Application of Linear Regression in Latency Estimation in Packet-Switched 5G xHaul Networks

Mirosław Klinkowski[1*], **Jordi Perelló**[2], **Davide Careglio**[2]
[1] National Institute of Telecommunications, Warsaw, Poland
[2] Universitat Politècnica de Catalunya (UPC) – BarcelonaTech, Barcelona, Spain
[*] *M.Klinkowski@il-pib.pl*

ABSTRACT In this work, we aim at investigating the applicability of machine learning (ML), namely, a linear regression (LR) model for estimation of flows latencies in packet-switched 5G xHaul networks. The analysis is performed in a network scenario in which the switches are connected using high-capacity fiber links.
Keywords: 5G; xHaul; packet-switched network; latency modeling, machine learning, linear regression.

1. Introduction

Packet-switched xHaul networks are a scalable solution enabling convergent transport of diverse types of radio data flows, such as fronthaul/midhaul/backhaul (FH/MH/BH) flows, between remote sites and a central site (hub) in 5G radio access networks (RANs) [1]. Such networks can be realized using the cost-efficient Ethernet technology, which enhanced with time-sensitive networking (TSN) features allows for prioritized transmission of latency-sensitive FH flows [2]. The major concern in such networks are non-deterministic latencies caused by packet buffering in Ethernet bridges (switches). For this reason, in latency-sensitive xHaul networks, resource allocation and network optimization tasks realized during both off-line network planning and online connection provisioning should take this issue into account, e.g., by making use of reliable models estimating flows latencies. Queuing theory-based models are not suitable in the majority of cases since they are oriented on estimation of average queuing latencies [3], whereas 5G RANs have strict requirements concerning maximum latencies. Concurrently, worst-case (WC) latency models, which guarantee that flows latencies do not exceed given limits, as they calculate an upper bound of the latencies experienced by packets, tend to overestimate the latencies [4]. Since overestimation of latencies may lead to over-dimensioning of allocated network resources, it is desirable to predict the latencies as accurate as possible.

The main goal of this work is to assess the applicability of LR in prediction of maximum flows latencies in packet-switched xHaul networks. The approach that we propose and study is based on the calculation of worst-case latencies by means of a deterministic model and, afterwards, on improving the accuracy of these estimations with the assistance of an LR model trained using the data obtained in an event-driven simulator of a packet xHaul network. To our best knowledge, this approach has not been considered in the literature so far.

2. Network Scenario and Main Assumptions

We assume a similar network and traffic model as in [5]–[7]. The elements of the 5G RAN considered include: (a) the remote units (RUs) located at antenna sites, (b) the distributed units (DUs) comprising a subset of baseband processing functions, which can be virtualized and performed at different sites of the network at processing pool (PP) facilities, and (c) the central units (CUs) that complete the radio processing at a hub node. The RUs, PPs, and the hub are connected using a convergent packet-switched xHaul network [1], which makes use of Ethernet switches [2] for multiplexing and routing of multiple data flows transported between the RAN elements. The xHaul network is connected using fiber links of capacity: 25 Gbps (RU–switch), 100 Gbps (switch–switch), and 400 Gbps (switch–PP and switch–hub). Two types of data flows of different bandwidth and latency requirements are considered: (a) FH – between RU and PP (DU), and (b) MH – between PP and the hub (CU). The transmission is realized in both uplink (RU→PP→hub) and downlink (hub→PP→RU) directions.

As in [3], the radio data are encapsulated and sent periodically in transmission windows (every $66.\bar{6}$ µs) as bursts of Ethernet frames (packets); each frame having a fixed size of 1542 bytes [2]. The data are transmitted with constant bit-rates corresponding to the capacity of a radio system consisting of 4 antennas with MIMO and 100 MHz channels. Option 7.2 and Option 2 are assumed for the functional split [8] in FH and MH, respectively. Table I shows the adequate bit-rates of xHaul flows, estimated using the model presented in [9], the number of frames forming each burst (denoted as burst size), and burst transmission delays in network links. The bursts are buffered and transmitted as entire in network switches. The selection of a burst for transmission is based on the strict priority algorithm [2], without preemption, where the FH bursts have assigned a higher priority than the MH bursts [2], and following the FIFO policy for the queued up bursts of the same priority.

