



HAL
open science

Low Complexity Approaches for End-to-End Latency Prediction

Pierre Larrenie, Jean-François Bercher, Olivier Venard, Iyad Lahsen-Cherif

► **To cite this version:**

Pierre Larrenie, Jean-François Bercher, Olivier Venard, Iyad Lahsen-Cherif. Low Complexity Approaches for End-to-End Latency Prediction. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Oct 2022, Kharagpur, France. pp.1-6, 10.1109/ICCCNT54827.2022.9984543 . hal-03957811

HAL Id: hal-03957811

<https://hal.science/hal-03957811>

Submitted on 30 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LOW COMPLEXITY APPROACHES FOR END-TO-END LATENCY PREDICTION

Pierre Larrenie

Thales SIX & LIGM
Université Gustave Eiffel, CNRS
Marne-la-Vallée, France
pierre.larrenie@esiee.fr

Jean-François Bercher

LIGM
Université Gustave Eiffel, CNRS
Marne-la-Vallée, France
jean-francois.bercher@esiee.fr

Olivier Venard

ESYCOM
Université Gustave Eiffel, CNRS
Marne-la-Vallée, France
olivier.venard@esiee.fr

Iyad Lahsen-Cherif

Institut National des Postes et Télécommunications (INPT)
Rabat, Morocco
lahsencherif@inpt.ac.ma

ABSTRACT

Software Defined Networks have opened the door to statistical and AI-based techniques to improve efficiency of networking. Especially to ensure a certain *Quality of Service* (QoS) for specific applications by routing packets with awareness on content nature (VoIP, video, files, etc.) and its needs (latency, bandwidth, etc.) to use efficiently resources of a network. Predicting various Key Performance Indicators (KPIs) at any level may handle such problems while preserving network bandwidth.

The question addressed in this work is the design of efficient and low-cost algorithms for KPI prediction, implementable at the local level. We focus on end-to-end latency prediction, for which we illustrate our approaches and results on a public dataset from the recent international challenge on GNN [1]. We propose several low complexity, locally implementable approaches, achieving significantly lower wall time both for training and inference, with marginally worse prediction accuracy compared to state-of-the-art global GNN solutions.

Keywords KPI Prediction · Machine Learning · General Regression · SDN · Networking · Queuing Theory · GNN

1 Introduction

Routing while ensuring Quality of Service (QoS) is still a great challenge in any networks. Having powerful ways to transmit data is not sufficient, we must use resources wisely. This is true for wide static networks but even more for mobile networks with dynamic topology.

The emergence of Software-Defined Networking (SDN) [2, 3] has made it possible to share data more efficiently between communication layers. Services are able to provide network requirements to routers based on their nature; routers acquire data about network performance, and finally allocate resources to meet these requirements. However, acquiring overall network performance can result in high consumption of network bandwidth for signaling; that is particularly constraining for networks with limited resources like Mobile *Ad-Hoc* Networks (MANET).

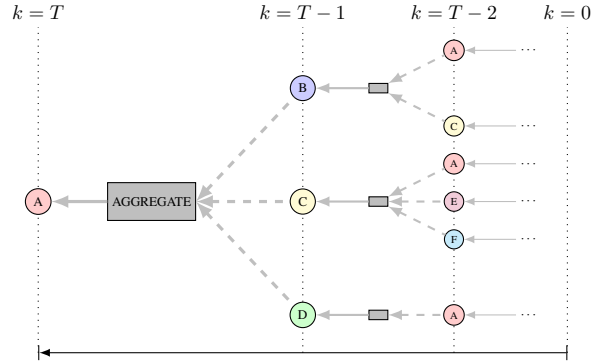


Figure 1: GNN repeating T Message Passing mechanisms: message propagation and aggregation (*inspired from [7]*)

We consider network for which we wish to reduce the amount of signalization and perform intelligent routing. In order to limit signalization, a first axis is to be able to estimate some key performance indicators (KPI) from other KPIs. A second point would be to be able to perform this prediction locally, at the node level, rather than a global estimation of the network. Finally, if predictions are to be performed locally, the complexity of the algorithms will need to be low, but still preserve good prediction quality. The question we address is thus the design of efficient and low-cost algorithms for KPI prediction, implementable at the local level. We focus on end-to-end latency prediction, for which we illustrate our approaches and results on a public dataset from the recent international challenge [1].

