Reforge: Low-Latency Distributed GNN Serving with Selective Embedding Recomputation

Geon-Woo Kim, Donghyun Kim, Jeongyoon Moon, Henry Liu, Tarannum Khan,
Anand Iyer†, Daehyeok Kim, Aditya Akella
The University of Texas at Austin, †*Georgia Institute of Technology*
{gwkim, donghyun, jeongyoonm, henry.liu, tarannum.khan, daehyeok}@utexas.edu,
anand.iyer@gatech.edu, akella@cs.utexas.edu

***Abstract*—Graph Neural Networks (GNNs) have been widely adopted for their ability to compute expressive node representations in graph datasets. However, serving GNNs on large graphs is challenging due to the high communication, computation, and memory overheads of constructing and executing computation graphs, which represent information flow across large neighborhoods. Existing approximation techniques in training can mitigate the overheads but, in serving, still lead to high latency and/or accuracy loss. To this end, we propose Reforge, a system that enables low-latency GNN serving for large graphs with minimal accuracy loss through two key ideas. First, Reforge employs *selective recomputation of precomputed embeddings*, which allows for reusing precomputed computation subgraphs while selectively recomputing a small fraction to minimize accuracy loss. Second, we develop *computation graph parallelism*, which reduces communication overhead by parallelizing the creation and execution of computation graphs across machines. Our evaluation with large graph datasets and GNN models shows that Reforge significantly outperforms state-of-the-art techniques.**

***Index Terms*—Graph Neural Networks, Low-Latency Serving, Efficient Inference, Parallelization Technique**

I. Introduction

Graph Neural Networks (GNNs) have gained widespread attention due to their ability to provide breakthrough results in diverse domains [1]–[6]; they have been instrumental in discovering life-saving drugs [7], [8], empowered recommendation systems [9], [10], and helped in traffic planning [11]–[13]. Recent studies have enabled reasoning on graph-structured data [14]–[18] by enhancing large language models (LLMs) [19]–[22] with GNNs to effectively capture intricate structural information.

The GNN lifecycle consists of two phases: *training* and *serving*. In the training phase [23]–[29], GNNs take as input graph-structured data and feature vectors of each node (Fig. 1 (left)), and learn the *embeddings* of each node in the graph. In the serving phase [9], [30]–[33], the trained model is used to predict the embeddings for previously unseen *query* nodes. Such predictions empower downstream GNN use-cases, such as malicious node prediction in financial applications and social media [34], [35], and integration with LLMs [36], [37].

Ensuring low latency in computing the embeddings of query nodes is key to GNN serving, as GNN models are increasingly used in time-sensitive tasks, including recommendation [9], [33], [38], fraud detection [39], [40], and traffic prediction [41]. At the same time, the underlying graphs are growing in size; modern industry graph datasets routinely span a few billion nodes and trillions of edges [42] and can consume hundreds of TB of memory. Distributing such large graph datasets and associated feature vectors across multiple machines is thus inevitable these days.

Unfortunately, achieving low-latency GNN serving for large graphs stored across machines is extremely challenging. To compute the embeddings of the query nodes, GNN models need the entire k-hop neighborhood. The number of k-hop neighbors can grow exponentially with k, known as the neighborhood explosion problem [43]–[45]. Serving query nodes can thus require excessive amounts of memory and computation costs due to the size of feature vectors and adjacency matrix in the large neighborhood.

Furthermore, unlike training, we do not know how an incoming query node will be connected to the rest of the graph; no matter how the large graph is partitioned and stored, the k-hop neighbors of a query node and their feature vectors are often spread across multiple machines. As a result, creating the query node's "computation graph" (Fig. 1 (right)), by which the node's embedding is computed based on the neighborhood's feature vectors and aggregation paths along the edges, involves fetching the associated data from remote partitions; the resulting massive data volumes (§III-A) cause high communication overheads, worsening serving latency.

GNN training systems facing similar issues adopt mitigation techniques based on *approximation*, such as reusing historical embeddings [46]–[48] or sampling [23], [27], [28], [42], [49]. However, we find they are not directly applicable to serving for two reasons: (a) they do not account for the *data dependence* between an incoming query node and the existing graph – i.e., determining how the query node's connectivity impacts the embeddings of the graph's existing nodes. Ignoring such dependence can lead to unacceptably high accuracy loss (§III-B). (b) While existing techniques reduce computation (either by reusing node embeddings or computing embeddings for fewer nodes by sampling), we find that the communication involved is still heavy at inference time because of the underlying graph's rich connectivity and the query node's neighborhood being spread across machines.

We present Reforge, a GNN serving system that combines data dependency-aware approximation with a new form of computation parallelism to systematically address the over-

heads, supporting low-latency GNN serving on modern large graphs with minimal accuracy loss. Existing serving systems either adopt sampling [9], [30], [33], [50] (with concomitant high latency or substantial accuracy loss) or apply to smaller graphs [31], [32], limiting their practical applicability (§IX).

Similar to some training systems [46], [47], Reforge reuses embeddings of existing nodes after training (which we call "precomputed embeddings (PEs)") instead of computing the embeddings in the entire k-hop neighborhood; this saves network, compute, and memory resources. Because naïvely reusing PEs ignores data dependence and undermines accuracy, Reforge uses *Selective Recomputation of Precomputed Embeddings* (SRPE) (§V). Here, we identify a small number of neighborhood nodes that impact accuracy substantially and focus on recomputing their embeddings while reusing PEs for other nodes. We prove that this statistically minimizes approximation errors. We develop a practical heuristic to perform recomputation on such nodes within a budget, where the budget trades off latency with accuracy.

SRPE significantly reduces the size of computation graphs, but it still entails substantial communication (§VIII-C) due to fetching remote PEs and feature vectors. To mitigate this overhead, we develop *Computation Graph Parallelism* (CGP) (§VI). Unlike existing GNN systems where a single machine solely creates computation graphs by fetching required data from remote machines, CGP allows each machine to build a partitioned computation graph using data available in local partitions. CGP then aggregates the partitioned computations using efficient all-to-all collectives and finally applies *custom merge functions* (tailored to the specific type of GNN model) to the aggregations to compute the final outputs.

We implement Reforge based on DGL [24], [28], a popular GNN training framework, and evaluate it on representative graph datasets and GNN models. Our evaluation results show that the combination of SRPE and CGP helps Reforge outperform DGL-based full-graph and approximation-driven baseline serving systems by up to 159× and 10.8× in latency, respectively, with minimal accuracy loss (§VIII).

We make the following contributions in this paper:

- We show that naïvely applying approximation techniques and reusing computations leads to large errors, and propose SRPE, which statistically minimizes these errors by identifying the parts of computation graphs that need recomputation to avoid accuracy losses (§V).
- We co-design CGP, a new parallelism technique that reduces communication overheads of SRPE with local aggregation by parallelizing the creation and execution of computation graphs for serving various GNN models (§VI).
- We show that Reforge is able to outperform state-of-the-art techniques and to achieve up to orders of magnitude lower latency with minimal accuracy loss (§VIII).

II. Background

A Primer on GNNs. Unlike conventional deep neural networks (DNNs) that process independent input vectors without considering their inter-dependencies, GNNs are designed to handle graph datasets. These datasets consist of node feature vectors and a graph structure that expresses the relationships between nodes, as depicted in Fig. 1 (left). For instance, in social networks, users are the nodes, their profiles are the feature vectors, and friendships are the edges.

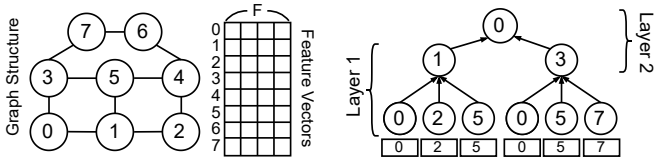


Fig. 1: (*Left*) An example graph dataset with 8 nodes and F-dimensional feature vectors. We use this graph dataset as our running example. (*Right*) The 2-hop computation graph for node 0 where the boxes below represent feature vectors.

GNN models aggregate the neighborhood information of each node to leverage the relationships and compute highly expressive *embeddings* than individual feature vectors alone. To achieve this, the first step is to create a *computation graph* that contains the k-hop neighborhood of each node along with the associated feature vectors, where k is typically 2 or more [51]. Fig. 1 (right) illustrates an example of a 2-hop computation graph for a node.

The *execution* of the computation graphs leverages *message passing* [28], [52], which we can generally express as follows. For a target node v and layer $1 \leq l \leq k$, define:

$$h_v^{(l)} = U^{(l)}(h_v^{(l-1)}, \bigoplus\nolimits_{u \in N(v)}^{(l)} \{M^{(l)}(h_u^{(l-1)})\}) \qquad (1)$$

where $h_u^{(0)}$ is the feature vector for node u, $N(v)$ represents direct neighbors of the node v, and $h_v^{(k)}$ is the final embedding for the node v. Each GNN layer defines (U, $\bigoplus$, M) functions, which represent *update*, *aggregate*, and *message* function, respectively. The message function is applied to the neighbors of node v to create *messages*, which are aggregated by the aggregate function. The layer embedding of each layer is then obtained by applying the update function to the previous layer embedding of node v and the aggregated messages.

Distributed GNN Serving for Large Graphs. Similar to conventional DNN models, GNN models can be trained on a static graph dataset in the background. However, unlike DNN serving, which only needs a trained model, GNN serving additionally requires the graph dataset used in training because of the data dependency in GNN computation. To serve a large graph dataset (e.g., 2 billion nodes and 2 trillion edges [42]), the dataset needs to be distributed across multiple machines [23], [42], [53], [54], and the embeddings for the new *query* nodes are computed with this distributed data.

Fig. 2 illustrates a workflow of distributed GNN serving systems. Here, a training graph dataset is partitioned across multiple machines. A *serving request* consists of the feature vectors of query nodes and the edges between the query nodes and the existing training nodes. When one of the machines receives the request, it creates computation graphs by fetching the required feature vectors and edges from the other machines, computes the embeddings of the query nodes,

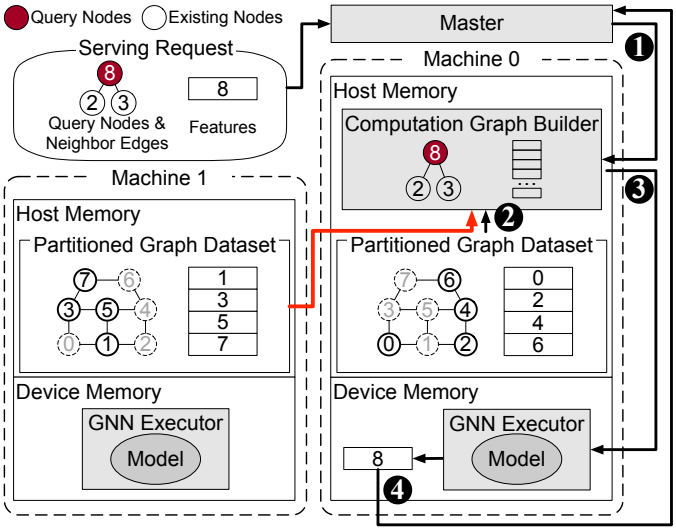


Fig. 2: Distributed GNN serving end-to-end workflow. To generate the embeddings of (batched) query nodes, ❶ the master forwards a serving request to one machine. ❷ A computation graph builder then creates k-hop computation graphs for the query nodes by loading from the local partition and fetching required edges and feature vectors from remote partitions. The red line represents remote communication. ❸ The computation graphs are then executed by a GNN executor after being copied into GPU device memory. ❹ Finally, the embeddings of the query nodes are returned.

and returns the result.

