INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2023, VOL. 69, NO. 2, PP. 335–340
Manuscript received January 20, 2023; revised April, 2023. DOI: 10.24425/ijet.2023.144369

Analysis of Latency-Aware Network Slicing in 5G Packet xHaul Networks

Mirosław Klinkowski

***Abstract*—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). 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 fronthaul flows. Provisioning of multiple types of 5G services of different service requirements in a shared network, commonly referred to as network slicing, requires adequate handling of transported data flows in order to satisfy particular service / slice requirements. In this work, we investigate two traffic prioritization policies, namely, flow-aware (FA) and latency-aware (LA), in a packet-switched xHaul network supporting slices of different latency requirements. We evaluate the effectiveness of the policies in a network-planning case study, where virtualized radio processing resources allocated at the processing pool (PP) facilities, for two slices related to enhanced mobile broadband (eMBB) and ultra-reliable low latency communications (URLLC) services, are subject to optimization. Using numerical experiments, we analyze PP cost savings from applying the LA policy (vs. FA) in various network scenarios. The savings in active PPs reach up to $40\% - 60\%$ in ring scenarios and 30% in a mesh network, whereas the gains in overall PP cost are up to 20% for the cost values assumed in the analysis.**

***Keywords*—5G; radio access networks; packet-switched xHaul; latency-sensitive network; network slicing; traffic prioritization; network optimization**

I. Introduction

The 5th generation mobile networks (5G) enable provisioning of diverse wireless communication services of diversified throughput and latency requirements, such as eMBB, URLLC, and massive machine type communications (mMTC). To facilitate the deployment of these services, the technical specifications of the 5G RAN architecture [1], [2] defined by the 3GPP organization allow for splitting of radio baseband processing functions between separate entities, namely, radio (RU), distributed (DU), and central (CU) unit. The DUs and CUs can be virtualized and executed on general-purpose processors available at processing pool (PP) facilities [3] located at different sites of the network, as shown in Fig. 1.

The IEEE 1914.1 standard [4] defines the next generation fronthaul interface (NGFI) architecture, which assures scalable and flexible connectivity between the RAN elements by means of a packet-switched xHaul transport network. NGFI allows for convergent transport of diverse radio data flows, such as fronthaul (RU-DU), midhaul (DU-CU), and backhaul (CU-5G Core/5GC) data flows, between the RUs, PPs, and a hub site / data center (DC), in which the traffic is aggregated. The NGFI architecture supports also the realization of multiple types of services with very different bandwidth and latency requirements (such as eMBB, URLLC, mMTC, 4G) in a shared physical network infrastructure, which is commonly referred to as network slicing [5], [6].

The enabling technology for packet-based xHaul networks is the well-known Ethernet, which has been adapted for fronthaul networks in the IEEE standard 802.1CM [7]. The 5G radio data are encapsulated into Ethernet frames (packets) according to the enhanced common public radio interface (eCPRI) protocol [8]. Ethernet allows for statistical multiplexing of multiple xHaul data flows, by these means improving the utilization of link bandwidth. However, the nondeterministic nature of packet transmission requires packet buffering in Ethernet bridges (switches), whenever contention in the access to the output link occurs. The buffering introduces unpredictable latencies, which may affect the quality of service (QoS) of data flows. Therefore, TSN mechanisms enabling prioritized transmission of latency-sensitive fronthaul traffic have been introduced in [7].

Standard IEEE 802.1CM specifies a relatively small number of priority classes. In particular, it discerns three classes of priority, a high priority class associated with the fronthaul traffic of low maximum end-to-end one-way latency requirements (100 μs is assumed), and two lower priority classes for the data flows of much lower latency requirements (at the level of 1 ms and 100 ms). Although this definition suits a single-slice case, where the priority classes might correspond to the FH and MH / BH flows, the efficiency of such approach is not obvious in network slicing scenarios. Namely, specific 5G services, such as eMBB and URLLC [4], may have different fronthaul latency requirements and the assignment of the same priority to the fronthaul flows of both services may have a negative impact on network performance.