The best performances of the state-of-the-art are obtained with Graph Neural Networks (GNNs) [4, 5, 1]. Although this is a global method while we favor local methods, we use these performances as a benchmark. We first propose to use standard machine learning regression methods, for which we show that a careful feature engineering and feature selection (based on queue theory and the approach in [6]) allows to obtain near state-of-the-art performances with a very low number of parameters and very low computational cost, with the ability to operate at the link level instead of a whole-graph level. Building on that, we show that it is even possible to obtain similar performances with a single feature and curve-fitting methods.

The presentation is structured as follows. In section 2, we first recall the key concepts on GNNs and queues; present some related works in the literature, before introducing the dataset used for the validation of our proposals. In section 3, we present the different approaches proposed, starting with the choice of features for machine learning methods, followed by general curve fitting methods. We then compare in section 4 the performances of these different approaches, in terms of performance as well as in terms of learning time and inference time. Finally, we conclude, discuss the overall results and draw some perspectives.

2 Related work and dataset

2.1 Graph Neural Networks (GNNs)

GNN [7, 8] is a machine learning paradigm that handles non-euclidean data: graphs. A graph is defined as a set of nodes and edges with some properties on its nodes and its edges. The key point in GNNs is the concept of Message Passing: each node of the graph will update its state according to states of its neighborhood by sending and receiving *messages* transmitted along edges. By repeating this mechanism T times, a node is able to capture states of its T -hop neighborhood as shown in Figure 1.

2.2 Queue Theory

Queue Theory is a well studied domain and for most of simple queue systems, explicit equations exist [9]. Further, we will refer to queue systems by using their Kendall's notation. We often take at reference $M/M/1$ and $M/M/1/K$ for their markovian property, since equations are particularly easy to handle in this case.

However, for more general queue systems such as $M/G/1$ and $M/G/1/K$, equations are getting more complex. Whereas closed formulas exist for $M/G/1$ queues, $M/G/1/K$ queues require to solve an equation system with $K + 1$ unknowns.

Queue systems analysis focus on stable queue, i.e. when the ratio $\rho = \frac{\lambda}{\mu} \leq 1$ where λ (resp. μ) is the expected value of the arrival rate process (resp. service time). But finite queue systems are always stable since the maximal number of pending items is always finite and are subject to loss instead. To model the drop of incoming item in the queue we use the ratio $\rho_e = \frac{\lambda_e}{\mu}$ where λ_e is known as the effective arrival rate and can be determined thanks to equation (1).

$$\lambda_e = \lambda(1 - \pi_K) = \mu(1 - \pi_0) \quad (1)$$

Where π_0 (resp. π_K) in the above equation (1) refers to the probability to the queue at equilibrium to be empty (resp. full).

2.3 Related Work

Chua, Ward, Zhang, *et al.* [10] present an heuristic and an Mixed Integer Programming approach to optimize Service Functions Chain provisioning when using Network Functions Virtualization for a service provider. Their approach relies on minimizing a trade-off between the expected latency and infrastructures resources.

Such optimization routing flow in SDN may need additional information to be exchanged between the nodes of a network. This results in an increase of the volume of signalization, by performing some measurements such as in [11]. This is not a consequent problem in unconstrained networks, i.e. static wired networks with near-infinite bandwidth but may decrease performance of wireless network with poor capacity. An interesting solution to save bandwidth would be to predict some of the KPIs from other KPIs and data exchanged globally between nodes.

In [12, 13], authors proposed a MANETs application of SDN in the domain of tactical networks. They proposed a multi-level SDN controllers architecture to build both secure and resilient networking. While orchestrating communication efficiently under military constraints such as: high-level of dynamism, frequent network failures, resources-limited devices. The proposed architecture is a trade-off between traditional centralized architecture of SDN and a decentralized architecture to meet dynamic in-network constraints.

Jahromi, Hines, and Delanev [14] proposed a Quality of Experience (QoE) management strategy in a SDN to optimize the loading time of all the tile of a mapping application. They have shown the impact of several KPIs on their application using a Generalized Linear Model (GLM). This mechanism make the application aware of the current network state.

