SHORT-PAPER

Scaling Graph Neural Networks (GNN) for Real-Time Modeling of Network Behaviour

MARYAM ASGARI, University of Technology Sydney, Sydney, NSW, Australia

RENPING LIU, University of Technology Sydney, Sydney, NSW, Australia

RAYMOND OWEN, University of Technology Sydney, Sydney, NSW, Australia

TANZEELA ALTAF, University of Technology Sydney, Sydney, NSW, Australia

Published: 08 September 2025

Citation in BibTeX format

SIGCOMM '25: ACM SIGCOMM 2025 Conference
September 8 - 11, 2025
Coimbra, Portugal

Conference Sponsors:
SIGCOMM

ACM SIGCOMM Posters and Demos '25: Proceedings of the ACM SIGCOMM 2025 Posters and Demos (September 2025)
https://doi.org/10.1145/3744969.3748415
ISBN: 9798400720260

POSTER: Scaling Graph Neural Networks (GNN) for Real-Time Modeling of Network Behaviour

Maryam Asgari
University of Technology Sydney
Maryam.Asgari@student.uts.edu.au

Ren Ping Liu
University of Technology Sydney
RenPing.Liu@uts.edu.au

Raymond Owen
University of Technology Sydney
Ray.Owen@uts.edu.au

Tanzeela Altaf
University of Technology Sydney
Tanzeela.Altaf@uts.edu.au

Abstract

Network modeling has long been a well-established field of study. More recently, Graph Neural Network based models have demonstrated remarkable capability in capturing complex interactions in network data without assumptions about physical networks. While this characteristic facilitates integration across various telecom access networks, current benchmark models remain impractical for real-world deployment, due to real-time demands of modern infrastructure.

This research develops a scalable solution for network modeling in large-scale domains such as telecommunication networks. By incorporating distributed learning into the architecture, we propose a novel framework that addresses computational inefficiency without compromising the accuracy offered by benchmark GNN-based models. The proposed architecture supports deeper and larger graphs, and natively handles fragmented datasets, reducing reliance on centralized aggregation and improving compatibility with real-world infrastructure. Beyond scalability, the design emphasizes stable optimization and resilience to enhance reliability in production environments. When applied to the state-of-the-art model, our proposed architecture outperforms the original, achieving a Pearson correlation of 0.999 with MSE under 0.0005. It also converges faster, with inference speedup scaling proportionally to the number of nodes. In a single-node, two-worker setup, it achieves ~48% inference speedup, with overall training efficiency improving by 20%, highlighting practical benefits for real-world scenarios.

CCS Concepts

• **Computing methodologies → Machine learning**; • **Networks → Network performance evaluation**.

Keywords

Broadband Telecommunications, Network Modeling, Graph Neural Network, Distributed Learning, Data Parallelism, Layer-wise Optimisation

SIGCOMM Posters and Demos '25, Coimbra, Portugal

ACM ISBN 979-8-4007-2026-0/2025/09
https://doi.org/10.1145/3744969.3748415

ACM Reference Format:
Maryam Asgari, Ren Ping Liu, Raymond Owen, and Tanzeela Altaf. 2025. POSTER: Scaling Graph Neural Networks (GNN) for Real-Time Modeling of Network Behaviour. In *ACM SIGCOMM 2025 Conference (SIGCOMM Posters and Demos '25), September 8–11, 2025, Coimbra, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3744969.3748415

1 Introduction

Network modeling is arguably one of the most critical components in enabling autonomous networks. Without accurate, real-time models of network behaviour, intelligent decision-making and proactive management become infeasible. As networks expand and traffic patterns shift rapidly, traditional rule-based approaches—whether human-driven decisions or classical simulators—are no longer sufficient. Queuing-theoretic models remain computationally intensive at scale, while digital twins often lack awareness of complex topologies. Meanwhile, cutting-edge academic models [4] exemplified by RouteNet [9, 11] and its successors [2, 3, 8], demonstrate strong performance in controlled environments, but a persistent gap remains between research prototypes and deployable systems in real-world telecom settings. This gap stems from computational inefficiencies, scalability bottlenecks, lack of robustness under evolving network conditions, and the inherent difficulty of integrating machine learning into legacy, heterogeneous network infrastructures [7].

To enable real-world deployment, key limitations must be addressed: in-memory designs that restrict scalability [6], computational overhead from single-threaded execution, and an inability to handle decentralized data sources. Working with telecom network data [12] also presents distinct challenges. This data is typically high-volume, transient, and often streamed from OSS systems, telemetry platforms, or vendor-specific APIs. Due to privacy, cost, and storage constraints, raw traffic data is rarely archived long-term, requiring models to learn from time-buffered telemetry. In addition, telecom networks themselves are highly dynamic: infrastructure evolves, routing policies change, and regional traffic patterns shift frequently. These factors create a fundamental requirement: any model intended for real-world telecom deployment must support frequent retraining to maintain predictive accuracy. Equally important is preserving end-to-end topology awareness to ensure coordinated network-wide decision-making. This brings us to a critical architectural question: how do we design a learning framework that can adapt effectively to these dynamics?