III. Challenges

Unfortunately, the size of graph datasets that distributed GNN serving systems can handle is limited due to unique properties of modern graph data, especially, the data dependency between query nodes and existing graph nodes, which leads to high overheads. Existing mitigation techniques often fall well short due to fundamental drawbacks.

A. Overheads from Large Neighborhoods

GNN serving requires the entire k-hop neighborhood in computing the embeddings of query nodes. Since the number of k-hop neighbors can grow exponentially with k, memory and compute costs can be also significant to create and execute computation graphs, which contain the feature vectors and edges corresponding to the neighborhood. For example, serving with the full (neighborhood) computation graph in Table I can take 711 ms, likely violating the tight latency SLOs of real-world applications [55]–[57]. Also, the memory overhead from these large neighborhoods can easily lead to GPU out-of-memory errors despite their tens of GB memory capacity (§VIII-B).

Moreover, it is hard to predict how query nodes are connected to the rest of the graph and as such feature vectors and edges can be stored across multiple machines (as shown in Fig. 2); thus, creating computation graphs involves fetching neighborhood information and the corresponding feature vectors from remote partitions (Step ❷), resulting in high communication overhead.

Method	Latency (ms)	Accuracy (%)
Full Computation Graph (Full)	711.1	56.9
Neighborhood Sampling (NS)	168.8	50.9 (-6.0)
Historical Embeddings (HE)	66.4	50.6 (-6.3)

TABLE I: Latency and accuracy of different serving methods. We run serving requests of batch size 1,024 with 4 machines. A 3-Layer Graph Attention Networks [59] trained on the Yelp dataset [43] is used. We describe the detailed setup in §VIII-A.

B. Limitations of Today's Approximations

To mitigate problems due to large neighborhoods and overheads, one could employ approximation taking inspiration from similar techniques in GNN *training* systems. Unfortunately, this can lead to significant model accuracy loss (Table I) and, in some cases, may not improve latency.

Historical Embeddings. Instead of creating full computation graphs every iteration, GNN training systems [46]–[48] reuse the layer embeddings from previous iterations, called *historical embeddings* [46], [58]. This technique avoids the exponential expansion of computation graphs, allowing the systems to construct computation graphs only with direct neighbors. One may consider applying this technique to serving by reusing the layer embeddings from the last training epoch. However, as the embeddings are data dependency-unaware, i.e., they do not account for how new query nodes impact existing embeddings, using them for serving can hurt accuracy. As shown in Table I, although serving can be 10.7× faster with this technique, it results in a 6.3% points drop in accuracy.

Sampling. GNN training systems [23]–[25], [27], [28], [42], [49], [53], [60]–[62] employ *sampling* with which they aggregate only a small number of sampled neighbors at each hop [43]–[45], [63]. This drastically reduces the size of the computation graph at the cost of accuracy due to approximation errors from sampling. Multiple training epochs are executed to account for the accuracy loss, and each epoch samples the computation graph differently. Running many epochs leads to convergence. However, in serving, we find that a sampled computation graph can hurt accuracy significantly (6.0% points drop in Table I) since we do not have multiple chances to recover approximation errors from sampling. Moreover, latency can be still high because, even with sampling, the size of computation graphs and the associated communication overhead can be significant (§VIII-B).

IV. Reforge Overview

Reforge addresses GNN serving challenges and overheads due to large neighborhoods through two complementary techniques: smart data dependency-aware approximation and distributed computation graph creation and execution.

T-1: Selective Recomputation of Precomputed Embeddings (§V). To deal with neighborhood explosion, we design a technique called selective recomputation of precomputed embeddings (SRPE) that reduces redundant computations of

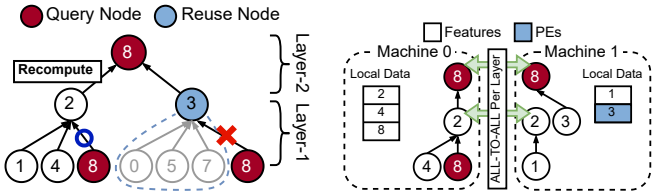


Fig. 3: High-level illustration of (*Left*) Selective Recomputation of Precomputed Embeddings (SRPE) and (*Right*) Computation Graph Parallelism (CGP). The query node 8 is connected to the existing nodes 2 and 3 of the example graph dataset in Fig. 1.

layer embeddings by effectively reusing previously computed embeddings named *precomputed embeddings* (PE). Unlike historical embeddings used in training (§III-B), SRPE mitigates accuracy drops by *selectively recomputing* a portion of the PEs that contributes to high approximation errors. SRPE decides on recomputation nodes based on a simple but effective heuristic policy that statistically minimizes the approximation errors. Fig. 3 (left) illustrates an example usage of SRPE. After training, SRPE precomputes the embedding of existing training nodes. When computing the embedding of query node 8, SRPE reuses the PE of node 3 while *recomputing* the PE of node 2 to better reflect the contribution of node 8 (i.e., the edge from node 8 to 2 at Layer-1).

T-2: Computation Graph Parallelism (§VI). While SRPE reduces the size of computation graphs and corresponding computational overheads, it still has high communication overhead due to fetching PEs and feature vectors; e.g., we find that SRPE alone can still yield latencies as high as 978 ms (§VIII-C) due to the overhead. To further minimize this overhead, we develop *computation graph parallelism* (CGP), which distributes both *creation* and *execution* of computation graphs across multiple machines. In CGP, each machine carefully constructs a *partitioned computation graph* to use only local feature vectors and PEs during execution, thus avoiding heavy remote communication. For instance, in Fig. 3 (right), to compute the Layer-1 embedding of node 2, Machine 0 uses local feature vectors of nodes 4 and 8, while Machine 1 uses the local feature vector of node 1. However, since each machine has a partial view of the entire neighborhood, conventional message passing computation (Eq. 1) fails to generate correct outputs. To overcome this challenge, CGP employs a form of distributed execution that extends message passing with collective communications and *custom merge functions* tailored to various GNN models.

End-to-end Workflow. Fig. 4 illustrates the end-to-end workflow of Reforge. The PEs of all nodes in the training graph dataset are computed after training and distributed across machines along with the graph dataset. ❶ Upon the arrival of a serving request containing a batch of query nodes, the master server evenly splits the query nodes, their feature vectors, and the neighbor edges. ❷ The partitioned requests are then distributed to the computation graph builder in each machine, ❸ which creates a partitioned computation graph.

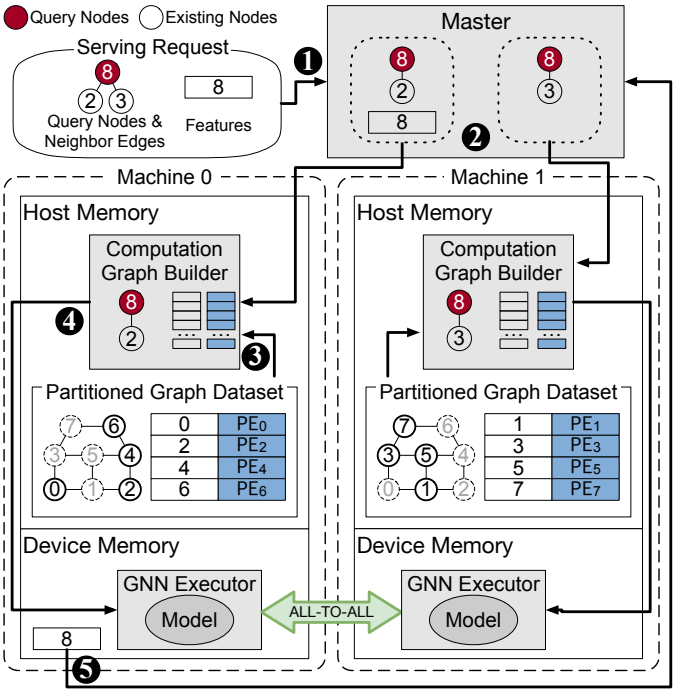


Fig. 4: Reforge end-to-end workflow.

When building the computation graph, each builder determines which PEs to *recompute* and adds the edges required for the recomputation to the computation graph. ❹ Copied to GPU memory, the partitioned computation graphs are executed by GNN executors, which employ local aggregations followed by collective communications to generate the final output embeddings with customized merge functions for GNN models; ❺ the master server collects and returns the output embeddings.

Problem Scope. This paper focuses on achieving low-latency, large-scale GNN serving for *new* query nodes that are unseen during training, while mitigating large computational overheads and approximation errors (§III). We assume that GNN models are trained on a fixed graph (e.g., the example graph dataset in Fig. 1), and during the serving phase, query nodes arrive with edges connecting them to training nodes in the fixed graph (e.g., query node 8 in Fig. 4 is connected to nodes 2 and 3). Addressing potential staleness problems and enabling dynamic updates of graph datasets, GNN models, and PEs is left for future work.

V. Selective Recomputation of PE

To mitigate the overheads from large neighborhoods (§III), Reforge reuses precomputed embeddings (PEs), $p_u^{(l)}$ for each node u and layer l ($1 \leq l \leq k-1$), which are captured from the layer embeddings (i.e., $h_u^{(l)}$ in Eq. 1) after training. During the serving phase, for a query node v, Reforge computes the final embedding $h_v^{(k)}$ without fully recalculating the embeddings for the full k-hop neighborhood. Instead, it substitutes the embeddings of v's direct neighbors with their respective PEs, utilizing $p_u^{(l)}$ in place of computing $h_u^{(l)}$. This approach significantly reduces both computational complexity

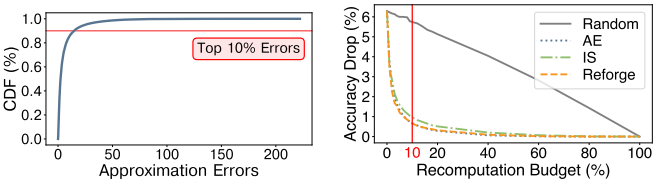


Fig. 5: (*Left*) CDF of the approximation errors of PEs, derived from the workload in Table I. We randomly selected 25% of test nodes as query nodes, computed PEs with the remaining nodes, and aggregated the errors using the query nodes. (*Right*) Accuracy recovered by recomputation policies as budgets grow. AE and IS represent recomputation based on actual approximation errors and node importance scores. Reforge denotes the proposed query edge ratio-based policy (§V-B2).

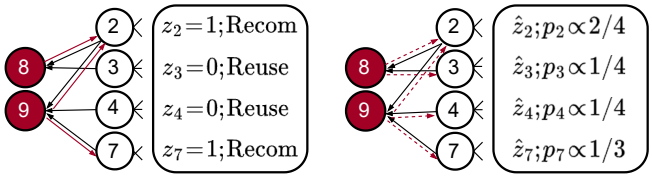


Fig. 6: Selective recomputation in a batch of two query nodes: 8 (connected to nodes 2 and 3) and 9 (connected to nodes 2, 4, and 7). (*Left*) Recomputation of two candidates given a 50% recomputation budget. (*Right*) Random variables with recomputation probabilities such that $\sum p_i = 0.5$.

and memory requirements from $O(N^k)$ to $O(N \times k)$, where N represents the average number of neighbors.