In related works, the authors of [9] focused on the design of a fronthaul network based on the IEEE 802.1CM standard assuming a single low-latency profile of fronthaul traffic. In [10], optimal allocation of radio resources for eMBB and URLLC slices was studied, however, without considering the xHaul transport domain. In [11], a slice-aware optimization of the placement of DU and CU processing resources was

This work was supported by National Science Centre, Poland under Grant No. 2018/31/B/ST7/03456.

M. Klinkowski is with the National Institute of Telecommunications, Warsaw, Poland (e-mail: M.Klinkowski@il-pib.pl).

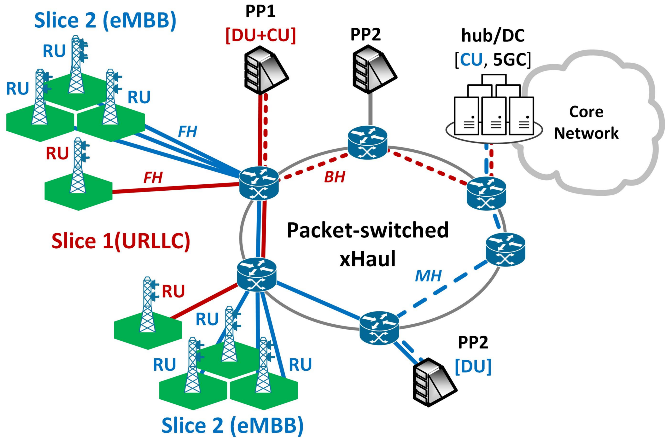

Fig. 1. Network slicing for URLLC and eMBB in a packet xHaul network

addressed, but without accounting for the transport of traffic in a packet-based xHaul network. The authors of [12] presented experimental results obtained in a test-bed of a TSN-aware xHaul transport network supporting eMBB and URLLC slices. In [12], the eMBB traffic was handled as the best-effort traffic without latency guarantees. To our best knowledge, in the literature there is lack of studies evaluating potential performance gains from the differentiation of the fronthaul traffic with low-latency guarantees in packet-switched xHaul network scenarios.

The main goal of this work is to investigate whether, in which cases, and to what extent, it is beneficial to extend the definition of the IEEE 802.1CM standard and consider additional priority classes in network slicing scenarios with diversified fronthaul latency requirements in a packet-switched xHaul network. To this end, we study two traffic prioritization policies: flow-aware, where the whole fronthaul traffic has the same priority, and latency-aware, where the flows of different latency limits have different priorities. The analysis is performed for a slice-aware xHaul network planning case study, which allows us to evaluate the impact of both policies on the network cost related to the required PP processing resources. The main conclusion is that the prioritization of traffic according to latency requirements of networks slices (5G services) allows to better optimize radio processing resources in a packet xHaul network than when a single low-latency fronthaul profile is considered for all slices.

The paper is organized as follows. In Section II, we present the details of the network, traffic, and latency model as well as the network planning case study considered in this work. In Section III, we describe the traffic prioritization policies. In Section IV, we present and discuss numerical results and, in Section V, we conclude this work.

II. Network Scenario

A. Main Assumptions

The network and traffic models are similar as in [13], [14]. The network implements the NGFI architecture [4], where the RUs, PPs, and the hub are connected using a packet-switched

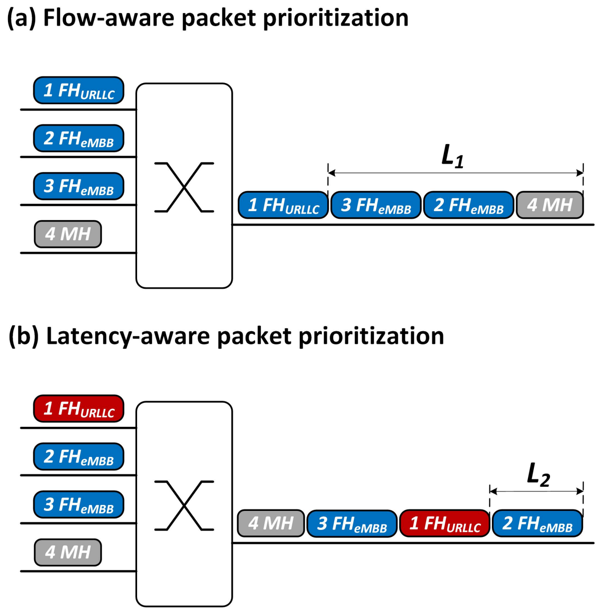

Fig. 2. xHaul traffic prioritization policies: (a) flow-aware and (b) latency-aware