Promising works rely on estimating KPIs at a graph-level. Note that it is very difficult, if not impossible, to address this analytically since computer networks models a complex structure of chained interfering queues for each flow in the network.

Rusek, Suárez-Varela, Mestres, *et al.* [4] used GNNs for predicting KPIs such as latency, error-rate and jitter. They relied on the *Routenet* architecture of Figure 2. The idea is to model the problem as a bipartite hypergraph mapping flows to links as depicted on Figure 3. Aggregating messages in such graph may result in predicting KPIs of the network in input. The model needs to know the routing scheme, traffic and links properties. Their result is very promising and has been the subject of two ITU Challenge in 2020 and 2021 [5, 1]. These ITU challenges have very good results since the top-3 teams are around 2% error in delay prediction in the sense of Mean-Absolute Percentage Error (MAPE).

In [6], very promising results were obtained with a a near 1% GNN model error (in the sense of MAPE) on the test set. The model mix analytical $M/M/1/K$ queueing theory used to create extra-features to feed GNN model. In order to satisfy the constraint of scalability proposed by the challenge, the first part of model operates at the link level.

2.4 Dataset

We use public data from the challenge [1] The dataset models static networks that have run for a certain amount of time; the obtained data is a mean of the global working period. The data contains information about

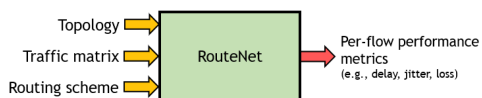


Figure 2: Routenet Architecture [4]

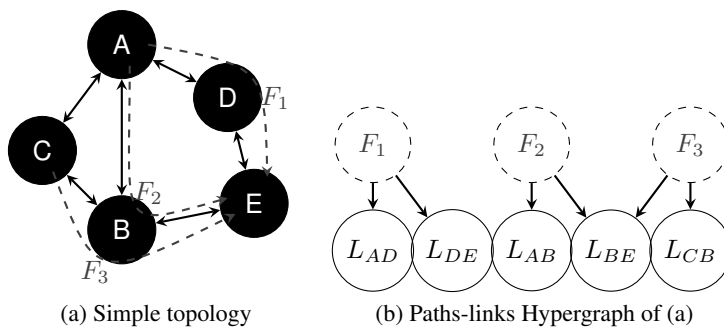


Figure 3: Routenet [4] paths-links hypergraph transformation applied on a simple topology graph carrying 3 flows.

(a) Black circles represents communication node, double headed arrows between them denotes available symmetric communications links and dotted arrows shows flows path. (b) Circle (resp. dotted) represents links (resp. flows) entities defined in the first graph (L_{ij} is the symmetric link between node i and node j). Unidirectional arrows encode the relation " $\langle \text{flow} \rangle$ is carried by $\langle \text{link} \rangle$ ".

the topology of the network, participants and available link characteristics, traffic and routing information. The aim of the GNN ITU Challenge [1] was to build a scalable GNN model in order to predict end-to-end flow latency. Nevertheless, train on one hand and test and validation on the other hand model very different networks. Whereas training dataset models network between 25 and 50 nodes (120,000 samples), test (1,560 samples) and validation (3,120 samples) datasets model networks up to 300 nodes. This results in a very different distribution among these different splits as shown on Figure 4.

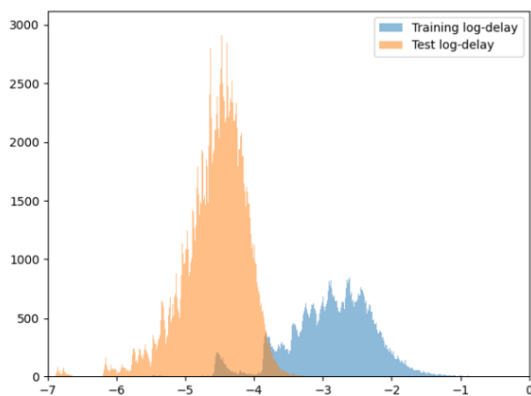


Figure 4: End-to-end latency distribution on train and test datasets of ITU Challenge 2021, where train and test datasets describe networks of very different sizes.[1]

It is important to point out that the proposed data is not in accordance with $M/M/1/K$ queue models since process service time depends on the size of the packet. The size of the packet for each flow follows a Binomial distribution; it can be approximated by a Normal distribution inducing a general service time.