A. Skewed Approximation Errors of PEs

However, naïvely reusing PEs for serving drops accuracy by up to 6.3% points (HE in Table I) because the stored PEs exclude embeddings from new query nodes. For example, the PE of node 3 shown in Fig. 3 (left) is computed considering only the neighbors in the training graph (nodes 0, 5, 7), excluding the new query node 8.

We quantify the approximation errors by computing the difference between the *full embeddings*, $f_u^{(l)}$, which include all the edges from query nodes in aggregation, and the PEs. When u is a direct neighbor of query nodes, we compute its PE approximation error as $\sum_{l=1}^{k-1} ||f_u^{(l)} - p_u^{(l)}||$.

We observe that the approximation errors are *highly skewed* – a small number of PEs dominate the accuracy losses. Fig. 5 (left) shows a distribution of the approximation errors of PEs. The top 10% of PE approximation errors are several orders larger than the other 90%. Recomputing just that 10%, while reusing the rest, recovers accuracy drop from -6.3% points to -0.7% points (AE in Fig. 5 (right)). Exploiting the skew in approximation errors is key to the effectiveness of recomputation; e.g., randomly selecting the 10% recomputation targets (Random in Fig. 5 (right)) yields a marginal accuracy benefit (from -6.3% points to -5.7% points).

B. Policy for Selective PE Recomputation

Based on the observation, we *selectively* recompute PEs that show high approximation errors when serving query nodes. The key challenge lies in devising an effective *recomputation policy* that can identify PEs prone to high approximation errors.

1) Problem Formulation: We illustrate the problem with query nodes 8 and 9 in Fig. 6 (left). Their direct neighbors (nodes 2, 3, 4, 7) form the *recomputation candidates*. A policy decides whether to recompute or reuse each PE. Recomputing includes the query edges, whereas reuse ignores them. Then, given a recomputation budget, the objective is to find the best targets that minimize the approximation errors of the candidates. Fig. 6 (left) illustrates a case where the budget allows for two out of four PEs to be recomputed.

The recomputation policy can be formulated as a constrained optimization problem as follows:

$$\min_{z\in\{0,1\}^{|R|}} \sum_{u\in R}\sum_{l=1}^{k-1} ||f_u^{(l)} - h_{G_z,u}^{(l)}|| \quad \textit{sub. to.} \quad \sum_{u\in R} z_u \le \gamma|R| \tag{2}$$

Here, R is the recomputation candidates, z is a $|R|$-sized 0-1 vector deciding recomputation of each candidate, G_z denotes the corresponding graph, $h_{G_z,u}$ represents the approximated embeddings in G_z, and f_u are the full embeddings. Finally, γ refers to the budget.

However, directly solving the problem at serving time is challenging because it is impractical to compute the full embeddings. Instead, we stochastically approximate the constrained optimization problem. As illustrated in Fig. 6 (right), we assign independent binary random variables $\hat{z}_u$ to each candidate u such that it has a recomputation probability of $p_u = \mathbb{E}[\hat{z}_u]$, within the budget $\gamma = \sum_u p_u$.

In this setting, our goal is to find the optimal probabilities that minimize the approximation errors. To achieve this, we adopt the variance minimization technique from recent work in sampling-based GNN training [43]; we define unbiased estimators for the full embeddings and derive the optimal recomputation probabilities that minimize the variance of the estimators. By selecting candidates with high probabilities, we can indirectly reduce the approximation errors.

We develop unbiased estimators for the full embeddings. For each u, the full embedding $f_u^{(l)}$ is computed with aggregations from the query nodes ($q_u^{(l)}$) and those from the training nodes ($t_u^{(l)}$). Here, we simplify the aggregation as the mean of the messages from neighbors; for example, $q_u^{(l)} = \sum_{v\in N_Q(u)} m_v^{(l)}/|N(u)|$ where $N_Q(u)$ denotes the query nodes connected to u and $m_v^{(l)}$ are the messages. Then, we define the estimator $\hat{f}_u^{(l)} = \frac{1}{p_u}\hat{z}_u q_u^{(l)} + t_u^{(l)}$ such that $\mathbb{E}[\hat{f}_u^{(l)}] = f_u^{(l)}$. We obtain the optimal probabilities that minimize the variance of the estimators with the following theorem.

Theorem 1. *The sum of the variances of every dimension of the estimators* ($\sum_u \sum_{l=1}^{k-1} \hat{f}_u^{(l)}$) *is minimized when* $p_u \propto ||\sum_{l=1}^{k-1} q_u^{(l)}|| = ||\sum_{l=1}^{k-1}\sum_{v\in N_Q(u)} \frac{m_v^{(l)}}{|N(u)|}||$, *given* $\gamma = \sum_u p_u$.

Dataset	GCN		SAGE		GAT	
	Acc. (%)	γ (%)	Acc. (%)	γ (%)	Acc. (%)	γ (%)
Reddit	92.5 (-0.1)	0	96.3 (0.0)	0	95.5 (0.0)	0
Yelp	42.4 (-4.5)	20	63.7 (0.0)	0	56.9 (-6.3)	7
Amazon	41.8 (-3.9)	3	79.4 (0.0)	0	63.4 (-3.0)	1
Products	75.6 (+0.1)	0	77.8 (+0.1)	0	73.3 (0.0)	0
Papers	42.0 (+1.4)	0	50.0 (+0.3)	0	49.3 (+0.4)	0

TABLE II: Effectiveness of Reforge's recomputation policy (§V-B2). 'Acc.' column represents the accuracy with full computation graphs and the percentage of accuracy drops with PEs (without recomputation). 'γ' column shows the recomputation budget to achieve less than 1% points of accuracy drop.

We describe the high-level idea of the proof of the theorem. Given a constant sum of probabilities γ, we derive a lower bound of the estimator using the Cauchy–Schwarz inequality. Equality holds when the recomputation probabilities align with those specified in the theorem.

2) The Top-Query-Edges-Ratio Policy: Based on the analysis, a recomputation policy can calculate the optimal probabilities and recompute the PEs within the top γ% of these probabilities. However, the exact calculation requires the full embeddings of query nodes to obtain the messages ($m_v^{(l)}$) of query nodes, which is infeasible in serving. Thus, instead of computing the full embeddings, we adopt the simplification from prior work [43], whereby our recomputation policy approximates $p_u \propto \frac{|N_Q(u)|}{|N(u)|}$ without considering the message terms. We explain the intuition and highlight the policy's effectiveness below.

We name our policy *top-query-edges-ratio* since it recomputes PEs with higher ratios of edges from query nodes. For instance, in Fig. 6 (right), the policy recomputes the PEs of nodes 2 and 7. Intuitively, a higher ratio of query edges is likely to change the PE significantly and result in a large approximation error. Empirically, Fig. 5 (right) shows this policy (Reforge) can recover accuracy drops almost similarly to recomputation based on real approximation errors (AE).

To further assess the effectiveness of Reforge's policy, we compare it with two alternatives, IS and Random, as shown in Fig. 5 (right). IS chooses recomputation targets based on node importance scores, defined for each node v as $IS(v) = \frac{1}{\deg(v)} \sum_{u \in N(v)} \frac{1}{\deg(u)}$, following existing sampling-based training methods [43], [45], [58]. Though beneficial for reducing sampling variance during training, these scores focus more on existing nodes and are less effective for recomputation during serving, resulting in less optimal performance compared to Reforge's policy. Random selects targets randomly. This approach fails to identify error-prone PEs, leading to ineffective recovery of accuracy losses, even with large recomputation budgets.

In Table II, we evaluate Reforge's policy across representative GNN models and graph datasets, which shows it can effectively recover accuracy to within 1% point with minimal budget. Our policy consistently outperforms the alternatives, IS and Random, similar to the result in Fig. 5 (right). We extensively evaluate the recomputation policies across diverse GNN models and graph datasets, and observe consistent results. Detailed results are omitted due to space constraints.

Finally, Reforge allows for changing the amount of recomputation through the recomputation budget parameter γ to navigate the tradeoff between accuracy and latency. Users can adjust the parameter depending on the model, dataset, and acceptable percentage of accuracy drop and latency. We evaluate this tradeoff in §VIII-D and show that Reforge can recover accuracy drops with minimal latency increase. We plan to investigate automatic selection of the recomputation budget based on runtime distributions of query nodes in future work.

Algorithm 1 Reforge Distributed Execution of CGP

Params: n_parts: number of partitions, γ: recomputation ratio, n_layers: number of GNN layers, q_n: query nodes, q_e: query edges, q_f: query features

def Serving_Request_Partitioning(q_n, q_e, q_f):
  $part\_reqs \leftarrow [\,]$
  for $p\_i \leftarrow 0$ to $n\_parts - 1$ **do**
    $p\_e \leftarrow \{(u \rightarrow v) \in q\_e | part(u) = p\_i\}$
    $p\_f \leftarrow \{q\_f[w] | w \in q\_n, part(w) = p\_i\}$
    $part\_reqs \leftarrow part\_reqs \cup \{q\_n, p\_e, p\_f\}$
  return $part\_reqs$

def Computation_Graph_Generation(q_n, p_e):
  $last\_cg \leftarrow (q\_n, p\_e)$ ▷ Destination nodes and edges
  $rc\_cands \leftarrow srcs(last\_cg) \backslash q\_n$
  $rc\_n \leftarrow$ `Nodes with top` γ `% query-edge-ratio in` rc_cands
  $reuse\_n \leftarrow rc\_cands \backslash rc\_n$
  $rc\_n \leftarrow ALL\_GATHER(rc\_n)$
  $rc\_e \leftarrow local\_part.edges(rc\_n)$
  $rc\_cg \leftarrow (q\_n \cup rc\_n, p\_e \cup rc\_e)$
  return $last\_cg, rc\_cg, reuse\_n$

def Computation_Graph_Execution($last\_cg, rc\_cg, reuse\_n, p\_f$):
  $n \leftarrow srcs(rc\_cg) \backslash q\_n$
  $embs \leftarrow local\_part.feats(n) \cup p\_f$
  for $l \leftarrow 1$ to $n\_layers - 1$ **do**
    $embs \leftarrow Dist\_Layer\_Exec(rc\_cg, embs)$
    if $l < n\_layers - 1$ **then**
      $n \leftarrow srcs(rc\_cg) \backslash dsts(rc\_cg)$
      $embs \leftarrow embs \cup local\_part.pes(n, l)$
    else
      $embs \leftarrow embs \cup local\_part.pes(reuse\_n, l)$
  return $Dist\_Layer\_Exec(last\_cg, embs)$

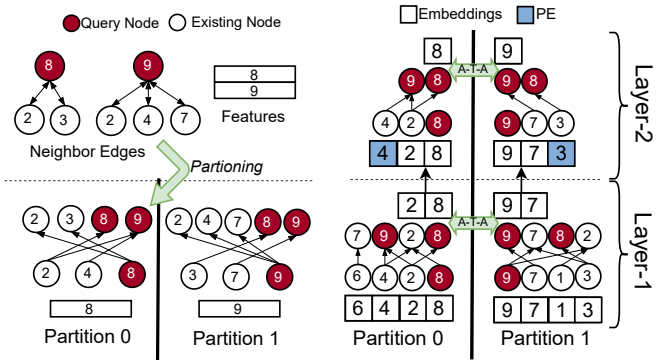
VI. Computation Graph Parallelism

While SRPE mitigates the neighborhood explosion problem, communication overhead remains significant. We observe that over 80% of the latency is due to creating computation graphs, which involves fetching feature vectors, PEs, and neighbor edges from remote machines and copying them to GPU memory (§VIII-C). We now describe *computation graph parallelism* (CGP) to further reduce communication overhead by distributing computation graph creation and execution.