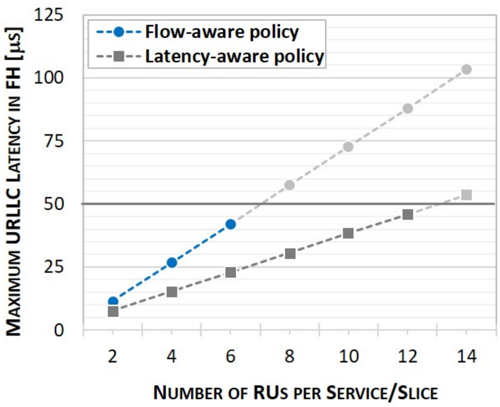

Fig. 3. Estimated worst-case latencies of URLLC flows in a fronthaul link aggregating the eMBB and URLLC traffic, in a function of the number of RUs (per slice) for the FA and LA traffic prioritization policies

xHaul network making use of Ethernet switches [7] for multiplexing and routing of the FH, MH, and BH data flows. Both uplink (RU→PP→hub) and downlink (hub→PP→RU) transmission directions are considered. Also, we assume that clusters of RUs might require joint DU processing, at the same PP node, for multi-cell coordination purposes [15].

We assume that the xHaul network operates in a slice-aware mode [4] and it supports two types of 5G services (slices), namely, eMBB and URLLC. The RUs are associated with particular slices. Following the deployment scenarios specified in [4], the URLLC slice is implemented with a single split between RU and DU/CU, the latter located at a PP, whereas for the eMBB slice we assume a double-split scenario with the DU and CU disaggregated and located, respectively, at a PP and at the hub. In network evaluation, we assume the one-way latency limits equal to 50 μs and 100 μs, respectively, for the URLLC and eMBB flows in FH [4], and 0.5 ms for MH and BH flows. Also, we consider that bandwidth requirements may differ in both slices (see Sec. IV).

As in [9], radio data are encapsulated [8] and sent periodically (every 66.$\bar{6}$ μs) as bursts of Ethernet frames (packets); each frame having a fixed size of 1542 bytes [7]. The data are transmitted with constant bit-rates assuming full utilization of radio resources. The bursts are buffered and transmitted as entire in network switches, and they are selected for transmission based on their priorities, as described in more details in Sec. III.

In estimation of flows latencies, we include both static and dynamic latencies that the bursts of Ethernet frames undergo during their transmission through the network, similarly as in [13], [14]. The static latencies account on propagation delays in network links (assuming 2×10^5 km/s speed), storing-and-forwarding delays in switches (5μs per a switch), and burst transmission times in links (dependent on burst lengths and link bit-rates). The dynamic latencies represent buffering delays of the bursts at switch output links and, to estimate them, we apply the worst-case latency calculation model described in Section 7.2 in [7]. Namely, for a given flow, a buffering delay in a link is produced by: (a) the 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 links e belonging to path p, dynamic latency $L^{\mathrm{d}}(f)$ is expressed as:

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

where $Q^{\mathrm{HEP}}(f,e)$ are higher/same priority flows and $Q^{\mathrm{LP}}(f,e)$ 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.