This study focuses on the maximum flow latency, which is defined as the maximum latency that a burst belonging to the flow may experience during its transmission through the network. As in [5]–[7], the latency of a burst consists of a static (fixed) and a dynamic (dependent on the network state) component. The static latency includes the propagation delay in network links (assuming 2×10^5 km/s speed), the store-and-forward

DOI: 10.1109/ICTON59386.2023.10207222

TABLE I: Assumed parameters of FH and MH flows.

Direction	Flow type	Flow bit-rate [Gbps]	Burst size	Burst delay in link [µs]		
				25 Gbps	100 Gbps	400 Gbps
Uplink	FH	5.496	31	15.30	3.82	0.96
	MH	0.774	5	2.47	0.62	0.15
Downlink	FH	6.076	34	16.78	4.19	1.05
	MH	1.016	6	2.96	0.74	0.18

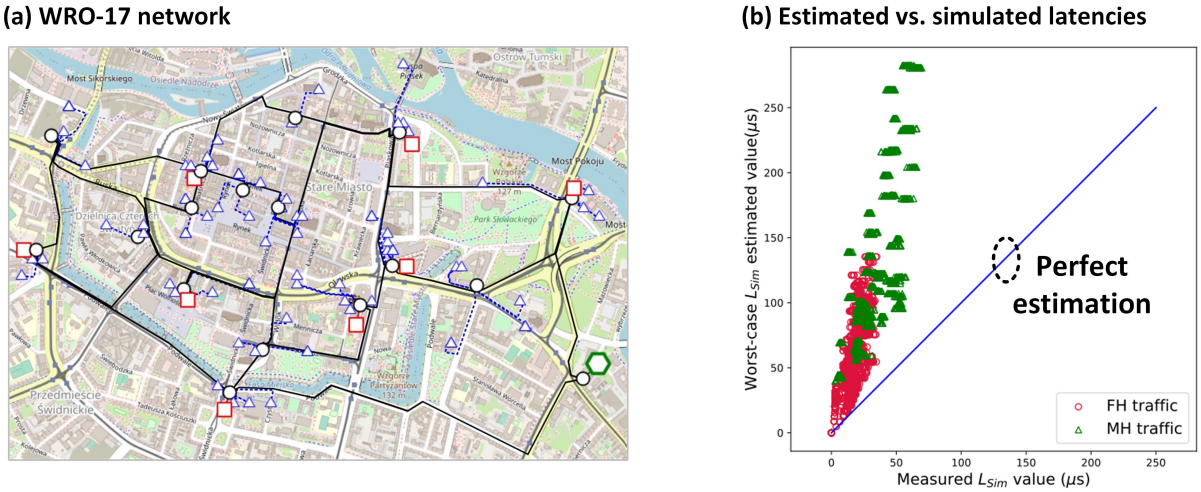

Figure 1: (a) WRO-17 network topology; RUs, PPs, switches, and the hub are marked by triangles, squares, circles, and hexagon, respectively; (b) estimated vs. measured latencies of FH and MH flows.

delay in switches (5 µs per switch), and the burst transmission delay in links (shown in Table I), whereas the dynamic latency represents the burst buffering delay at the output links of switches. The maximum buffering latency is calculated using the WC latency model described in [2]. Namely, for a given flow, a buffering delay in a link is produced by: (a) all bursts that belong to other flows of either higher or equal priority, which might be selected for transmission before the burst considered, and (b) the largest burst of a lower priority flow, which might be in-transmission. For flow f routed through links e belonging to path p, WC buffering latency $L^{\text{buf}}(f)$ is expressed as:

$$L^{\text{buf}}(f) = \sum_{e \in p} \left(\sum_{q \in Q^{\text{HEP}}(f,e)} L(q,e) + \max_{q \in Q^{\text{LP}}(f,e)} L(q,e) \right) \tag{1}$$

where $Q^{\text{HEP}}(f,e)$ are higher/same priority flows and $Q^{\text{LP}}(f,e)$ are lower priority flows interfering with f in link e, and $L(q,e)$ is the latency introduced by interfering flow q in link e.