Nevertheless, it turns out that the system does not have the behavior of a $M/G/1/K$ queue system globally but that of a complex system with interconnected queues that cannot be easily modeled.

Hence, approximating the system locally by a mixed of a simple analytical theory ($M/M/1/K$) and black-box optimization (GNNs), as was proposed in [6], is a good approach despite the lack of explicability or interpretability and the high-computational requirements with a lot of parameters to train. We show below that it is possible to obtain comparable performances with other regression approaches.

3 Our approaches

The main question is to define an estimator \hat{y} of the occupancy y according to the various available characteristics of the system, with a joint objective of low complexity and performance. In the following, we present regression approaches based on machine learning and then approaches based on curve-fitting.

Once an estimate of occupancy is obtained, it is possible to get the latency prediction \hat{d}_n for a specific link n by the simple relation

$$\hat{d}_n = \hat{y}_n \frac{\mathbb{E}(|P_n|)}{c_n}$$

where $\mathbb{E}(|P_n|)$ is the observed average packet size on link n and c_n the capacity of this link.

Performances will be evaluated using the MAPE loss-function

$$\mathcal{L}(\hat{y}, y) = \frac{100\%}{N} \sum_{n=1}^N \left| \frac{\hat{y}_n - y_n}{y_n} \right| \quad (2)$$

which is preferred to Mean Squared Error (MSE) because of its scale-invariant property.

3.1 Feature Engineering and Machine Learning

Based on the assumption that the system may be approximated by a model whose essential features come from $M/M/1/K$ and $M/G/1/K$ queue theory,

we took essential parameters characterizing queueing systems, such as: ρ , ρ_e , π_0 , π_K , etc. and built further features by applying interactions and various non-linearities (powers, log, exponential, square root). Then, we selected features in this set by a forward step-wise selection method; i.e. by adding in turn each feature to potential models and keeping the feature with best performance. Finally, we selected the model with best MAPE error. For a linear regression model, this led us to select and keep a set of 4 simple features, which interestingly enough, have simple interpretations:

$$\begin{cases} \pi_0 = \frac{1-\rho}{1-\rho^{K+1}} \\ L = \rho + \pi_0 \sum_k k \rho^k \\ \rho_e = \frac{\lambda_e}{\lambda} \rho = \frac{\lambda_e}{\mu} \\ S_e = \sum_k k \rho_e^k \end{cases} \quad (3)$$

where L is the expected number of packets in the queue according to $M/M/1/K$, π_0 the probability that the queue is empty according to $M/M/1/K$ theory, ρ_e the effective queue utilization, and S_e the unnormalized expected value of the effective number of packet in the queue buffer. These features can be thought as a kind of data preprocessing, before applying ML algorithms, and this turns out to be a key to achieving good performances. The 4 previous features have been kept as input for all the machine learning models.

Next we considered several machine learning algorithm, fitted on the training split and performances were evaluated by test split of a public dataset [1]. Algorithms that were considered are: Multi-Layer Perceptron model (MLP) with 4 layers and with ReLU activation function, Linear Regression, Gradient Boosting Regression Tree (GBRT) with an ensemble of $n = 100$ estimators, Random Forest of $n = 100$ trees and Generalized Linear Model (GLM) with Poisson family and exponential link. All results of these methods are shown in Table 1.

3.2 Curve Regression for occupancy prediction

There is a high interdependence of the features we selected in Equation 3, since all these features can be expressed in term of ρ_e . Furthermore, it is confirmed by data exploration that ρ_e is the prominent feature for occupancy prediction (and in turn latency prediction), as exemplified in Figure 5.

It is then tempting to try to further simplify our features space and to try to estimate the occupancy from a non-linear transformation of the single feature ρ_e , as:

$$\hat{y} = g(\rho_e) \quad (4)$$

where \hat{y} is the estimate of the occupancy y . The concerns are of course to define simple and efficient functions g , with a low number of parameters, that can model the kind of growth shown in Figure 5, and of course to check that the performance remains interesting.

We followed three approaches to design the estimator g in order to predict links occupancy and end-to-end flow latency. In all cases, the parameters of g were computed by minimizing the mean squared of the regression error.