B. Case study: slice-aware xHaul network planning

To evaluate the xHaul traffic prioritization policies studied in this work (see next Section), we consider a slice-aware xHaul network planning case study. The planning problem concerns the selection of PP nodes for placement of DU (in eMBB) and DU / CU (in URLLC) entities for given set of RUs, and the routing of corresponding FH, MH, and BH flows between the RUs, selected PPs, and the hub. As in [14], we consider that the flows are routed over shortest paths. The problem constraints are related to: (a) the selection of a common PP for each cluster of RUs, (b) capacity limits of network links, (c) latency limits of particular flows. The problem objective is to minimize the overall PP cost, which is expressed as:

$$cost^{\mathrm{PP}} = \sum_{v \in \mathcal{V}^{\mathrm{PP}}} \left(\kappa^{\mathrm{activ}}(v) \cdot \alpha(v) + \kappa^{\mathrm{proc}}(v) \cdot \rho(v) \right) \quad (2)$$

where $\mathcal{V}^{\mathrm{PP}}$ denotes the set of candidate PP nodes, $\alpha(v)$ indicates if PP node v is active (used), and $\kappa^{\mathrm{activ}}(v)$, $\kappa^{\mathrm{proc}}(v)$, and $\rho(v)$ are, respectively, the activation cost, the processing cost, and the processing load at node v. The activation cost may correspond, e.g., to renting the space for the placement of servers. The processing cost may represent the cost of energy consumed by servers when executing virtualized DUs/CUs. The optimization oriented on selecting the cheapest locations for the activation of the PPs and the allocation of the DU/CU workloads will let the network operator to optimize the network cost. Note that it goes beyond previous studies (e.g., see [13], [16]) in which only the number of active PPs was subject to optimization without accounting for their costs.

We model and solve the slice-aware xHaul network planning problem as a mixed-integer programming (MIP) problem. The MIP model is an extension of the MIP formulation proposed in [14] for a single-slice scenario. The modifications concern:

- the fixed placement of CUs for eMBB at the hub node, which is achieved by setting the CU placement variable u^C_{dv} equal to 1 for the PP node v representing the hub node for all demands d realizing the eMBB service;
- the collocation of DU/CU entities for URLLC, which is imposed by setting the DU placement variable u^D_{dv} equal to the CU placement variable u^C_{dv} for all candidate PP nodes v and all demands d realizing the URLLC service;
- the admission of different FH latency limits (denoted in [14] by $L^{max}(f)$, where $f = \{\mathrm{FH}\}$) for the demands belonging to different slices/services.

The above-mentioned variables u^D_{dv} and u^C_{dv} form part of the optimization model formulated in [14]. Due to slight differences, and in order not to repeat the model formulation, we refer to [14] for a complete description of the MIP model.

III. xHaul Traffic Prioritization

In standard IEEE 802.1CM [7], the selection of packets for transmission at switch output ports is realized according to the strict priority algorithm, which makes decisions based on the priority levels of the queued up packets. In particular, higher priority packets are always selected before lower priority packets at given output port, whereas the order of transmission of the packets of same priority is arbitrary. In this work, we assume such a mode of switch operation. Moreover, the switches apply Profile A of operation, defined in Sec. 8.1 in [7], according to which preemption of lower priority frames during their transmission is not allowed.

A basic approach to prioritization of xHaul traffic in switches is to assign priorities to particular types of flows. Such a flow–aware (FA) policy is considered in [7], where the highest priority is assigned to latency-sensitive fronthaul flows (100 μs maximum end-to-end one-way latency is assumed), whereas other types of flows with much lower maximum latency limits (namely, 1 ms and 100 ms) are associated with two lower priority classes. Taking into account that particular 5G services might have diverse fronthaul latency requirements, the FA policy may not be suitable for network slicing scenarios. In particular, we consider to make a distinction between such services and assign priorities in accordance to the latency requirements of particular fronthaul flows, namely, the flows with lower maximum latencies have higher priorities. We refer to it as a latency–aware (LA) policy. Taking the above into consideration, in the network scenario assumed in this work, the relation between priorities of flows is the following:

- FA policy: $\mathrm{FH}_{\mathrm{URLLC}}$ = $\mathrm{FH}_{\mathrm{eMBB}}$ > MH = BH;