2 System Design

The design centers on a synchronous data-parallel strategy [1, 5], where multiple model replicas are instantiated across available computational units. Each replica applies the forward pass independently using a shared core GNN that encodes the model's predictive logic. An adapted regularization is applied during this stage to support more stable optimization. Synchronization maintains variable consistency, enabling full topology awareness and coordinated learning. This parallelism boosts training throughput, reduces inference latency, and improves resource efficiency — delivering capabilities not attainable through batch scaling alone. A custom data pipeline supports fragmented, transient telemetry by ingesting decentralized, regionally scoped datasets without prior aggregation. Parallel loading, prefetching, and dynamic tuning ensure high-throughput ingestion under diverse infrastructure constraints. Since custom GNN-based models often exhibit heterogeneous layer structures, the architecture includes layer-wise optimization, allowing each layer to adapt independently. This mitigates instability caused by global optimization under distributed conditions, especially when network dynamics shift due to traffic bursts or data noise. The workflow is illustrated below:

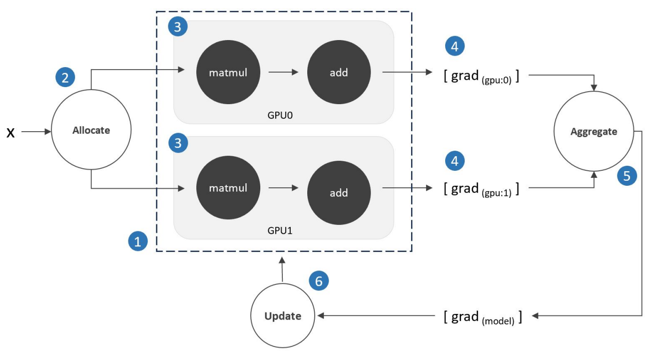

Figure 1: Synchronous data-parallel training

(1) The model is replicated across all processing units. (2) Input is allocated via the custom data pipeline, enabling dataset partitioning or per-replica assignment. (3) Each replica performs a forward pass and computes the loss, incorporating an adapted regularization. (4) Gradients are computed independently on each replica. (5) Gradients are aggregated and used to update model parameters. (6) A layer-wise optimizer applies updates, keeping all replicas synchronized.

3 Experiment Design and Results

To evaluate effectiveness, we applied the framework to RouteNet - a state-of-the-art GNN model for SDN - and observed measurable improvements over the benchmark. While the experiment preserves RouteNet's predictive core, the framework itself is designed to generalize to other GNN-based models with similar constraints. Training and evaluation were conducted on a single RHEL 8 machine with two Nvidia L4 GPUs, modeling per-path delay using the 14-node NSF network. Each dataset sample represents an independent source-destination traffic scenario, enabling a data-parallel approach where batches are distributed across replicas. RouteNet was trained under identical conditions to ensure a fair performance baseline. Both models are trained for 50k steps and evaluated every 1k to observe model behaviour comprehensively. A summary of the training run is presented in the following table, with source code available in [10]:

Metric	RouteNet	Proposed Framework
Lowest MSE	0.0013	0.0001
Highest ρ	0.993	0.999
Training step/sec	$\sim$ 2.13	$\sim$ 2.05
Inference Time	$\sim$ 356 sec	$\sim$ 184 sec
Convergence Step	14000	7000
Time to Convergence	3h22m	1h18m
MSE StdDev	0.00251	0.00155
Wall Time (50k steps)	11:42:12	9:20:27

Table 1: Performance summary

Although the distributed framework introduces a modest synchronization overhead of ~4% (reflected in reduced training step/sec), it delivers substantial performance gains — including ~48% faster inference, ~62% faster convergence, and a ~92% reduction in MSE. As shown in the figures below, the proposed model (blue curve) reaches stability at 7k steps (1h18m) versus 14k steps (3h22m) for the original (black curve), achieving a lower MSE floor and consistently higher ρ, contributing to improved reliability under fluctuating network conditions. It also exhibits fewer oscillations throughout training (lower MSE StdDev), reducing training instability by 38% and maintaining smoother, more stable convergence, demonstrating resilience in a production-like environment.