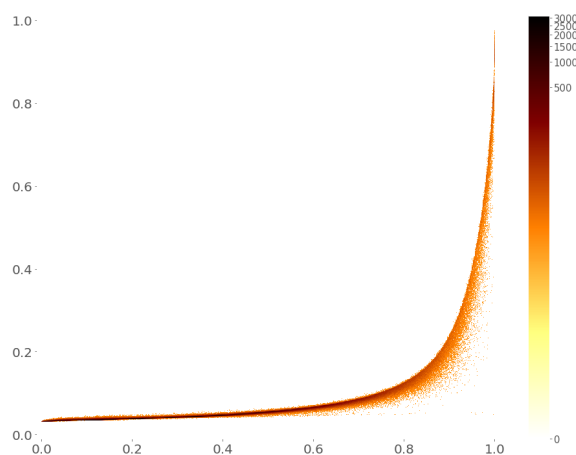


Figure 5: Data of ITU Challenge 2021 [1], ρ_e vs queue occupancy. Color-scale is an indicator of points cloud density.

3.2.1 Exponential of polynomial

The simplest approach is to use a curve-fitting regression of the form

$$\hat{y} = g(\rho_e) = e^{p_n(\rho_e)} \quad (5)$$

where $p_n(x) \in \mathbb{R}_n[x]$ is a polynomial of degree n with real coefficients.

In order to find coefficients of p_n one can obviously consider predicting $\log(y)$ (where y denotes the queue occupancy). Choosing an arbitrary high polynomial degree results to oscillations and increases largely computation time. However choosing a too small degree does not allow the prediction of high occupancy.

3.2.2 Generative polynomials

The estimator g is defined as a linear combination of simple functions (f_n):

$$\hat{y} = g(\rho_e) = \sum_n \alpha_n \cdot f_n(\rho_e) \quad (6)$$

Generative polynomial similar to $M/M/1/K$ theory The idea here is to use a polynomial f_n^K that will match approximately the expression in Equation 3 of the expected number of packets in the queue L .

$$\begin{cases} f_n^K(x) = \frac{\phi_n^K(x)}{\gamma_n} \\ \phi_n^K(x) = x^n \frac{(1-x)}{1-x^{K+1}} & n \geq 0 \forall x \in [0; 1[\\ \gamma_n = \phi_n^K\left(\frac{n}{n+1}\right) & n \geq 0 \end{cases} \quad (7)$$

where K is the size of the queue¹ The sequence of $(f_n)_{n=0}^K$ is finite and defined in interval $[0; 1[$.

In order to improve regression capabilities, each f_n^K is defined as ϕ_n^K normalized by γ_n , a local maximum of ϕ_n^K in the interval $[0; 1[$.

Bernstein Polynomials The previous method relies on polynomial approximation. Since the expected value L can be expressed theoretically in terms of a polynomial of degree K , we are driven to the Bernstein polynomials that form a basis in the set of polynomial in the interval $[0; 1[$:

$$f_n^K(x) = \binom{K}{n} x^n (1-x)^{K-n} \quad (8)$$

The approximation of any continuous function on $[0; 1[$ by a Bernstein polynomial converges uniformly.

3.2.3 Implicit function

The idea here is to define a set of N points $\theta_n = (a_n, b_n)$ and approximate the underlying function by linear interpolation between those points. To obtain a good positioning of these points, we select them as the solution of the following optimization problem:

$$\begin{aligned} \min_{\theta} \quad & \mathcal{L}(f_{\theta}(x), y) + \frac{\alpha}{N} \sum_n \frac{\|\vec{u}_n \times \vec{u}_{n+1}\|^2}{\|\vec{u}_n\|^2 \|\vec{u}_{n+1}\|^2} \\ \text{s.t.} \quad & \vec{u}_n = \theta_{n+1} - \theta_0 n \\ & a_0 = 0 \\ & a_N = b_N = 1 \\ & a_{n+1} - a_n \geq 0 \\ & \theta_n = (a_n, b_n)^T \in [0; 1]^2 \end{aligned} \quad (9)$$

Equation 9 includes a first term for minimizing the interpolation error, and a second term weighted by a parameter $\alpha \geq 0$, to force θ_n sequence to be as aligned and as far as possible. This implies that our sampling will be refined in high curvature zone of our function. The constraint formulated makes θ_n an increasing sequence along the feature axis in order to get a correct interpolation of the curve, especially when N is high enough.