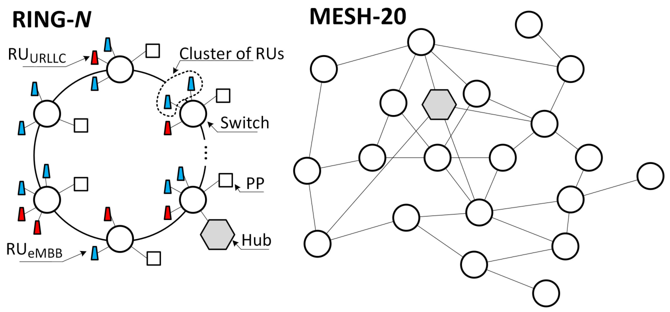

Fig. 4. Network topologies: RING-N and MESH-20

- LA policy: FH$_{\text{URLLC}}$ > FH$_{\text{eMBB}}$ > MH = BH.

In Fig. 2, we illustrate the difference between both policies by an example of a packet switch aggregating four flows: FH$_{\text{URLLC}}$ with 50 μs max. latency ($\times$1), FH$_{\text{eMBB}}$ with 100 μs max. latency ($\times$2), and MH with a high latency limit ($\times$1). If FA is applied, it may happen that the FH$_{\text{URLLC}}$ packet is delayed by time L_1 related to the transmission of both FH$_{\text{eMBB}}$ packets and, additionally, of the MH packet, which might be in-transmission at the moment of the arrival of the FH packets. In case of LA, the FH$_{\text{URLLC}}$ packet has the highest priority and, at the worst case, it has to wait for time $L_2 < L_1$ until the transmission of the longest packet of a lower priority is accomplished.

In Fig. 3, we show maximum buffering latencies estimated for URLLC flows in a 100 Gbps uplink link aggregating FH traffic from the eMBB and URLLC RUs, in a function of the number of RUs (the same in both slices), assuming the traffic and latency models described in Sec. II, and the FH flow bit-rate shown in Table I. We can see that the LA policy allows to serve the URLLC slice consisting of up to 12 RUs with flow latency guarantees below a 50 μs limit, whereas FA is able to support 6 RUs at most.

In the next Section, we analyze potential performance gains from using LA in network scenarios.

IV. Numerical Results

The flow-aware (FA) and latency-aware (LA) traffic prioritization policies are evaluated in a ring network topology (RING-N) and a 20-node mesh network (MESH-20), both shown in Fig. 4. RING-N consists of N switching nodes, where we assume $N \in \{6, 8, 10\}$ in accordance to [4]. Rings are common in access / aggregation networks [4], [17], whereas MESH-20 was studied in [18]. A given number of RUs (denoted by R) are connected (randomly) to the switches, where α percent of RUs belongs to the URLLC slice and the rest to eMBB. The clusters of RUs consist of the RUs belonging to the same slice and linked with the same switch. A PP node, which can be activated and used for DU&CU (in URLLC) and DU (in eMBB) processing, is connected to each switch. The PP activation and processing costs are uniformly distributed, respectively, between 50–100 and 5–10 of cost units. The lengths of links (in km) are random within the following limits: $[0.2 \ldots 0.5]$ for RU and PP connections, $[1 \ldots 3]$ between switches, and $[4 \ldots 6]$ for the hub. The link capacities are 25 Gbit/s for RUs, 100 Gbit/s between switches, and 400 Gbit/s for PP and hub connections. The generate paths, we used Dijkstra's shortest-path algorithm.

TABLE I
Reference bit-rates (in Gbit/s) of xHaul data flows

	Fronthaul	Midhaul	Backhaul
Uplink	5.496	0.774	0.750
Downlink	6.076	1.016	1.000

The assumed reference bit-rates of xHaul data flows, shown in Table I, have been estimated using the model presented in [19] for a radio system of 4 antennas with MIMO and 100 MHz channels, and assuming the functional split Option 7.2 for RU–DU and Option 2 for DU–CU [1]. We consider different xHaul bit-rates for eMBB and URLLC slices. In particular, parameter γ represents the relative difference between the eMBB and URLLC bit-rates, and the reference values in Table I are multiplied by $\sqrt{\gamma}$ and $\sqrt{\gamma^{-1}}$ for eMBB and URLLC, respectively. Note that the xHaul bit-rates of both services are the same for $\gamma = 1$. As mentioned in Section II, 50 μs and 100 μ limits are assumed, respectively, for URLLC and eMBB fronthaul flows.