A. Distributed Creation and Execution

We describe the key steps in CGP (Algorithm 1) using an example serving request with two batched query nodes (i.e., nodes 8 and 9) at the top of Fig. 7a.

Serving Requests Partitioning. Reforge's master first splits incoming serving requests. Specifically, the master evenly assigns partitions to every query node in a request. In our example, nodes 8 and 9 are assigned to partitions 0 and 1, respectively. The edges in a request are then split based on the source nodes' partitions to enable local aggregation. For instance, the edges whose source nodes are 2, 4, and 8 are



(a) Serving Request (b) Computation Graph

Fig. 7: (a) An example serving request on the graph dataset in Fig. 1 with the partitioned requests depicted at the bottom. (b) Partitioned computation graphs for the request.

Algorithm 2 Reforge Distributed Layer Execution with Attention Mechanism

Params: M: weights matrix, $\vec{a}$: learnable attention weights, n_parts: number of partitions

1: **def** Dist_Layer_Exec($dst\_nodes, edges, embs$):
2: **for** $dst\_node \in dst\_nodes$ **do** ▷ Compute local aggregations
3: $h\_dst \leftarrow M(embs[dst\_node])$
4: $local\_srcs \leftarrow \{u | (u, dst\_node) \in edges\}$
5: $h\_local\_srcs \leftarrow M(embs[local\_srcs])$
6: $logits \leftarrow \vec{a}(h\_dst || h\_local\_srcs)$
7: $logit\_max \leftarrow max(logits)$
8: $exp\_sum \leftarrow \sum_{l \in logits} exp(l - logit\_max)$
9: $aggr \leftarrow \sum_{u \in local\_srcs} \frac{exp(logits[u] - logit\_max)}{exp\_sum} h\_local\_srcs[u]$
10: ALL-TO-ALL for local $aggr$, exp_sum, and $logit\_max$ across partitions into $aggrs$, exp_sums, and $logit\_maxs$
11: **for** $dst\_node \in dst\_nodes$ **do** ▷ Merge local aggregations
12: $g\_logit\_max \leftarrow \max(logit\_maxs)$
13: $rescales \leftarrow exp(logit\_maxs - g\_logit\_max)$
14: $g\_exp\_sum \leftarrow \sum_{p \in n\_parts} exp\_sums[p] * rescales[p]$
15: $outputs[dst\_node] \leftarrow \sum_{p \in n\_parts} \frac{aggrs[p] * rescales[p]}{g\_exp\_sum}$
16: **return** $outputs$

included in the partitioned request for partition 0, as depicted at the bottom left of Fig. 7a. The feature vectors of query nodes follow the partitions of the query nodes. The master sends the partitioned edges and features to corresponding computation graph builders.

Computation Graph Generation. Among the edges, each builder uses those having query nodes as destination nodes to create the last layer of a computation graph (e.g., Layer-2 in Fig. 7b). Each builder then applies Reforge's recomputation policy (§V-B2) to identify recomputation target nodes among the input nodes of the last layer. In Layer-2 of Fig. 7b, while the PEs of nodes 3 and 4 are reused, the embeddings of nodes 2 and 7 are recomputed. For the recomputation, all of the edges terminating at the target nodes need to be included in the other layers. Since the edges are distributed across different partitions, the builders first broadcast their recomputation target nodes to each other through all-gather collective (e.g., nodes 2 and 7), extract the required edges from local graphs (e.g., $4 \rightarrow 2$ and $6 \rightarrow 7$ in Layer-1 of partition 0 in Fig. 7b) and from the partitioned requests (e.g., $8 \rightarrow 2$), then create the remaining layers (e.g., Layer-1 in Fig. 7b).

Layer-wise Distributed Execution. With the computation graphs, GNN executors compute the embeddings of the query nodes by executing GNN layers sequentially. To first compute Layer-1 embeddings, each executor computes partial aggregations using its local features, shuffles them with the other executors through collective communications, and merges them. The output embeddings are concatenated with the local PEs, which are used as the input for the next layer (e.g., Layer-2 input embeddings in Fig. 7). The process is repeated until the query nodes' embeddings are computed in the last layer execution.

To enable the distributed execution, Reforge extends conventional message passing (Eq. 1). Instead of applying the aggregation function to the entire neighborhood, Reforge generates a local aggregation for each partition using the neighbors placed in each partition. Reforge further defines a merge function to compute global aggregations with the local aggregations. The l-th layer execution is as follows:

$$a_{v,p}^{(l)} = \hat{\bigoplus}_{u \in N_p(v)}^{(l)} \{M^{(l)}(h_u^{(l-1)})\}, 0 \leq p < P$$
$$h_v^{(l)} = U^{(l)}(h_v^{(l-1)}, \biguplus^{(l)} \{a_{v,p}^{(l)}\}_{0 \leq p < P}) \quad (3)$$

Here, P is the number of partitions, $N_p(v)$ returns the neighbors of a node v in a partition p, $a_{v,p}^{(l)}$ represents a local aggregation for v in p, $\hat{\bigoplus}$ is the local aggregation function, and $\biguplus$ is the merge function.

B. Custom Merge Functions

Applying CGP is straightforward for GNN models with commutative and associative aggregations, such as sum or max, but not so for more complex generalized arithmetic aggregations [64], [65] or softmax-based aggregation with learned weights [59], [66]. We explain how to handle them below.

Generalized Arithmetic Aggregation. Beyond simple average functions, some GNN models leverage power mean aggregation [65] or normalized moments aggregation [64]. For power mean aggregation (i.e., $(\frac{1}{|N(v)|}\sum_{u \in N(v)} m_u^p)^{\frac{1}{p}}$), Reforge computes the local aggregations by summing local messages after applying the `pow()` function with p. To merge them, Reforge adds the aggregations and then applies the `pow()` function with $1/p$. Normalized moments aggregation (i.e., $(\frac{1}{|N(v)|}\sum_{u \in N(v)}(m_u - \bar{m})^n)^{\frac{1}{n}}$) requires the mean of the messages ($\bar{m}$). Reforge first computes the mean values for each destination node then broadcasts the values with an all-gather operation. Then, Reforge computes the global aggregations with the same procedure used in the power mean aggregation.

Softmax-based Aggregation. GNN models can utilize the attention mechanism [59], [66], which learns attention weights using the softmax function [67] to identify important nodes in the neighborhood. To employ the aggregation, a learned attention weight needs to be computed for each edge, which requires the embeddings of destination nodes [59]. As described in Algorithm 2, Reforge optionally employs an all-

Dataset	Nodes	Edges	Avg. Deg.	Features	Hiddens
Reddit [44]	232 K	115 M	492	602	128
Yelp [43]	717 K	14.0 M	20	300	512
Amazon [43]	1.6 M	264 M	168	200	512
Products [74]	2.4 M	124 M	52	100	128
Papers [74]	111 M	1.6 B	14	128	512
FB10B [75]	30 M	10 B	333	1024	128

TABLE III: Graph datasets used in the evaluation. The first four columns represent the number of nodes, number of directed edges, average degrees, and feature dimensions. The last column denotes the hidden dimensions of GNN models.

gather operation for destination embeddings (i.e., $h_8^{(1)}$ and $h_9^{(1)}$) to enable local aggregation. Next, the attention weights are normalized using the softmax function. For merging these softmax-based local aggregations while maintaining numerical stability, Reforge additionally generates the exponential sum of logits and the maximum logits in each partition and adds them to the local aggregations, which are then utilized in the merging step.

We note there are some stateful aggregations [44], [68] to which CGP cannot be directly applied; e.g., those leveraging recurrent neural networks (RNNs) [69] to aggregate messages with recurrent hidden states. The strict dependency on aggregation prevents CGP from computing local aggregations in parallel. Since the RNN model must be sequentially applied to each neighbor in the resulting central aggregation, serving latency can become excessive. For instance, in our measurements of serving latency for a 3-Layer GraphSAGE [44] model with RNN-based and mean aggregations under identical settings in Table I, we observe that the RNN-based aggregation requires substantially more time, taking 41× longer (28.9 s) compared to mean aggregation (702 ms).

VII. Implementation

We implement Reforge in C++ and Python on top of the distributed Deep Graph Library (DGL) [24] and PyTorch [70]. Specifically, Reforge reuses DGL's graph structures and message-passing APIs together with PyTorch's DNN operators and communication back-ends. Computation graph builders communicate via PyTorch GLOO [71], while the master node coordinates with workers through PyTorch RPC [72]. During execution, Reforge customizes DGL's message-passing primitives to replace global aggregation with local aggregation, all-to-all communication, and merging. It uses DGL for local aggregation, PyTorch NCCL [73] for cross-GPU communication, and PyTorch's DNN operations for the merging.

VIII. Evaluation

We evaluate the performance of Reforge using popular benchmark graph datasets and representative GNN models. We compare DGL-based [24], [28] baseline serving systems with our techniques, SRPE (§V), CGP (§VI), and their combination. Our evaluation addresses the following key questions:

- How much does Reforge improve latency and accuracy compared to the baseline systems (§VIII-B)?
- To what extent do SRPE and CGP contribute to the reduction in Reforge's latency (§VIII-C)?
- What are the recomputation and memory overheads of using PEs (§VIII-D)?
- How well does Reforge scale with additional GPUs and machines (§VIII-E)?

A. Experimental Setup

Testbed. We conducted our experiments on a GPU cluster with 4 servers, each equipped with two 32-core AMD 7542 CPUs, 512 GB of main memory, and two NVIDIA V100S GPUs with 32 GB of memory. The servers are connected with 25 Gbps Ethernet links via a switch. All servers run 64-bit Ubuntu 22.04 and DGL v1.0.2.

GNN Models. We evaluate Reforge on three representative GNN models, Graph Convolutional Networks (GCN) [76], GraphSAGE (SAGE) [44], and Graph Attention Networks (GAT) [59]. We use two layers for GCN and three layers for SAGE and GAT models to avoid the over-smoothing problem [77]–[79] with deeper GNN models.

Datasets. We evaluate Reforge on six graph datasets detailed in Table III. Products [74] and Papers [74] are widely used for assessing GNN performance [23], [42], [46], [61]. Reddit [44], Yelp [43], and Amazon [43] represent real-world web services where GNNs power applications like recommendations. For testing Reforge on a larger scale, we include FB10B, a synthetic dataset with 10 billion edges modeled after Facebook's social network [75] with randomly generated 1,024-dimensional features, reflecting the high dimensionality in real-world scenarios [9], [62], [80].

Workloads. We synthesize serving workloads because no public large-scale GNN traces exist. For each dataset, we hold out 25% of test nodes, remove their incident edges, and later form requests by sampling query nodes from this set and reinstating their edges to the retained graph. We generate 500 requests, each with a batch size of 1,024, for each graph dataset and reuse them across experiments. Requests run sequentially, and the averaged latencies are reported.