4 Comparison and Discussion

In this section, we evaluate our methods on the data from the GNN ITU Challenge 2021, described in subsection 2.4. We compare our results to those of the challenge winners, which establish the state-of-the-art in terms of pure performance. Since the actual labeled test dataset used for the challenge was released after the end of the challenge, all evaluations are performed on this particular dataset. The Table 1 presents the characteristics of the methods, in terms of the number of input features and parameters to be learned; their performance in the sense of MAPE and MSE; and the values of the execution times, both in learning time and inference time. All results were obtained with the same computer configuration: 120 Go RAM, 1 CPU Intel i9-9920X @ 3.50 GHz with 24 cores and 2 GPUs Nvidia TITAN RTX2080 24Go.

The methods used for comparison are divided into 3 groups, the first being the set of GNN approaches.

¹In results shown Table I., we consider $K = 32$ in order to match the data contained in the ITU challenge dataset [1].

Approaches	Input Features	Model Parameters	MAPE	MSE	Wall Training Time	Wall Inference Time
Routenet [4]	Topology	-	$\gg 100\%$	(N/A)	$\approx 12\text{h}$	-
Top-1 ITU Challenge Team (PARANA)[6]	Traffic matrix Routing Scheme	654006	1.27%	1.10e-5	$\approx 8\text{h}$	214s
MLP		291	1.91%	3.18e-5	$\approx 45\text{min}$	8.26s
Linear Regression ⁺	π_0	4	1.74%	3.20e-5	<1sec	0.296s
GBRT (n=100) [*]	L	4	1.73%	2.90e-5	$\approx 1\text{min}$	0.867s
Random Forest (n=100) [*]	S_e	4	1.69%	3.00e-5	<1sec	0.994s
GLM - Poisson ⁺	ρ_e	4	3.68%	5.09e-4	$\approx 1\text{min}$	0.481s
Curve-fitting exponential (deg=3) ⁺		4	3.94%	3.75e-4	$\approx 1\text{sec}$	0.311s
Curve-fitting exponential (deg=8) ⁺		9	1.70%	3.53e-5	$\approx 5\text{secs}$	0.320
Curve-fitting M/M/1/K ^{**+}		33	2.04%	4.42e-5	$\approx 3\text{min}$	3.55s
Curve-fitting Bernstein ^{**+}		33	1.68%	3.13e-5	$\approx 2\text{min}$	3.14s
Sampling Optimization ($N = 12, \alpha = 0$) ^{**+}		24	1.77%	3.18e-5	$\approx 1\text{min}$	0.281s
Sampling Optimization ($N = 12, \alpha = 1e-5$) [*]		24	1.77%	3.18e-5	$\approx 1\text{min}$	0.306s

Table 1: Results Synthesis of various models for flow latency prediction. Test dataset from [1]
^{*}only 500,000 samples used for training (2.25% of training dataset); ^{**}only 5,000,000 samples used for training (22.5% of training dataset); ⁺under-estimation/over-estimation occurs on high queue occupancy prediction

In the second group, we used classical machine learning models with only 4 input features obtained by stepwise selection, as presented in Equation 3.

In the third group, we group curve regression models using a single well-chosen feature, namely ρ_e , as presented in subsection 3.2.

As we can observe, the proposed approaches achieve a much lower computational time than the GNN approaches, both in terms of learning time and inference time; this at the cost of a marginal performance degradation.

Moreover, non-GNN approaches provide a more local solution since predictions are performed at the link level and not at the whole graph level (Models predict queues occupancy, then compute analytically delay for each link and finally aggregate along path). This would allow to use them for simple local predictions, without having to rely on the global knowledge and prediction of the network.

The consequent gain in computational time of our low-complexity approaches is that they use far fewer parameters, which reduces the amount of data needed for training. The reduction in the number of parameters and the architecture (number of operations) of the solutions explains the drop in learning and inference times.

Nevertheless, when we match the distribution as presented in Figure 5, we notice that most of our data are on a low occupancy level. In practice, some models have a kind of limited behavior when the occupation of the targeted queue is close to 100%: there is a significant over- or under-prediction. However, this behavior does not really affect the overall performance due to the low density of this scenario in our dataset and the predicted values are close enough to the targets.

