Exploring Machine Learning Models with Spatial-Temporal Information for Interconnect Network Traffic Forecasting

Xiongxiao Xu
xxu85@hawk.iit.edu
Illinois Institute of Technology
Chicago, Illinois, United States

ABSTRACT

Interconnect networks are an essential component of high-performance computing (HPC) systems. To study large-scale networking systems, parallel discrete event simulation (PDES) has been widely used to simulate real-world HPC behaviors. However, PDES simulation requirements and computational complexity are increasing rapidly, making it challenging to achieve accurate results. Therefore, researchers have been exploring a surrogate-ready PDES framework that utilizes machine learning-based surrogate models to accelerate PDES. In this paper, we present our vision and initial step to leverage machine learning models to utilize spatial-temporal information to forecast interconnect network traffic. The preliminary results show that it is promising to explore machine learning models for interconnect network traffic forecasting.

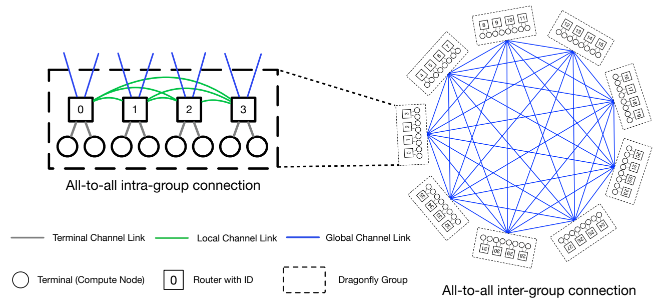

Figure 1: The illustration of the 1D Dragonfly network.

CCS CONCEPTS

• **Computing methodologies → Discrete-event simulation**.

ACM Reference Format:
Xiongxiao Xu. 2023. Exploring Machine Learning Models with Spatial-Temporal Information for Interconnect Network Traffic Forecasting. In *ACM SIGSIM Conference on Principles of Advanced Discrete Simulation (SIGSIM-PADS '23), June 21–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3573900.3593635

1 INTRODUCTION

Efficient and scalable interconnect networks are critical for high-performance computing (HPC) systems to support unprecedented system sizes at a reasonable cost. The Dragonfly network is an example of a hierarchical, high-radix and low-diameter topology that offers high-bandwidth and low-latency service while reducing network costs [4, 5]. The network topology is as shown in Figure 1. This unique network topology has been widely adopted by various HPC facilities, including the National Energy Research Scientific Computing Center and the Argonne Leadership Computing Facility.

Parallel discrete event simulation (PDES) has been successfully applied to model network data flows and hierarchical storage systems in various applications, including science enterprise design and provisioning, transportation and mobility, internet and cyber-security simulations, materials science, and hardware co-design [9]. While PDES modeling frameworks such as ROSS [1] and CODES [8] can simulate Dragonfly networks, their high computational complexity limits practical deployment. For example, simulating the 4,096-node MILC for 122 milliseconds using PDES takes up to four hours [11].

In order to tackle this issue, there is a high demand for high-fidelity surrogate models that have the potential to replace billions or even trillions of PDES events. However, developing effective surrogate models poses several challenges. Firstly, when simulating multiple HPC applications on a large-scale interconnect network, the fierce contention among the applications for shared resources may have a negative impact on the forecasting accuracy of surrogate models [6, 12]. Secondly, interconnect networks, e.g., dragonfly network, exhibit a unique network topology, making it daunting to capture the correlation between different ports in the unique topology to improve the forecasting performance. Thirdly, PDES events are generated at a fast pace, and surrogate models need to ensure prediction efficiency while maintaining effectiveness.

2 PRELIMINARY RESULTS

To capture the complex network traffic, we investigate a deep learning-based model, LSTM (Long Short-Term Memory), due its wide adoption in a variety of traffic prediction scenarios such as road traffic forecasting [2, 7, 13]. LSTM is developed based on the LSTM [3] framework. LSTM is shown to be effective in addressing the gradient vanishing problem and capable of capturing long-term dependency [10]. The forecasting result of LSTM on a port of a router is shown in Figure 2. The LSTM is trained on the training data (the blue line) and tested on the test data (the orange line). We can clearly observe that the forecast data (the green line) is close to the test data (the orange line). It indicates the LSTM can accurately forecast the network traffic.