3. Data Model

The data used in this study was obtained in a made-up mesh network WRO-17, shown in Fig. 1a, consisting of 17 switches (depicted by circles) and PP nodes (squares) placed in the proximity of real antenna locations (79 RUs in total, marked by triangles) in the center of city Wrocław in Poland and connected using links driven along streets. The one-way latency limits of the FH flows were randomly selected for particular RUs, between 100 µs and 250 µs, whereas the MH flows had a fixed limit of 1 ms. The selection of PP nodes for the DU processing for particular RUs and the routing of related FH and MH flows was planned using the latency-aware optimization model presented in [7], which resulted in 316 flows in total (assuming both uplink and downlink).

Next, the network flows were deployed in an event-driven simulator implemented in the OMNET++ environment [10], which simulated the transmission and buffering of the xHaul data bursts in a packet-switched transport network. The simulation was performed for 1e7 generated bursts. To suppress the repeatability of buffering states due to the periodic character of transmission (see Section 2), the bursts were transmitted with random offset times in consecutive transmission windows. During the simulation, the latency information of each transmitted burst was collected and, after the simulation, the largest latency of a burst belonging to a flow represented the measured (simulated) latency of the flow.

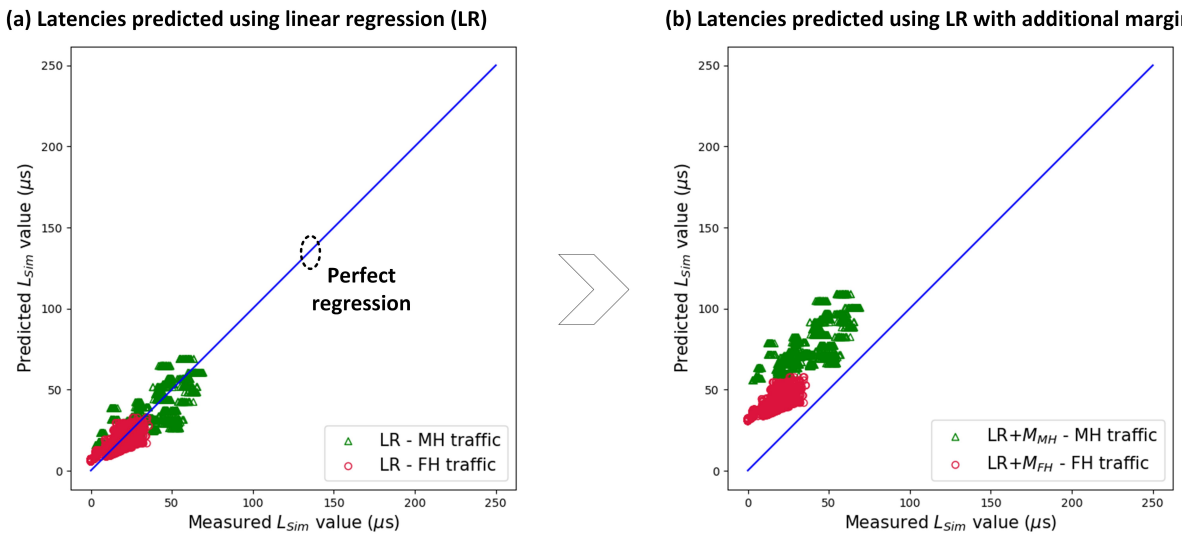

Figure 2: Predicted vs. measured flows latencies: (a) obtained using LR and (b) after including additional margin.

The network planning and simulation process was repeated 10 times for different random FH latency limits, which allowed to create a data set comprising 3160 samples, each sample corresponding to a single data flow. Fig. 1b shows a plot of the measured values (L_{sim}) and the WC estimations (L_{wc}) of the buffering latency, in the form of (L_{sim}, L_{wc}) points, for all 3160 samples included in the dataset, where L_{wc} was calculated using Equation (1). It can be seen that the WC estimation model tends to significantly overestimate the measured L_{sim} values, particularly for the flows with larger latencies.

