficiently routes the intermediate result from Scatter PEs to Gather PEs. On CPU/GPU platforms, the data layout optimization can potentially be adopted to optimize the scatter phase. However, the gather phase is hard to be optimized on CPU/GPU since the data communication among the computation cores is through a complex cache hierarchy. Since the scatter phase and gather phase need be optimized simultaneously, the proposed optimizations may lead to limited performance improvement on CPU/GPU platforms.

8 CONCLUSION

In this paper, we proposed HP-GNN, a general framework to generate high-throughput GNN training implementation on a given CPU-FPGA heterogeneous platform. Based on the high-level abstraction of GNN computation, we designed a host program template and hardware templates to support various GNN models. Our proposed data layout and internal representation improve the throughput of GNN training. The implementations generated by HP-GNN achieve 55.67× and 2.17× throughput compared with state-of-the-art CPU-only and CPU-GPU platforms. Compared with state-of-art accelerators, our framework achieves up to 4.45× throughput. In the future, we plan to extend our framework to multi-FPGA platforms by exploiting model parallelism.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their comments which greatly improved the presentation. This work has been supported by the U.S. National Science Foundation under grant numbers OAC-1911229, CNS-2009057 and CCF-1919289. Equipment and support by Xilinx are greatly appreciated.

REFERENCES

- [1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *USENIX OSDI 16*.
- [2] Jie Chen, Tengfei Ma, and Cao Xiao. 2018. Fastgcn: fast learning with graph convolutional networks via importance sampling. *arXiv preprint arXiv:1801.10247* (2018).
- [3] Jianfei Chen, Jun Zhu, and Le Song. 2017. Stochastic training of graph convolutional networks with variance reduction. *arXiv preprint arXiv:1710.10568* (2017).
- [4] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. 2015. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. *CoRR* (2015).
- [5] Xinyu Chen, Hongshi Tan, Yao Chen, Bingsheng He, Weng-Fai Wong, and Deming Chen. 2021. ThunderGP: HLS-Based Graph Processing Framework on FPGAs. In *Proceedings of the 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA '21)*. Association for Computing Machinery, New York, NY, USA.
- [6] Xiaobing Chen, Yuke Wang, Xinfeng Xie, Xing Hu, Abanti Basak, Ling Liang, Mingyu Yan, Lei Deng, Yufei Ding, Zidong Du, and Yuan Xie. 2021. Rubik: A Hierarchical Architecture for Efficient Graph Neural Network Training. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2021).
- [7] Yao Chen, Jiong He, Xiaofan Zhang, Cong Hao, and Deming Chen. 2019. Cloud-DNN: An open framework for mapping DNN models to cloud FPGAs. In *Proceedings of the 2019 ACM/SIGDA international symposium on field-programmable gate arrays*.
- [8] Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. 2019. Cluster-gcn: An efficient algorithm for training deep and large graph convolutional networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
- [9] Young-kyu Choi, Yuze Chi, Weikang Qiao, Nikola Samardzic, and Jason Cong. 2021. Hbm connect: High-performance hls interconnect for fpga hbm. In *The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. 116–126.
- [10] Hanjun Dai, Zornitsa Kozareva, Bo Dai, Alex Smola, and Le Song. 2018. Learning steady-states of iterative algorithms over graphs. In *International conference on machine learning*. PMLR, 1106–1114.
- [11] Matthias Fey and Jan E. Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.
- [12] Tong Geng, Ang Li, Runbin Shi, Chunshu Wu, Tianqi Wang, Yanfei Li, Pouya Haghi, Antonino Tumeo, Shuai Che, Steve Reinhardt, et al. 2020. AWB-GCN: A graph convolutional network accelerator with runtime workload rebalancing. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE.
- [13] Gidel. 2021. Proc10S High Performance Scalable Compute Accelerators. [online] Available at: <https://gidel.com/acceleration-platforms/proc10s-high-performance-scalable-compute-accelerators> [Accessed 11 September 2021].
- [14] William L Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.
- [15] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2021. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *arXiv:2005.00687 [cs.LG]*
- [16] Weiwei Jiang and Jiayun Luo. 2021. Graph neural network for traffic forecasting: A survey. *arXiv preprint arXiv:2101.11174* (2021).
- [17] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017*.
- [18] Shengwen Liang, Cheng Liu, Ying Wang, Huawei Li, and Xiaowei Li. 2020. DeepBurning-GL: an Automated Framework for Generating Graph Neural Network Accelerators. In *2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*.
- [19] Yi Chien Lin, Bingyi Zhang, and Viktor Prasanna. 2021. GCN Inference Acceleration using High-Level Synthesis. In *2021 IEEE High Performance Extreme Computing Conference (HPEC)*.
- [20] Xin Liu, Mingyu Yan, Lei Deng, Guoqi Li, Xiaochun Ye, and Dongrui Fan. 2021. Sampling methods for efficient training of graph convolutional networks: A survey. *CoRR* (2021).
- [21] Alec Lu, Zhenman Fang, Weihua Liu, and Lesley Shannon. 2021. Demystifying the memory system of modern datacenter FPGAs for software programmers through microbenchmarking. In *The 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*.
- [22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*.
- [23] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2019. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. *arXiv preprint arXiv:1909.01315* (2019).
- [24] Yuke Wang, Boyuan Feng, Gushu Li, Shuangchen Li, Lei Deng, Yuan Xie, and Yufei Ding. 2021. GNNAdvisor: An Adaptive and Efficient Runtime System for GNN Acceleration on GPUs. In *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*. 515–531.
- [25] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks? (2019).
- [26] Mingyu Yan, Lei Deng, Xing Hu, Ling Liang, Yujing Feng, Xiaochun Ye, Zhimin Zhang, Dongrui Fan, and Yuan Xie. 2020. HyGCN: A GCN Accelerator with Hybrid Architecture. In *2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*.
- [27] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
- [28] Hanqing Zeng and Viktor Prasanna. 2020. GraphACT: Accelerating GCN Training on CPU-FPGA Heterogeneous Platforms. In *Proceedings of the 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA '20)*. Association for Computing Machinery, New York, NY, USA.
- [29] Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. 2020. GraphSAINT: Graph Sampling Based Inductive Learning Method. In *International Conference on Learning Representations*.
- [30] Bingyi Zhang, Rajgopal Kannan, and Viktor Prasanna. 2021. BoostGCN: A Framework for Optimizing GCN Inference on FPGA. In *2021 IEEE 29th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE.
- [31] Bingyi Zhang, Hanqing Zeng, and Viktor Prasanna. 2020. Hardware acceleration of large scale GCN inference. In *2020 IEEE 31st International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. IEEE.
- [32] Shijie Zhou, Rajgopal Kannan, Viktor K Prasanna, Guna Seetharaman, and Qing Wu. 2019. Hitgraph: High-throughput graph processing framework on fpga. *IEEE Transactions on Parallel and Distributed Systems* 30, 10 (2019), 2249–2264.
- [33] Rong Zhu, Kun Zhao, Hongxia Yang, Wei Lin, Chang Zhou, Baole Ai, Yong Li, and Jingren Zhou. 2019. AliGraph: a comprehensive graph neural network platform. *Proceedings of the VLDB Endowment* (2019).