5 Conclusion

In this paper, we considered the problem of designing efficient and low-cost algorithms for KPI prediction, implementable at the local level. We have argued and proposed several alternatives to GNNs for predicting the queue occupancy of a complex system using simple ML models with carefully chosen features or general curve-fitting methods.

At the cost of a marginal performance loss, our proposals are characterized by low complexity, significantly lower learning and inference times compared to GNNs, and the possibility of local deployment. Thus, this type of solution can be used for continuous performance monitoring.

The low complexity and structures of linear regression algorithms or curve-fitting solutions should also be suitable for adaptive formulations. These last two points are current perspectives of this work. Of course, the approaches considered here will have to be considered and adapted for other types of KPI, such as error-rate or jitter.

Finally, a last point that deserves interest is the fact that these low complexity models can be interpreted/explained either by direct inspection (visualization), or by using tools such as Shapley values [15] which allow to interpret output values by measuring contributions of each input feature on the prediction.

References

- [1] J. Suárez-Varela *et al.*, “The graph neural networking challenge: A worldwide competition for education in AI/ML for networks,” *ACM SIGCOMM Computer Communication Review*, vol. 51, no. 3, pp. 9–16, 2021. DOI: 10.1145/3477482.3477485.
- [2] S. Singh and R. K. Jha, “A survey on Software Defined Networking: Architecture for next generation network,” *Journal of Network and Systems Management*, vol. 25, no. 2, pp. 321–374, 2017.
- [3] R. Amin, M. Reisslein, and N. Shah, “Hybrid SDN networks: A survey of existing approaches,” *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3259–3306, 2018. DOI: 10.1109/COMST.2018.2837161.
- [4] K. Rusek, J. Suárez-Varela, A. Mestres, P. Barlet-Ros, and A. Cabellos-Aparicio, “Unveiling the potential of graph neural networks for network modeling and optimization in SDN,” in *Proceedings of the 2019 ACM Symposium on SDN Research*, 2019, pp. 140–151.
- [5] *The graph neural networking challenge 2020*. <https://bnn.upc.edu/challenge/gnnet2020>.
- [6] B. K. de Aquino Afonso, *GNNet challenge 2021 report (1st place)*, <https://github.com/ITU-AI-ML-in-5G-Challenge/ITU-ML5G-PS-001-PARANA>, 2021.
- [7] W. L. Hamilton, “Graph representation learning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 14, no. 3, pp. 1–159, 2020.
- [8] D. Bacciu, F. Errica, A. Micheli, and M. Podda, “A gentle introduction to deep learning for graphs,” *Neural Networks*, vol. 129, pp. 203–221, 2020.
- [9] R. B. Cooper, “Introduction to queueing theory,” *Edward Arnold, London*, 1981.
- [10] F. C. Chua, J. Ward, Y. Zhang, P. Sharma, and B. A. Huberman, “Stringer: Balancing latency and resource usage in service function chain provisioning,” *IEEE Internet Computing*, vol. 20, no. 6, pp. 22–31, 2016.
- [11] S. T. V. Pasca, S. S. P. Kodali, and K. Kataoka, “AMPS: Application aware multipath flow routing using machine learning in SDN,” in *2017 Twenty-third National Conference on Communications (NCC)*, IEEE, 2017, pp. 1–6.
- [12] K. Poularakis, G. Iosifidis, and L. Tassiulas, “SDN-enabled tactical ad hoc networks: Extending programmable control to the edge,” *IEEE Communications Magazine*, vol. 56, no. 7, pp. 132–138, 2018.
- [13] K. Poularakis, Q. Qin, E. M. Nahum, M. Rio, and L. Tassiulas, “Flexible SDN control in tactical ad hoc networks,” *Ad Hoc Networks*, vol. 85, pp. 71–80, 2019, ISSN: 1570-8705. DOI: 10.1016/j.adhoc.2018.10.012.
- [14] H. Z. Jahromi, A. Hines, and D. T. Delanev, “Towards application-aware networking: ML-based end-to-end application KPI/QoE metrics characterization in SDN,” in *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, IEEE, 2018, pp. 126–131.
- [15] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.