SIGSIM-PADS '23, June 21–23, 2023, Orlando, FL, USA

ACM ISBN 979-8-4007-0030-9/23/06.
https://doi.org/10.1145/3573900.3593635

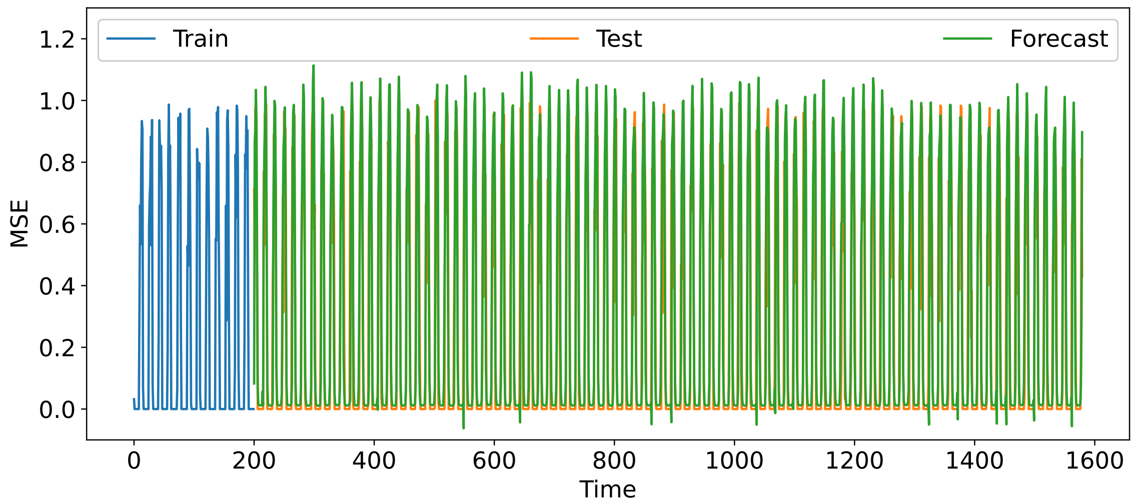

Figure 2: The forecasting result of the LSTM.

3 THE INTEGRATION OF SURROGATE MODEL AND PDES

We envision a surrogate-ready PDES where the simulation alternates between two phases: a detailed PDES phase where the application workload is simulated fully and a surrogate phase where the induced traffic is forecasted to fast-forward PDES through the time period forecasted by the surrogate.

4 CONCLUSION AND FUTURE WORK

In this paper, we explore the possibility of spatial-temporal machine learning models as surrogate models to leverage the PDES modeling, by forecasting the intricate traffic in the Dragonfly network. Our preliminary analysis demonstrates the the potential of LSTM as a surrogate model. However, our current version of LSTM does not utilize the spatial information of the interconnect network and further investigation is needed to determine the combination of forecasted results from surrogate models and PDES.

For future work, we plan to incorporate more network features (e.g., bandwidth consumed and busy time) and consider the spatial information (e.g., correlation between different ports within the same router, group or network) of the interconnect network for machine learning models. Furthermore, exploring the potential of the advanced machine learning models in interconnect network traffic forecasting, e.g., transformers and graph neural networks (GNNs), is also a promising research direction.

REFERENCES

[1] Christopher D Carothers, David Bauer, and Shawn Pearce. 2002. ROSS: A high-performance, low-memory, modular Time Warp system. *J. Parallel and Distrib. Comput.* 62, 11 (2002), 1648–1669.

[2] Xingyi Cheng, Ruiqing Zhang, Jie Zhou, and Wei Xu. 2018. Deeptransport: Learning spatial-temporal dependency for traffic condition forecasting. In *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[3] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[4] Yao Kang, Xin Wang, and Zhiling Lan. 2022. Study of Workload Interference with Intelligent Routing on Dragonfly. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (Dallas, Texas) *(SC '22)*. Article 20, 14 pages.

[5] John Kim, Wiliam J Dally, Steve Scott, and Dennis Abts. 2008. Technology-driven, highly-scalable dragonfly topology. *ACM SIGARCH Computer Architecture News* 36, 3 (2008), 77–88.

[6] Boyang Li, Sudheer Chunduri, Kevin Harms, Yuping Fan, and Zhiling Lan. 2019. The effect of system utilization on application performance variability. In *Proceedings of the 9th International Workshop on Runtime and Operating Systems for Supercomputers*. 11–18.

[7] Xiaolei Ma, Zhimin Tao, Yinhai Wang, Haiyang Yu, and Yunpeng Wang. 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies* 54 (2015), 187–197.

[8] Misbah Mubarak, Christopher D Carothers, Robert B Ross, and Philip Carns. 2017. Enabling parallel simulation of large-scale HPC network systems. *IEEE Transactions on Parallel and Distributed Systems* 28, 1 (2017), 87–100.

[9] Kalyan Perumalla, Maximilian Bremer, Kevin Brown, Cy Chan, Stephan Eidenbenz, K Scott Hemmert, Adolfy Hoisie, Benjamin Newton, James Nutaro, Tomas Oppelstrup, et al. 2022. *Computer Science Research Needs for Parallel Discrete Event Simulation (PDES)*. Technical Report. Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).

[10] Igor V Tetko, David J Livingstone, and Alexander I Luik. 1995. Neural network studies. 1. Comparison of overfitting and overtraining. *Journal of chemical information and computer sciences* 35, 5 (1995), 826–833.

[11] Xin Wang, Misbah Mubarak, Yao Kang, Robert B Ross, and Zhiling Lan. 2020. Union: An automatic workload manager for accelerating network simulation. In *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 821–830.

[12] Xin Wang, Misbah Mubarak, Xu Yang, Robert B Ross, and Zhiling Lan. 2018. Trade-off study of localizing communication and balancing network traffic on a dragonfly system. In *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 1113–1122.

[13] Huaxiu Yao, Xianfeng Tang, Hua Wei, Guanjie Zheng, and Zhenhui Li. 2019. Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5668–5675.