4. Linear Regression Model

In this work, we make use of a linear regression model in the prediction of maximum latencies of xHaul flows (i.e., the values of L_{sim}). The static component of the flow latency is fixed as it does not depend on the network state and its exact value can be calculated easily. Hence, our focus is on predicting buffering delays, which in many cases are overestimated significantly when using deterministic model (1), as shown in Fig. 1b.

The LR model was implemented using the widely used scikit-learn Python library. The model was trained and tested using the dataset defined in Section 3. Three input features were used to predict the values of L_{sim}:

- Flow type – with the assigned numerical values: 0 for FH uplink, 1 for FH downlink, 2 for MH uplink, and 3 for MH downlink;
- Number of buffers appearing along the flow path;
- Estimated worst-case latency value L_{wc}.

K-Fold cross validation (K = 5) was applied to verify the accuracy of the regression model and the 70% mean accuracy score was obtained.

Fig. 2a shows a plot of the measured values (L_{sim}) and the values predicted using LR (L_{pred}) of the buffering latency as (L_{sim}, L_{pred}) points for all samples in the dataset. The observed differences between L_{sim} and L_{pred} are much smaller than between L_{sim} and L_{wc} in Fig. 1b, which indicates that the LR model provides a more accurate prediction of latencies. Still, the application of regression introduces some underestimation of maximum latencies for the points in the chart below the ”perfect regression” line, which might result in unreliable network operation. To solve this issue, additional latency margins need to be added to prevent falling into underestimating the L_{sim} values. For the dataset considered, the sufficient margins are, e.g., $M_{\text{FH}} = 25$ µs and $M_{\text{MH}} = 40$ µs for FH and MH data flows, respectively.

Fig. 2b presents the plot of measured and predicted latency values after increasing the L_{pred} value by the above margins. As can be seen, all the depicted points lay above the ”perfect regression” line, which means that the predictions are not underestimated. Moreover, even including the additional latency margins, more accurate maximum latency values can be delivered than with the WC latency estimation analytical model. In particular, the mean absolute error (MAE) is 33 µs for the LR model, whereas it is 68 µs for the WC model, which results in the MAE reduction of about 52%.

To assess the potential gains in network performance from applying the LR-aided latency predictions, we analyze the overall (end-to-end, E2E) flows latencies, which include both latency components, i.e., static and dynamic. In Fig. 3, we show histograms illustrating the distribution of the amount of FH flows with particular E2E latencies for the WC estimations (in Fig. 3a) and the LR-aided predictions (in Fig. 3b). Also, the ”actual”