Baselines. For a fair comparison with Reforge, which is based on Distributed DGL (§VII), we implement the following two baseline GNN service systems using DGL [24].[1]

- **DGL (Full)** constructs and executes full computation graphs that consist of the entire k-hop neighborhood of query nodes. This approach is typically preferred for relatively small graphs [31], [32].
- **DGL (NS)** employs neighborhood sampling to reduce latencies at the expense of accuracy [9], [30], [33]. We use widely adopted sampling fanouts [30], [44], [50], [81] of (25, 10)[2] and (15, 10, 5) for models with 2 and 3 layers unless otherwise specified.

[1]Since DGL is primarily designed for training, we adapt distributed DGL [24] by removing the backward pass and using only the forward pass in both baselines.

[2]The (25, 10) fanout means sampling at most 10 neighbors at the first hop and sampling at most 25 neighbors for each direct neighbor.

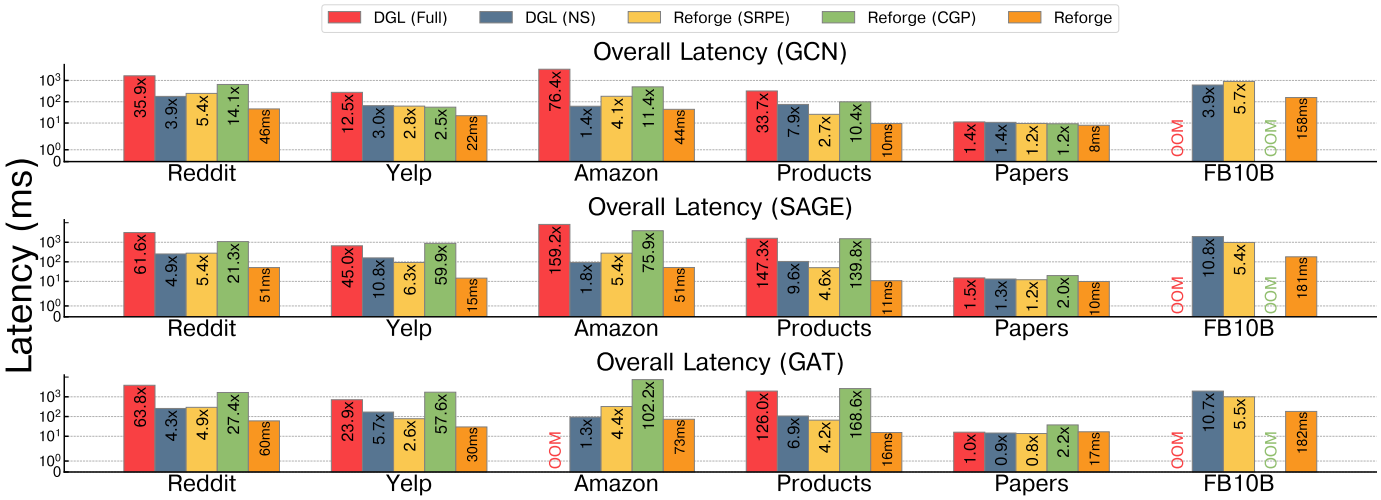


Fig. 8: End-to-end serving latency of Reforge and baseline systems (in log scale) across three models and six datasets. The numbers within each bar represent the latency values for Reforge or the relative speedup of Reforge compared to each system. 'OOM' indicates that a system failed to execute the workload due to a CUDA Out-of-Memory error.

Default Configurations. Unless noted otherwise, we use a batch size of 1,024 and run on 4 machines, each with one GPU. Reforge applies the recomputation policy (§V-B2), with budget γ chosen to keep the accuracy loss under 1% (Table II). For FB10B (which uses synthetic feature vectors), we disable recomputation. We use random hash partitioning by default for load balancing. We do not use locality-aware partitioners such as Metis [82], as they are applied before the arrival of query nodes and therefore do not improve the locality of neighbors for the queried nodes, yielding marginal benefit in our setting.

B. Overall Performance

We evaluate the end-to-end serving latency and accuracy of Reforge in comparison to DGL (Full) and DGL (NS). In Fig. 8, we report the latencies of Reforge and the baseline systems, along with relative speedups, across six datasets listed in Table III and GCN, SAGE, and GAT models. To highlight the contributions of SRPE and CGP, we also include latencies for Reforge (SRPE) and Reforge (CGP), representing Reforge with only SRPE and CGP, respectively.

Compared to DGL (Full). Reforge significantly reduces latency while minimizing accuracy drop. For instance, as shown in Fig. 8, *Reforge achieves 159× lower latency than DGL (Full) for the SAGE model with the Amazon dataset*. Reforge handles large graphs with higher degrees efficiently, leveraging SRPE's precomputation and CGP's local aggregation. For the FB10B dataset, Reforge successfully runs GNNs on the largest dataset by significantly reducing computation graph sizes, whereas DGL (Full) fails due to CUDA out-of-memory (OOM) errors. The Papers dataset shows little benefit from Reforge since its serving nodes have low degrees (2.4 on average), and DGL (Full) shows low latency of 15 ms.

Compared to DGL (NS). We first study how neighborhood sampling affects accuracy. For GCN and GAT models on Amazon and Yelp datasets, accuracy loss from sampling compared to Reforge is significant (1.7% to 5.0% in Table IV),

Systems	Yelp			Amazon		
	GCN	SAGE	GAT	GCN	SAGE	GAT
DGL (NS)	36.2%	63.7%	50.9%	36.1%	79.4%	60.8%
Reforge	41.5%	63.7%	55.9%	40.9%	79.4%	62.5%

TABLE IV: Test accuracies of Reforge and DGL (NS) using three GNN models (GCN, SAGE, GAT) on the Yelp and Amazon datasets. Reforge employs SRPE with the recomputation budgets from Table II, maintaining less than 1 % point of accuracy loss. Latency results are provided in Fig. 8.

while Reforge achieves substantial latency speedups (1.3× to 5.7× in Fig. 8) with less than 1% point accuracy loss. SAGE models are more resilient to sampling, typically showing negligible accuracy losses, as reported in prior work [81]. Overall, Reforge demonstrates minimal accuracy drop while significantly reducing latency across all cases.

Additionally, sampling allows DGL (NS) to handle the largest dataset, FB10B, without CUDA OOM. However, even with sampling, the total neighbors can be large. For instance, with a (15, 10, 5) sampling configuration, one query node can have up to 750 3-hop neighbors, leading to high communication overheads for fetching associated feature vectors, resulting in a latency of 2.0 seconds for FB10B. In contrast, *Reforge effectively reduces computation graph sizes through SRPE and minimizes communication overheads with CGP, outperforming DGL (NS) by up to* 10.8×.

C. Contributions of Reforge's Techniques

To better understand the performance benefits of Reforge, we analyze the respective contributions of SRPE and CGP.

Contribution of SRPE. We present the latency breakdown and required communication size of DGL (Full), DGL (NS), Reforge (SRPE), and Reforge in Fig. 9. DGL (Full) suffers from fetching feature vectors and edges for the entire k-hop neighborhood during computation graph creation (`Fetch`),

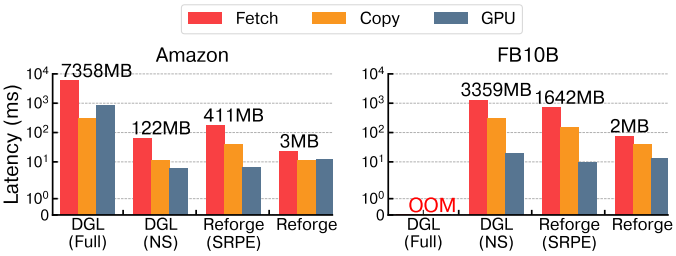


Fig. 9: Latency breakdown of SAGE model with Amazon and FB10B datasets in Fig. 8, which consists of three components: 'Fetch' builds computation graphs by fetching necessary data (i.e., remote feature vectors, edges, and PEs), 'Copy' transfers the data into GPU device memory, and 'GPU' runs GNN computations in GPUs (including collective communications for Reforge's CGP). The size of data in fetching and collective communications is noted on top of each bar.

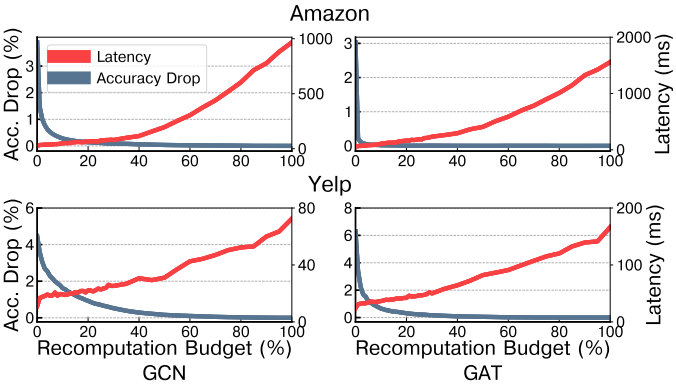


Fig. 10: Latency and accuracy trade-off of Reforge varying recomputation budget with GCN and GAT models and Amazon and Yelp datasets.

taking several seconds to serve a single request or causing CUDA OOM. In contrast, SRPE reduces computation graph sizes and communication overhead. For example, in the Amazon dataset, *SRPE reduces the communication volume by 18×, from 7.4GB to 411MB, and latency by 29×, from 8.2s to 278ms*.

Contribution of CGP. CGP's local aggregation effectively reduces the communication overhead of SRPE, as it only requires collective communications for the query nodes and a small number of recomputation target nodes (e.g., nodes 2, 7, 8, and 9 in Fig. 7b). As shown in Fig. 9 (denoted Reforge), *CGP's local aggregations minimize communication to a few MBs of necessary collective communications, reducing latency of Reforge (SRPE) by 5.5× (from 278 ms to 51 ms) and 5.4× (from 978 ms to 181 ms)* for the Amazon and FB10B datasets.

We observe that CGP alone can provide limited benefit. As Fig. 8 shows (denoted Reforge (CGP)), while CGP reduces latency for 2-hop computation graphs (i.e., GCN) relative to DGL (Full), its performance for 3-hop graphs (i.e., SAGE, GAT) is poorer since deeper GNN models' neighborhoods expand to encompass most of a dataset, limiting local aggregation effectiveness.

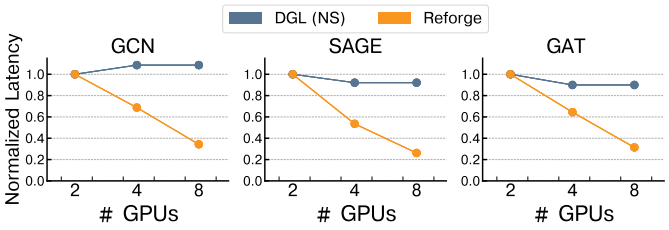


Fig. 11: Normalized Latency of Reforge and DGL (NS) varying number of GPUs with FB10B dataset.

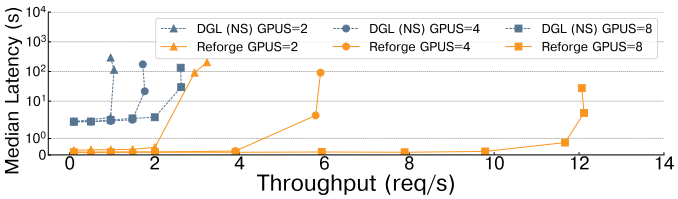


Fig. 12: Latency-throughput results of Reforge and DGL (NS) varying request rates with SAGE model and FB10B dataset.

