Transformer-based Predictions for Sudden Network Changes

UCHICAGO

Siddhant Ray, Xi Jiang, Zhuohan Gu, Junchen Jiang, Nick Feamster

Introduction

Motivation

- Inter-packet interactions lead to sudden changes in the network state. ***E.g., Newly starting flows cause a sudden spike in latency***
- Packet interactions are not arbitrary but exhibit recurring patterns. ***E.g., Repeated filling of buffers in a video session***
- Predicting sudden changes enables change in forwarding decisions. ***E.g., Optimized queuing for lower packet loss***
- Transformers can learn time series trend predictions using *attention*. ***Train them for sudden change prediction.***

Challenges

- Network traces are long and sudden changes in latency are rare ***E.g., <10 times in 10000s of packets***
- Predicting such rare events via a learnt model is ***hard.***
- Heuristic-based prediction models fail on tail cases by only learning ***averaged smoothened past values***
- General Transformers scale quadratically ***($O(n^2)$)*** with sequence size and need to be adapted to learn on long network traces.

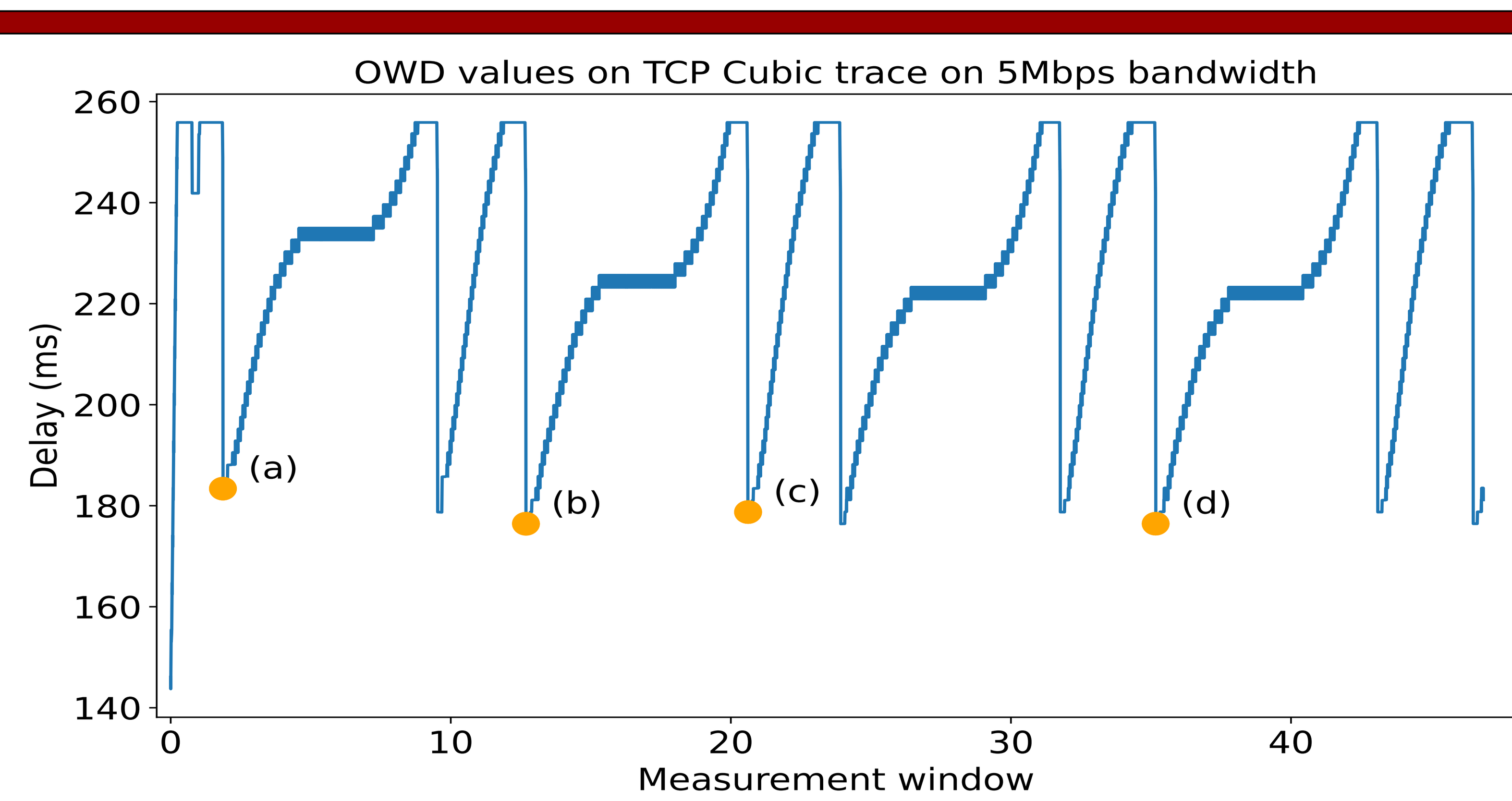
Design and Insights

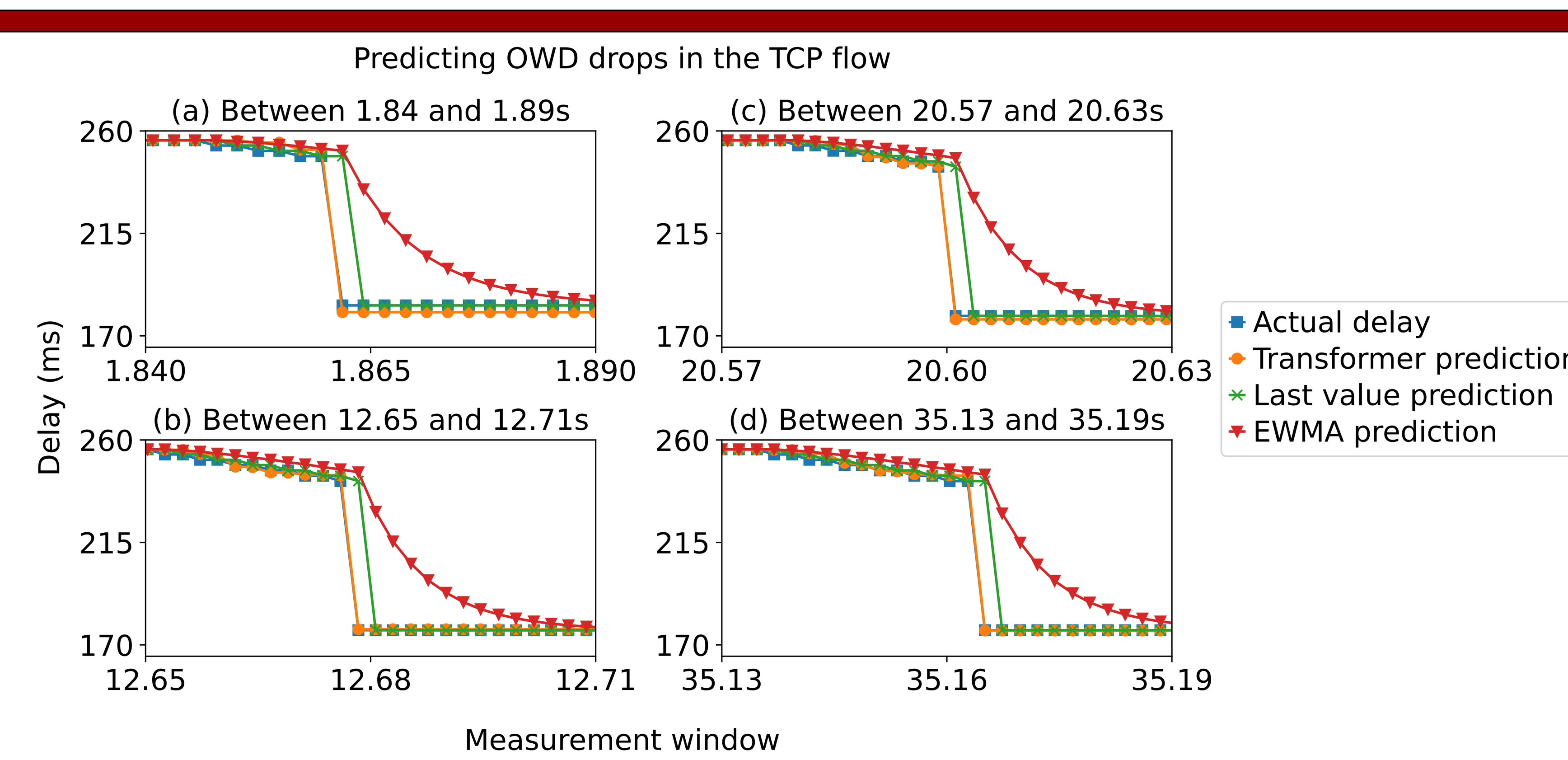
- Train Transformer to predict ***One Way Delay (OWD)*** for the next packet, based on the sequence of values from previous packets
- Features used : `relative_timestamp`, `packet_size`, `one_way_delay`
- Loss function penalizes sudden changes in ***OWD*** by a large factor α by tracking changes in value between packets
- Fixed lookback window size for training: matches ***queue size*** as packets queued together capture interactions much better