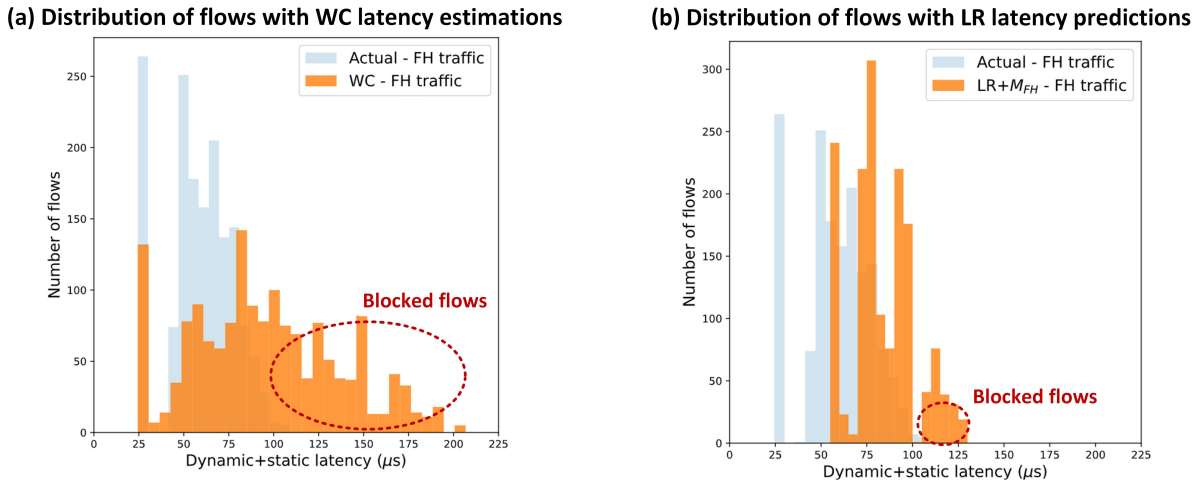

Figure 3: Distribution of FH flows with particular E2E latencies: (a) WC estimations and (b) LR predictions.

E2E latencies corresponding to the measured L_{sim} values are shown at the background of the figures. We can see that the proposed LR model delivers more precise E2E latency predictions. Moreover, the amount of flows that might fall in the range of unacceptable latencies is higher if the latencies are estimated using the WC model than when based on the LR predictions. For instance, assuming the one-way latency limit of 100 µs for FH flows, there are 662 flows and 208 flows (out of all 1580 flows) with the latencies exceeding this limit, respectively, if the WC estimations and LR predictions are considered. Note that if the decisions concerning the acceptance of particular flows have to be taken, about 69% less flows would be blocked after applying the LR model. Eventually, for the actual (measured) values of E2E latencies, only 7 flows have larger latencies than 100 µs, which still leaves a significant room for improvement.

5. Concluding Remarks

We have analyzed the applicability of linear regression in predicting the maximum latencies of xHaul flows in packet-switched 5G xHaul networks. We have shown that the use of LR has a potential to improve the prediction of flows latencies in such networks, which may lead to a better network performance. In future work, we will continue the study assuming more diversified traffic scenarios, also involving the LR prediction model directly in the network planning process, as well as we will examine the effectiveness of other ML methods.

Acknowledgments

The work was supported by National Science Centre, Poland under Grant No. 2018/31/B/ST7/03456. Moreover, it was also supported by the Spanish I+D+i project TRAINER-A (ref. PID2020-118011GB-C21), funded by MCIN/AEI/10.13039/501100011033.

References

[1] IEEE, "1914.1-2019 - IEEE standard for packet-based fronthaul transport networks," Nov. 2019.

[2] ——, "802.1CM-2018 – IEEE standard for local and metropolitan area networks – time-sensitive networking for fronthaul," Nov. 2018.

[3] G. O. Perez, D. Larrabeiti, and J. A. Hernandez, "5G new radio fronthaul network design for eCPRI-IEEE 802.1CM and extreme latency percentiles," *IEEE Access*, vol. 7, pp. 82 218–82 229, 2019.

[4] J.-Y. L. Boudec and P. Thiran, *Network Calculus – A Theory of Deterministic Queuing Systems for the Internet*, ser. Lecture Notes in Computer Science (2050). Springer Verlag, 2001.

[5] M. Klinkowski, "Optimization of latency-aware flow allocation in ngfi networks," *Comp. Commun.*, vol. 161, pp. 344–359, 2020.

[6] ——, "Latency-aware DU/CU placement in convergent packet-based 5G fronthaul transport networks," *Appl. Sci.*, vol. 10, no. 21, 2020.

[7] D. Mroziński, M. Klinkowski, and K. Walkowiak, "Cost-aware DU placement and flow routing in 5G packet xHaul networks," *IEEE Access*, vol. 11, pp. 12 710–12 726, 2023.

[8] 3GPP, "Study on new radio access technology: Radio access architecture and interfaces," Tech. Rep. TR 38.801, v14.0.0, Sophia Antipolis, France, 2017.

[9] S. Lagen, L. Giupponi, A. Hansson, and X. Gelabert, "Modulation compression in next generation RAN: Air interface and fronthaul trade-offs," *IEEE Comm. Mag.*, vol. 59, no. 1, pp. 89–95, 2021.

[10] A. Varga, "OMNeT++ discrete event simulator." [Online]. Available: https://omnetpp.org/