D. Recomputation and Memory Overhead of PEs

While SRPE provides significant latency benefits, it also introduces additional costs: recomputation (to mitigate accuracy losses) and memory (to store PEs on the host). We evaluate the impact of recomputation on latency and accuracy, and discuss the memory overhead of PEs.

Recomputation Overhead. We assess the trade-offs between latency and accuracy based on the budget of Reforge's recomputation policy (§V-B). Fig. 10 presents results for GCN and GAT models using the Yelp and Amazon datasets, which experience the largest accuracy drops without PE recomputation. *The results demonstrate that Reforge's policy effectively minimizes recomputation costs; for instance, a recomputation budget of 7% reduces accuracy losses from 6.3% points to just 1.0% points, while increasing latency by only 11 ms for the GAT model on the Yelp dataset.*

Memory Overhead. The memory required to store PEs is proportional to the number of layers, hidden dimensions, and data type size: $(L-1)*H*D$ bytes. Among the evaluated workloads, the Papers100M dataset results in the largest PE memory footprint. For a 3-layer GNN with 512-dimensional hidden states, storing PEs requires approximately 455 GB. Reforge stores PEs in host CPU memory and distributes them across machines using CGP; in our 4-machine evaluation cluster with 2 TB of aggregate host RAM, this corresponds to 22% of the available memory. This enables Reforge to eliminate recursive neighborhood fetching during serving, yielding order-of-magnitude latency reductions (Fig. 8).

E. Scalability Analysis

In this section, we examine the impact of varying the number of machines and GPUs on the latency and throughput of Reforge. Additionally, to minimize communication and host-to-device transfer overheads, existing GNN serving systems often employ GPU feature caching [30], [81]. We therefore

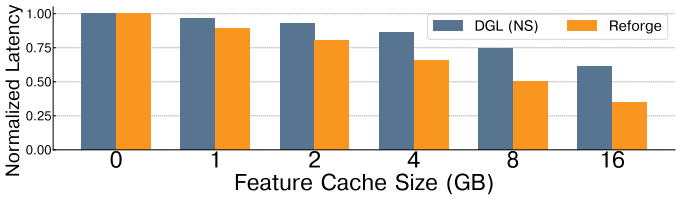


Fig. 13: Latency of Reforge and DGL (NS) varying feature cache size with SAGE model and FB10B dataset.

analyze the effect of GPU feature caching as part of our evaluation. We evaluate performance on the largest FB10B dataset and compare Reforge with DGL (NS), as DGL (Full) is unable to run the workload due to CUDA OOM errors.

Latency. In Fig. 11, we vary the number of GPUs: 2 (on 2 machines), 4 (on 4 machines), and 8 (with 4 machines each using 2 GPUs). While DGL (NS) does not benefit from additional resources due to its centralized execution, Reforge demonstrates strong scaling by distributing the host-to-GPU memory transfer across multiple GPUs, with each GPU handling only local feature vectors and PEs. *As a result, for instance, the latency of Reforge with the GAT model decreases by 67% (from 282ms to 88ms), whereas DGL (NS) only shows a 9% (from 2.2s to 2.0s) reduction in latency.*

Throughput. We analyze the throughput of Reforge and DGL (NS) by modeling request arrivals with a Poisson distribution and feeding workloads into both systems in Fig. 12. While DGL (NS) enables concurrent request handling by individual GPUs, significant network contention limits its scalability, achieving only a 2.6× increase in maximum throughput from 2 to 8 GPUs. In contrast, *Reforge leverages CGP to reduce communication overhead and avoid contention, resulting in a 3.8× increase in throughput. Furthermore, with 8 GPUs, Reforge outperforms DGL (NS) by 4.7× while maintaining significantly lower latency.*

Impact of Feature Cache. We evaluate how the feature caches in GPUs impact the serving latency of Reforge and DGL (NS). Following prior work [27], we sort nodes by their out-degrees, caching the highest-ranked nodes to maximize the chances of cache hits. As shown in Fig. 13, *both Reforge and DGL (NS) benefit from caching, but Reforge sees a much larger reduction in latency. For example, with a 16GB cache, Reforge's latency drops by 65%, compared to 39% for DGL (NS).* This is because Reforge takes advantage of CGP's distributed execution, where each GPU caches frequently accessed nodes from its local subset of data. In contrast, DGL (NS) uses a centralized execution approach, requiring every GPU to consider nodes from the entire dataset.

IX. Related Work

GNN Serving Systems. Several systems [9], [30], [33], [50] accelerate GNN serving for large graphs through sampling. However, as discussed in §III-B and §VIII-B, sampling often results in significantly higher latency and/or lower accuracy. Other systems [31], [32] focus on resource-efficient techniques for optimizing GNN serving in decentralized environments. Serving with the full computation graphs in these systems is prohibitively memory-intensive and slow for large-scale graphs (§VIII-B). In contrast, Reforge's SRPE efficiently reduces computation graph size through selective recomputation, achieving higher accuracy than sampling-based approaches while requiring smaller graphs (§VIII-B). Additionally, while existing GPU-based GNN serving systems such as Quiver [30] and SALIENT [81] rely on centralized execution with feature caching limited to single-GPU memory, Reforge employs CGP to enable distributed execution with pooled GPU memory across the cluster, which improves the effectiveness of feature caching at scale (§VIII-E).

RIPPLE [83] and InkStream [84] incrementally update node embeddings from the deltas caused by query nodes to address the overhead of GNN serving. However, those systems primarily support linear aggregation and have limited support for attention-based models. In contrast, Reforge systematically employs CGP with custom merge functions, supporting a wide range of GNN models.

GNN Kernel Optimizations. Existing systems optimize GNN execution on GPUs using techniques like kernel fusion and pipelining to enhance performance [26], [85]–[87]. For example, GNNAdvisor [85] reduces atomic operations and global memory access through runtime graph-aware warp and block-level organization. While these optimizations could be integrated into Reforge, they do not address communication overheads in serving. Similarly, HAG [88] introduces hierarchical aggregation to eliminate redundancy, which is comparable but distinct from CGP's local aggregation.

GNN Training Systems. To mitigate the overhead of large neighborhoods, many GNN training systems use approximations such as historical embeddings [46]–[48] or sampling [23]–[25], [27], [28], [42], [49], [53], [60]–[62]. However, both approaches often fall short of preserving full model accuracy. Historical embeddings do not account for dependencies introduced by query nodes, degrading accuracy (Table II), while sampling approximates the full neighborhood and causes accuracy loss (Table IV). Moreover, these systems rely on centralized execution, leading to high latency in GNN serving (Fig. 8). Reforge achieves minimal accuracy loss via SRPE and reduces latency via CGP, outperforming these techniques in both dimensions (§VIII-B). Full-graph training systems [26], [89]–[91] avoid computation graph construction by operating on the entire graph, but are designed for static graphs and lack dynamic computation graph creation for new query nodes, making their techniques either inapplicable or complementary to Reforge.

X. Conclusion

We present Reforge, a distributed GNN serving system designed to deliver low latency and minimal accuracy loss for large graphs. Reforge leverages two key techniques, SRPE (§V) and CGP (§VI), to mitigate the substantial overheads from large neighborhoods. Our evaluation shows that Reforge can achieve up to orders of magnitude lower latency than baseline serving systems with minimal accuracy loss.

REFERENCES

[1] Y. Wu, D. Lian, Y. Xu, L. Wu, and E. Chen, "Graph convolutional networks with markov random field reasoning for social spammer detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 01, 2020, pp. 1054–1061.

[2] A. Fout, J. Byrd, B. Shariat, and A. Ben-Hur, "Protein interface prediction using graph convolutional networks," *Advances in neural information processing systems*, vol. 30, 2017.

[3] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[4] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, "Graph neural networks for social recommendation," in *The World Wide Web Conference*, ser. WWW '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 417–426. [Online]. Available: https://doi.org/10.1145/3308558.3313488

[5] J. Suarez-Varela, P. Almasan, M. Ferriol-Galmes, K. Rusek, F. Geyer, X. Cheng, X. Shi, S. Xiao, F. Scarselli, A. Cabellos-Aparicio, and P. Barlet-Ros, "Graph neural networks for communication networks: Context, use cases and opportunities," *IEEE Network*, pp. 1–8, 2022.

[6] F. Geyer and S. Bondorf, "DeepTMA: Predicting effective contention models for network calculus using graph neural networks," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019, pp. 1009–1017.

[7] Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, "Machine learning in chemoinformatics and drug discovery," *Drug Discovery Today*, vol. 23, no. 8, pp. 1538 – 1546, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1359644617304695

[8] J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay, and J. J. Collins, "A deep learning approach to antibiotic discovery," *Cell*, vol. 180, no. 4, pp. 688 – 702.e13, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092867420301021

[9] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 974–983. [Online]. Available: https://doi.org/10.1145/3219819.3219890

[10] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. P. Hsu, and K. Wang, "An overview of Microsoft Academic Service (MAS) and applications," in *Proceedings of the 24th International Conference on World Wide Web*, ser. WWW '15 Companion. New York, NY, USA: Association for Computing Machinery, 2015, p. 243–246. [Online]. Available: https://doi.org/10.1145/2740908.2742839

[11] S. Deng, H. Rangwala, and Y. Ning, "Learning dynamic context graphs for predicting social events," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1007–1016. [Online]. Available: https://doi.org/10.1145/3292500.3330919

[12] W. Jiang and J. Luo, "Graph neural network for traffic forecasting: A survey," *Expert Systems with Applications*, p. 117921, 2022.

[13] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," *Advances in neural information processing systems*, vol. 33, pp. 17 804–17 815, 2020.

[14] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. D. Manning, and J. Leskovec, "GreaseLM: Graph REASoning enhanced language models," in *International Conference on Learning Representations*, 2021.

[15] J. Yang, Z. Liu, S. Xiao, C. Li, D. Lian, S. Agrawal, A. Singh, G. Sun, and X. Xie, "GraphFormers: GNN-nested transformers for representation learning on textual graph," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 28 798–28 810. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/f18a6d1cde4b205199de8729a6637b42-Paper.pdf

[16] Y. Tian, H. Song, Z. Wang, H. Wang, Z. Hu, F. Wang, N. V. Chawla, and P. Xu, "Graph neural prompting with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 080–19 088.

[17] Y. Shi, A. Zhang, E. Zhang, Z. Liu, and X. Wang, "ReLM: Leveraging language models for enhanced chemical reaction prediction," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://openreview.net/forum?id=gJZqSRfV21

[18] H. Abdine, M. Chatzianastasis, C. Bouyioukos, and M. Vazirgiannis, "Prot2text: Multimodal protein's function generation with gnns and transformers," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, 2024, pp. 10 757–10 765.

[19] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[21] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," *Transactions on Machine Learning Research*, 2022. [Online]. Available: https://openreview.net/forum?id=yzkSU5zdwD

[22] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[23] S. Gandhi and A. P. Iyer, "P3: Distributed deep graph learning at scale," in *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*, 2021, pp. 551–568.