$$Loss\,(p(x), t(x)) = \begin{cases} \alpha\big(t(x) - p(x)\big)^2, & if\ \ t(x-1) - t(x) \geq \delta \\ \big(t(x) - p(x)\big)^2, & else \end{cases}$$

- $t(x)$ = *OWD* value, $t(x-1)$ = previous *OWD* value, $p(x)$ = predicted value
- $\alpha = 100$ and $\delta = 60$ ms are workload based hyper-parameters
- Training over different workloads will help generalization

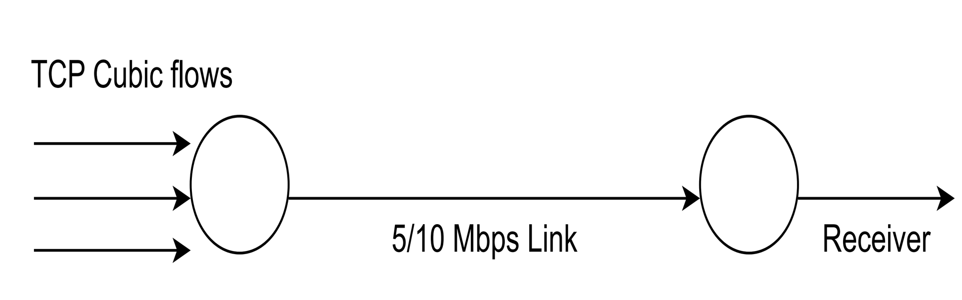
Preliminary Evaluation





Evaluation Details

- NS-3 simulation trace: TCP Cubic
- Trained on ***10 Mbps*** link bandwidth
- Tested on ***5 Mbps*** link bandwidth
- One Titan TRX GPU
- Training time ~***15 min***
- Inference time ~***1 ms***



Future Work

- Evaluate the model on more real-world workloads
- Applications of sudden change predictions. E.g., multi-level queue management, TCP multipath
- Generalize the model to predict application-level changes. E.g., user behaviour prediction, prediction of frame-skipping in video streaming

References

1. Alexander Dietmüller, Siddhant Ray, Romain Jacob, and Laurent Vanbever. A new hope for network model generalization. In Proceedings of the 21st ACM Workshop on Hot Topics in Networks, HotNets '22, page 152–159, New York, NY, USA, 2022.
2. Satyandra Guthula, Navya Battula, Roman Beltiukov, Wenbo Guo, and Arpit Gupta. netfound: Foundation model for network security, 2023