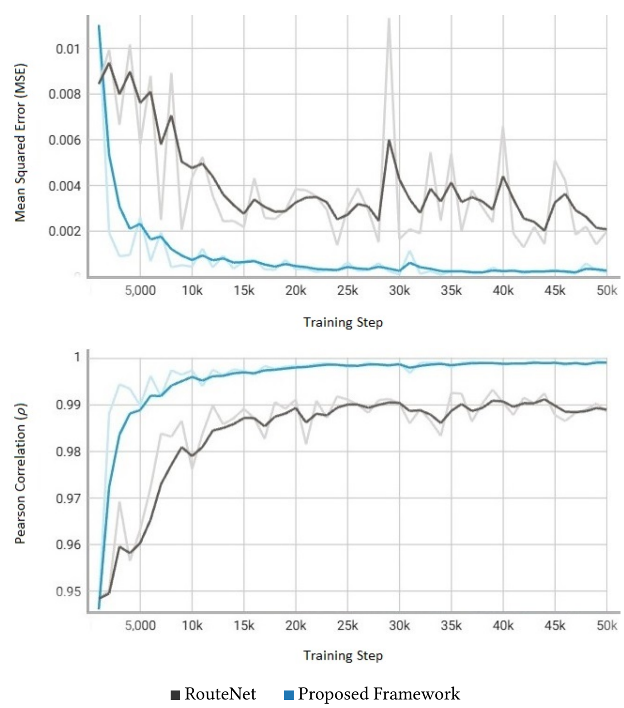

Figure 2: Performance metrics (MSE, ρ) comparison

References

[1] A. A. Awan, J. Bedorf, C.-H. Chu, H. Subramoni, and D. K. Panda. 2019. Scalable Distributed DNN Training using TensorFlow and CUDA-Aware MPI: Characterization, Designs, and Performance Evaluation. In *2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*. 498–507. doi:10.1109/CCGRID.2019.00064

[2] M. Ferriol-Galmes, J. Paillisse, J. Suarez-Varela, K. Rusek, S. Xiao, X. Shi, X. Cheng, P. Barlet-Ros, and A. Cabellos-Aparicio. 2023. RouteNet-Fermi: Network Modeling With Graph Neural Networks. *IEEE/ACM Transactions on Networking* 31, 6 (2023), 1–0. doi:10.1109/TNET.2023.3269983

[3] M. Ferriol-Galmes, K. Rusek, J. Suarez-Varela, S. Xiao, X. Shi, X. Cheng, B. Wu, P. Barlet-Ros, and A. Cabellos-Aparicio. 2022. RouteNet-Erlang: A Graph Neural Network for Network Performance Evaluation. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*. 2018–2027. doi:10.1109/INFOCOM48880.2022.9796944

[4] W. Jiang. 2022. Graph-based deep learning for communication networks: A survey. *Computer Communications* 185 (2022), 40–54. doi:10.1016/j.comcom.2021.12.015

[5] H. Lin, M. Yan, X. Ye, D. Fan, S. Pan, W. Chen, and Y. Xie. 2023. A Comprehensive Survey on Distributed Training of Graph Neural Networks. *Proc. IEEE* 111, 12 (2023), 1572–1606. doi:10.1109/JPROC.2023.3337442

[6] M. Liu, H. Gao, and S. Ji. 2020. Towards Deeper Graph Neural Networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*. 338–348. doi:10.1145/3394486.3403076

[7] A. R. Munappy, J. Bosch, H. H. Olsson, A. Arpteg, and B. Brinne. 2022. Data management for production quality deep learning models: Challenges and solutions. *The Journal of Systems and Software* 191 (2022), 111359. doi:10.1016/j.jss.2022.111359

[8] K. Rusek, J. Suarez-Varela, P. Almasan, P. Barlet-Ros, and A. Cabellos-Aparicio. 2020. RouteNet: Leveraging Graph Neural Networks for Network Modeling and Optimization in SDN. *IEEE Journal on Selected Areas in Communications* 38, 10 (2020), 2260–2270. doi:10.1109/JSAC.2020.3000405

[9] K. Rusek, J. Suárez-Varela, A. Mestres, P. Barlet-Ros, and A. Cabellos-Aparicio. 2019. Unveiling the potential of Graph Neural Networks for network modeling and optimization in SDN. In *Proceedings of the 2019 ACM Symposium on SDN Research*. 140–151. doi:10.1145/3314148.3314357

[10] Source Code: https://github.com/0009-0007-1725-7979/RN0-DL-LW 2025.

[11] J. Suárez-Varela, S. Carol-Bosch, K. Rusek, P. Almasan, M. Arias, P. Barlet-Ros, and A. Cabellos-Aparicio. 2019. Challenging the generalization capabilities of Graph Neural Networks for network modeling. In *Proceedings of the ACM SIGCOMM 2019 Conference Posters and Demos*. 114–115. doi:10.1145/3342280.3342327

[12] P. Tam, I. Song, S. Kang, S. Ros, and S. Kim. 2022. Graph Neural Networks for Intelligent Modelling in Network Management and Orchestration: A Survey on Communications. *Electronics (Basel)* 11, 20 (2022), 3371. doi:10.3390/electronics11203371