[24] D. Zheng, C. Ma, M. Wang, J. Zhou, Q. Su, X. Song, Q. Gan, Z. Zhang, and G. Karypis, "DistDGL: Distributed graph neural network training for billion-scale graphs," in *2020 IEEE/ACM 10th Workshop on Irregular Applications: Architectures and Algorithms (IA3)*. Los Alamitos, CA, USA: IEEE Computer Society, nov 2020, pp. 36–44. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/IA351965.2020.00011

[25] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[26] L. Ma, Z. Yang, Y. Miao, J. Xue, M. Wu, L. Zhou, and Y. Dai, "NeuGraph: Parallel deep neural network computation on large graphs," in *2019 USENIX Annual Technical Conference (USENIX ATC 19)*. Renton, WA: USENIX Association, Jul. 2019, pp. 443–458. [Online]. Available: https://www.usenix.org/conference/atc19/presentation/ma

[27] Z. Lin, C. Li, Y. Miao, Y. Liu, and Y. Xu, "PaGraph: Scaling gnn training on large graphs via computation-aware caching," in *Proceedings of the 11th ACM Symposium on Cloud Computing*, ser. SoCC '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 401–415. [Online]. Available: https://doi.org/10.1145/3419111.3421281

[28] M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, T. Xiao, T. He, G. Karypis, J. Li, and Z. Zhang, "Deep Graph Library: A graph-centric, highly-performant package for graph neural networks," *arXiv preprint arXiv:1909.01315*, 2019.

[29] P. Battaglia, J. B. C. Hamrick, V. Bapst, A. Sanchez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. E. Dahl, A. Vaswani, K. Allen, C. Nash, V. J. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu, "Relational inductive biases, deep learning, and graph networks," *arXiv*, 2018. [Online]. Available: https://arxiv.org/pdf/1806.01261.pdf

[30] Z. Tan, X. Yuan, C. He, M.-K. Sit, G. Li, X. Liu, B. Ai, K. Zeng, P. Pietzuch, and L. Mai, "Quiver: Supporting GPUs for low-latency, high-throughput GNN serving with workload awareness," *arXiv preprint arXiv:2305.10863*, 2023.

[31] H. Hu, F. Liu, Q. Pei, Y. Yuan, Z. Xu, and L. Wang, "λGrapher: A resource-efficient serverless system for gnn serving through graph sharing," in *The World Wide Web Conference*, ser. WWW '24. Association for Computing Machinery, 2024.

[32] L. Zeng, P. Huang, K. Luo, X. Zhang, Z. Zhou, and X. Chen, "Fograph: Enabling real-time deep graph inference with fog computing," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 1774–1784.

[33] D. Lin, S. Sun, J. Ding, X. Ke, H. Gu, X. Huang, C. Song, X. Zhang, L. Yi, J. Wen *et al.*, "Platogl: Effective and scalable deep graph learning system for graph-enhanced real-time recommendation," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3302–3311.

[34] M. Jin, Y. Liu, Y. Zheng, L. Chi, Y.-F. Li, and S. Pan, "ANEMONE: graph anomaly detection with multi-scale contrastive learning," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 3122–3126.

[35] O. Atkinson, A. Bhardwaj, C. Englert, V. S. Ngairangbam, and M. Spannowsky, "Anomaly detection with convolutional graph neural networks," *Journal of High Energy Physics*, vol. 2021, no. 8, pp. 1–19, 2021.

[36] Z. Wen and Y. Fang, "Augmenting low-resource text classification with graph-grounded pre-training and prompting," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 506–516. [Online]. Available: https://doi.org/10.1145/3539618.3591641

[37] E. Chien, W.-C. Chang, C.-J. Hsieh, H.-F. Yu, J. Zhang, O. Milenkovic, and I. S. Dhillon, "Node feature extraction by self-supervised multi-scale neighborhood prediction," in *International Conference on Learning Representations (ICLR)*, 2022.

[38] S. Virinchi, A. Saladi, and A. Mondal, "Recommending related products using graph neural networks in directed graphs," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, pp. 541–557.

[39] "Build a GNN-based real-time fraud detection solution using Amazon Sagemaker, Amazon Neptune, and the Deep Graph Library," https://aws.amazon.com/blogs/machine-learning/build-a-gnn-based-real-time-fraud-detection-solution-using-amazon-sagemaker-amazon-neptune-and-the-deep-graph-library/.

[40] M. Lu, Z. Han, S. X. Rao, Z. Zhang, Y. Zhao, Y. Shan, R. Raghunathan, C. Zhang, and J. Jiang, "Bright-graph neural networks in real-time fraud detection," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2022, pp. 3342–3351.

[41] A. Derrow-Pinion, J. She, D. Wong, O. Lange, T. Hester, L. Perez, M. Nunkesser, S. Lee, X. Guo, B. Wiltshire *et al.*, "Eta prediction with graph neural networks in google maps," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 3767–3776.

[42] T. Liu, Y. Chen, D. Li, C. Wu, Y. Zhu, J. He, Y. Peng, H. Chen, H. Chen, and C. Guo, "BGL: GPU-Efficient GNN training by optimizing graph data I/O and preprocessing," in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. Boston, MA: USENIX Association, Apr. 2023, pp. 103–118. [Online]. Available: https://www.usenix.org/conference/nsdi23/presentation/liu-tianfeng

[43] H. Zeng, H. Zhou, A. Srivastava, R. Kannan, and V. K. Prasanna, "GraphSAINT: Graph sampling based inductive learning method," in *8th International Conference on Learning Representations, ICLR, Addis Ababa, Ethiopia, April 26-30, 2020*. [Online]. Available: https://openreview.net/forum?id=BJe8pkHFwS

[44] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 1025–1035.

[45] J. Chen, T. Ma, and C. Xiao, "FastGCN: Fast learning with graph convolutional networks via importance sampling," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rytstxWAW

[46] M. Fey, J. E. Lenssen, F. Weichert, and J. Leskovec, "GNNAutoScale: Scalable and expressive graph neural networks via historical embeddings," in *International conference on machine learning*. PMLR, 2021, pp. 3294–3304.

[47] J. Peng, Z. Chen, Y. Shao, Y. Shen, L. Chen, and J. Cao, "Sancus: staleness-aware communication-avoiding full-graph decentralized training in large-scale graph neural networks," *Proceedings of the VLDB Endowment*, vol. 15, no. 9, pp. 1937–1950, 2022.

[48] K. Huang, H. Jiang, M. Wang, G. Xiao, D. Wipf, X. Song, Q. Gan, Z. Huang, J. Zhai, and Z. Zhang, "Freshgnn: Reducing memory access via stable historical embeddings for graph neural network training," *Proceedings of the VLDB Endowment*, vol. 17, no. 6, pp. 1473–1486, 2024. [Online]. Available: https://www.vldb.org/pvldb/vol17/p1473-huang.pdf

[49] W. Chen, S. He, H. Qu, and X. Zhang, "LeapGNN: Accelerating distributed GNN training leveraging Feature-Centric model migration," in *23rd USENIX Conference on File and Storage Technologies (FAST 25)*, 2025, pp. 255–270.

[50] J. Sun, Z. Shi, L. Su, W. Shen, Z. Wang, Y. Li, W. Yu, W. Lin, F. Wu, B. He *et al.*, "Helios: Efficient distributed dynamic graph sampling for online gnn inference," in *Proceedings of the 30th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming*, 2025, pp. 2–15.

[51] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *arXiv preprint arXiv:1709.05584*, 2017.

[52] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*. PMLR, 2017, pp. 1263–1272.

[53] H. Yang, "AliGraph: A comprehensive graph neural network platform," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 3165–3166. [Online]. Available: https://doi.org/10.1145/3292500.3340404

[54] D. Zheng, X. Song, C. Yang, D. LaSalle, and G. Karypis, "Distributed hybrid cpu and gpu training for graph neural networks on billion-scale heterogeneous graphs," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 4582–4591.

[55] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, "Deep interest network for click-through rate prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD, London, UK, August 19-23*, Y. Guo and F. Farooq, Eds. ACM, 2018, pp. 1059–1068. [Online]. Available: https://doi.org/10.1145/3219819.3219823

[56] S. Zhang, Y. Liu, Y. Sun, and N. Shah, "Graph-less neural networks: Teaching old mlps new tricks via distillation," in *The Tenth International Conference on Learning Representations, ICLR, Virtual Event, April 25-29*, 2022. [Online]. Available: https://openreview.net/forum?id=4p6_5HBWPCw

[57] J. Shi, V. Chaurasiya, Y. Liu, S. Vij, Y. Wu, S. Kanduri, N. Shah, P. Yu, N. Srivastava, L. Shi, G. Venkataraman, and J. Yu, "Embedding based retrieval in friend recommendation," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei*, H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, and B. Poblete, Eds. ACM, 2023, pp. 3330–3334. [Online]. Available: https://doi.org/10.1145/3539618.3591848

[58] J. Chen, J. Zhu, and L. Song, "Stochastic training of graph convolutional networks with variance reduction," in *International Conference on Machine Learning*, 2018, pp. 941–949.

[59] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[60] J. Yang, D. Tang, X. Song, L. Wang, Q. Yin, R. Chen, W. Yu, and J. Zhou, "GNNLab: a factored system for sample-based gnn training over gpus," in *Proceedings of the Seventeenth European Conference on Computer Systems*, 2022, pp. 417–434.

[61] J. Sun, L. Su, Z. Shi, W. Shen, Z. Wang, L. Wang, J. Zhang, Y. Li, W. Yu, J. Zhou, and F. Wu, "Legion: Automatically pushing the envelope of Multi-GPU system for Billion-Scale GNN training," in *2023 USENIX Annual Technical Conference (USENIX ATC 23)*. Boston, MA: USENIX Association, Jul. 2023, pp. 165–179. [Online]. Available: https://www.usenix.org/conference/atc23/presentation/sun

[62] D. Zhang, X. Huang, Z. Liu, J. Zhou, Z. Hu, X. Song, Z. Ge, L. Wang, Z. Zhang, and Y. Qi, "AGL: A scalable system for industrial-purpose graph machine learning," *Proc. VLDB Endow.*, vol. 13, no. 12, pp. 3125–3137, Aug. 2020. [Online]. Available: https://doi.org/10.14778/3415478.3415539

[63] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, "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*, ser. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 257–266. [Online]. Available: https://doi.org/10.1145/3292500.3330925

[64] G. Corso, L. Cavalleri, D. Beaini, P. Liò, and P. Veličković, "Principal neighbourhood aggregation for graph nets," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 260–13 271, 2020.

[65] G. Li, C. Xiong, A. Thabet, and B. Ghanem, "DeeperGCN: All you need to train deeper gcns," *arXiv preprint arXiv:2006.07739*, 2020.

[66] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" *arXiv preprint arXiv:2105.14491*, 2021.

[67] J. S. Bridle, "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing: Algorithms, architectures and applications*. Springer, 1990, pp. 227–236.

[68] K. Xu, C. Li, Y. Tian, T. Sonobe, K.-i. Kawarabayashi, and S. Jegelka, "Representation learning on graphs with jumping knowledge networks," in *International conference on machine learning*. PMLR, 2018, pp. 5453–5462.

[69] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[70] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[71] "Gloo: a collective communications library," https://github.com/facebookincubator/gloo.

[72] P. Damania, S. Li, A. Desmaison, A. Azzolini, B. Vaughan, E. Yang, G. Chanan, G. J. Chen, H. Jia, H. Huang, J. Spisak, L. Wehrstedt, L. Hosseini, M. Krishnan, O. Salpekar, P. Belevich, R. Varma, S. Gera, W. Liang, S. Xu, S. Chintala, C. He, A. Ziashahabi, S. Avestimehr, and Z. DeVito, "Pytorch RPC: Distributed Deep Learning Built on Tensor-Optimized Remote Procedure Calls," in *Proceedings of Machine Learning and Systems*, 2023.

[73] "NVIDIA NCCL," https://github.com/NVIDIA/nccl.

[74] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open Graph Benchmark: Datasets for machine learning on graphs," *Advances in neural information processing systems*, vol. 33, pp. 22 118–22 133, 2020.

[75] S. Edunov, D. Logothetis, C. Wang, A. Ching, and M. Kabiljo, "Generating synthetic social graphs with Darwini," in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, 2018, pp. 567–577.

[76] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *Proceedings of the 5th International Conference on Learning Representations*, ser. ICLR '17, 2017. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[77] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li, "Simple and deep graph convolutional networks," in *International conference on machine learning*. PMLR, 2020, pp. 1725–1735.

[78] G. Li, M. Muller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs go as deep as CNNs?" in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9267–9276.

[79] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[80] W. Hu, M. Fey, H. Ren, M. Nakata, Y. Dong, and J. Leskovec, "OGB-LSC: A large-scale challenge for machine learning on graphs," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, J. Vanschoren and S. Yeung, Eds., 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/db8e1af0cb3aca1ae2d0018624204529-Abstract-round2.html

[81] T. Kaler, N. Stathas, A. Ouyang, A.-S. Iliopoulos, T. Schardl, C. E. Leiserson, and J. Chen, "Accelerating training and inference of graph neural networks with fast sampling and pipelining," *Proceedings of Machine Learning and Systems*, vol. 4, pp. 172–189, 2022.

[82] G. Karypis and V. Kumar, "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM J. Sci. Comput.*, vol. 20, no. 1, p. 359–392, Dec. 1998.

[83] P. Naman and Y. Simmhan, "Ripple: Scalable incremental gnn inferencing on large streaming graphs," in *2025 IEEE 45th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2025, pp. 857–867.

[84] D. Wu, Z. Li, and T. Mitra, "Inkstream: Real-time gnn inference on streaming graphs via incremental update," *arXiv preprint arXiv:2309.11071*, 2023.

[85] Y. Wang, B. Feng, G. Li, S. Li, L. Deng, Y. Xie, and Y. Ding, "GNNAdvisor: An adaptive and efficient runtime system for GNN acceleration on GPUs," in *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*. USENIX Association, Jul. 2021, pp. 515–531. [Online]. Available: https://www.usenix.org/conference/osdi21/presentation/wang-yuke

[86] Y. Wang, B. Feng, Z. Wang, T. Geng, K. Barker, A. Li, and Y. Ding, "MGG: Accelerating graph neural networks with Fine-Grained Intra-Kernel Communication-Computation pipelining on Multi-GPU platforms," in *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*. Boston, MA: USENIX Association, Jul. 2023, pp. 779–795. [Online]. Available: https://www.usenix.org/conference/osdi23/presentation/wang-yuke

[87] Y. Wu, K. Ma, Z. Cai, T. Jin, B. Li, C. Zheng, J. Cheng, and F. Yu, "Seastar: vertex-centric programming for graph neural networks," in *Proceedings of the Sixteenth European Conference on Computer Systems*, 2021, pp. 359–375.

[88] Z. Jia, S. Lin, R. Ying, J. You, J. Leskovec, and A. Aiken, "Redundancy-free computation for graph neural networks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 997–1005. [Online]. Available: https://doi.org/10.1145/3394486.3403142

[89] Z. Jia, S. Lin, M. Gao, M. Zaharia, and A. Aiken, "Improving the accuracy, scalability, and performance of graph neural networks with ROC," in *Proceedings of Machine Learning and Systems*, I. Dhillon, D. Papailiopoulos, and V. Sze, Eds., vol. 2, 2020, pp. 187–198. [Online]. Available: https://proceedings.mlsys.org/paper/2020/file/fe9fc289c3ff0af142b6d3bead98a923-Paper.pdf

[90] C. Wan, Y. Li, A. Li, N. S. Kim, and Y. Lin, "BNS-GCN: Efficient full-graph training of graph convolutional networks with partition-parallelism and random boundary node sampling," *Proceedings of Machine Learning and Systems*, vol. 4, pp. 673–693, 2022.

[91] J. Thorpe, Y. Qiao, J. Eyolfson, S. Teng, G. Hu, Z. Jia, J. Wei, K. Vora, R. Netravali, M. Kim, and G. H. Xu, "Dorylus: Affordable, scalable, and accurate GNN training with distributed CPU servers and serverless threads," in *15th USENIX Symposium on Operating Systems Design and Implementation (OSDI 21)*. USENIX Association, Jul. 2021, pp. 495–514. [Online]. Available: https://www.usenix.org/conference/osdi21/presentation/thorpe

Appendix: Artifact Description

Artifact Description (AD)

I. Overview of Contributions and Artifacts

A. Paper's Main Contributions

This paper presents Reforge, a distributed GNN serving system that achieves low-latency inference by reusing and selectively recomputing precomputed node embeddings and by parallelizing computation graph construction across machines. In summary, our key contributions are as follows:

- C_1 **SRPE**: We propose Selective Recomputation of Precomputed Embeddings, which statistically minimizes accuracy loss by recomputing only a small, high-impact fraction of cached embeddings, dramatically reducing the communication volume required during GNN serving.
- C_2 **CGP**: We propose Computation Graph Parallelism, in which each machine builds a local partitioned computation graph using only locally available data and aggregates results via all-to-all collectives, avoiding centralized neighborhood fetching.
- C_3 **Reforge**: We implement and evaluate the full system on a four-machine GPU cluster across six datasets and three GNN models, achieving up to 159×/10.8× lower latency than DGL (Full)/DGL (NS) with less than 1% accuracy loss.

B. Computational Artifacts

A_1 is the sole artifact accompanying this paper and covers all three contributions listed above. The source code is not publicly available.

Artifact ID	Contributions Supported	Related Paper Elements
A_1	C_1 (SRPE)	Fig. 5, 8, 9, 10; Tables II, IV
	C_2 (CGP)	Figs. 8, 9, 11, 12
	C_3 (Reforge)	All figures/tables in §VIII

II. Artifact Identification

A. Computational Artifact A_1

Relation To Contributions

A_1 comprises the Reforge source code: a fork of Distributed DGL v1.0.2 extended with SRPE and CGP and packaged as the `reforge` Python module. In addition to the system implementation, the artifact ships dataset preparation scripts, model training scripts, multi-machine experiment launchers, which produce measuremetns (e.g., latency, throughput, and accuracy), that together reproduce every result reported in this paper.

Expected Results

Running A_1 should reproduce the following key results from §V: SRPE cuts communication volume by up to 18× and per-request latency by up to 29× over DGL (Full) by selectively recomputing only a small, high-impact fraction of cached embeddings. Adding CGP yields a further latency reduction of up to 5.5× by replacing centralized neighborhood fetching with distributed computation graph construction. Together, the full Reforge system reaches up to 159× and 10.8× lower latency than DGL (Full) and DGL (NS), respectively, at less than 1% accuracy loss across all six datasets and three GNN models.

Expected Reproduction Time (in Minutes)

Setup. Installing Reforge and its Python dependencies takes a few minutes.

Execution. Reproducing the paper's results involves four tasks described in detail below. Downloading and partitioning all six graph datasets across the cluster (T1) takes approximately 720 minutes. Training all GNN model and dataset combinations (T2) takes approximately 1440 minutes. Evaluating model accuracy and SRPE recomputation policies (T3) takes approximately 120 minutes. Running all serving performance experiments (T4) takes approximately 600 minutes. DGL (Full) on large graphs such as FB10B may be omitted when GPU memory is insufficient.

Artifact Setup (incl. Inputs)

Hardware: Four machines, each equipped with 2× NVIDIA V100S 32 GB GPUs, 2× AMD EPYC 7542 CPUs, and 512 GB host RAM, connected via 25 Gbps Ethernet. All machines must be reachable by passwordless SSH as `node0`–`node3` from `node0`.

Software: Tested on Ubuntu 22.04 with CUDA 11.8. Required libraries:

- PyTorch, 1.13+, https://pytorch.org
- Reforge/DGL, 1.0.2, built from A_1

Datasets / Inputs: Five standard datasets (Reddit, Yelp, Amazon, OGB-Products, OGB-Papers100M) are downloaded automatically via OGB during graph partitioning. FB10B [1] is a publicly available 10-billion-edge synthetic graph. Each machine requires at least 1 TB of disk space.

Installation and Deployment: After installing PyTorch, configure the following environment variables on all machines:

```
export DGL_HOME=$PWD
export DGL_LIBRARY_PATH=$DGL_HOME/build
export PYTHONPATH=$PYTHONPATH:$DGL_HOME/python
export DGL_DATA_HOME=<path with >=1 TB free>
export DGL_IFNAME=<network interface>
```

Then compile the C++ backend and install the `reforge` Python package:

```
mkdir build && cd build
```

```
cmake .. && make -j
cd ../python
python setup.py build_ext --inplace
```

Artifact Execution

Reproducing the paper's results requires executing four tasks in sequence from `node0`, each building on the outputs of the previous. We provide a driver script for each task; all scripts accept an output directory and coordinate the four-machine cluster via passwordless SSH.

T1: Dataset Preprocessing (`preprocess_data.py`). This task downloads all six graph datasets, partitions each graph into per-machine shards using distributed graph partitioning, and distributes the resulting partitions to all cluster nodes. After partitioning, it extracts query nodes from each dataset's test set (25% of test nodes held out as described in §V) and generates serving request batches of 1,024 query nodes each.

T2: Model Training (`train_models.py`). This task trains all GNN model and dataset combinations (3 models $\times$ 6 datasets) by sweeping hyperparameters over learning rates and dropout probabilities. Each combination is trained until validation loss no longer improves. The script selects the checkpoint with the highest test accuracy per combination, producing trained model weights and precomputed embeddings (PEs) used in subsequent tasks.

T3: Model Accuracy and SRPE Policy Evaluation (`evaluate_accuracy.py`). Using the trained models and PEs from T2, this task measures serving accuracy under varying recomputation budgets (γ). It evaluates the proposed top-query-edges-ratio heuristic, importance-score-based, and random recomputation policies. The results reproduce Table II (accuracy comparison between Reforge and DGL (NS)), Table IV (per-dataset/model recomputation budgets), and Fig. 10 (recomputation budget vs. latency/accuracy tradeoff).

T4: Serving Performance Measurement (`run_serving.py`). This task provides a distributed launcher that spawns processes across the cluster running Reforge and both baseline systems (DGL (Full) and DGL (NS)). Pre-generated serving requests from T1 are fed to each system to measure per-request latency, per-stage latency breakdown (Fetch / Copy / GPU), communication volume, and throughput under varying request rates. Experiments cover all dataset and model combinations, varying numbers of GPUs (2, 4, 8) and GPU feature cache sizes. The results reproduce Fig. 8 (end-to-end latency), Fig. 9 (latency breakdown and communication sizes), Fig. 11 (scalability with GPUs), Fig. 12 (latency-throughput), and Fig. 13 (impact of GPU feature caching).

REFERENCES

[1] S. Edunov, D. Logothetis, C. Wang, A. Ching, and M. Kabiljo, "Generating synthetic social graphs with Darwini," in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, 2018, pp. 567–577.