Regarding performance metrics, our main focus is on the overall PP cost (z^{cost}) and the number of active PPs (z^{PP}) obtained for both traffic prioritization policies in optimized networks. Also, we present the relative difference in performance (denoted as Δ^{cost} and Δ^{PP}, respectively) between the policies. The results were obtained using the CPLEX MIP solver v.12.9 [20] run on a 3.7 GHz 32-core Threadripper-class machine with 128 GB RAM. We report that MIP solutions were optimal and the computations did not exceed 140 and 480 seconds in the most demanding scenario in RING-N and MESH-20, respectively.

In Fig. 5, we analyze the impact of the eMBB to URLLC bit-rate ratio (γ) on the performance metrics considered in the RING-8 network with $R = 60$ RUs and assuming $\alpha = 50\%$, i.e., the same number of RUs in the eMBB and URLLC slices. We can see that the FA policy results in a very high number of active PPs ($z^{\text{PP}} \geq 7$ in left chart), which means that either all or almost all PPs (8 are available in total) have to be used to support the RUs of both network slices. This is caused by large buffering delays of the whole FH traffic, without distinction on particular services, which leads to the placement of DU/CU entities of URLLC at the PPs located close to the RUs due to the low-latency requirements in the URLLC slice. The LA policy, which assigns the highest priority to the URLLC fronthaul flows, results in lower buffering latencies of the URLLC FH packets. This enables the placement of DU/CUs for the URLLC slice at more distant PP locations which, consequently, decreases the number of active PPs. In Fig. 5, such an effect is visible for $\gamma \geq 2$ (i.e., for the eMBB xHaul bit-rate exceeding twice the URLLC bit-rate) and it increases with γ up to $\gamma = 8$ for which the relative difference in z^{PP} between LA and FA reaches $\Delta^{\text{PP}} = 60\%$. For $\gamma \geq 9$, the amount of eMBB traffic is $\sqrt{9} = 3$ times the initial reference value (corresponding to $\gamma = 1$) and it is so high that the

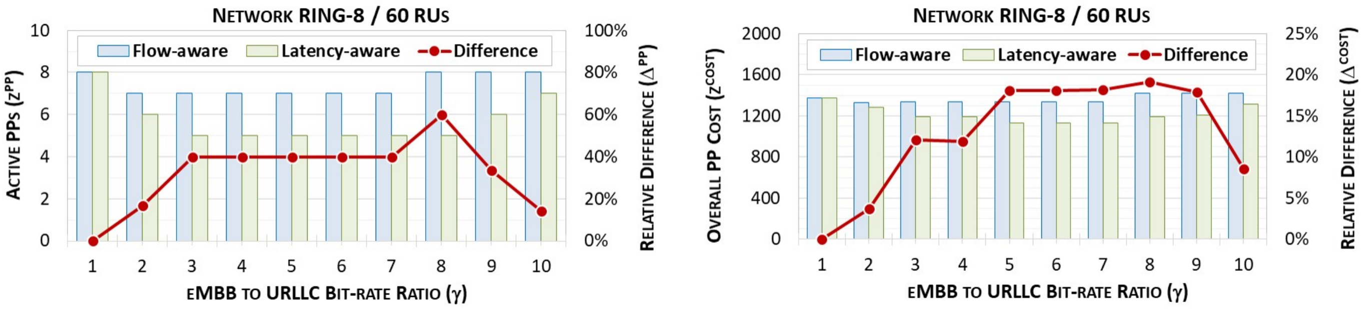

Fig. 5. Number of active PPs (bars in left chart), overall PP cost (bars in right chart), and relative difference in performance (lines) in a function of the eMBB to URLLC bit-rate ratio (γ) in network RING-8 with $R = 60$ RUs assuming $\alpha = 50\%$ of URLLC RUs

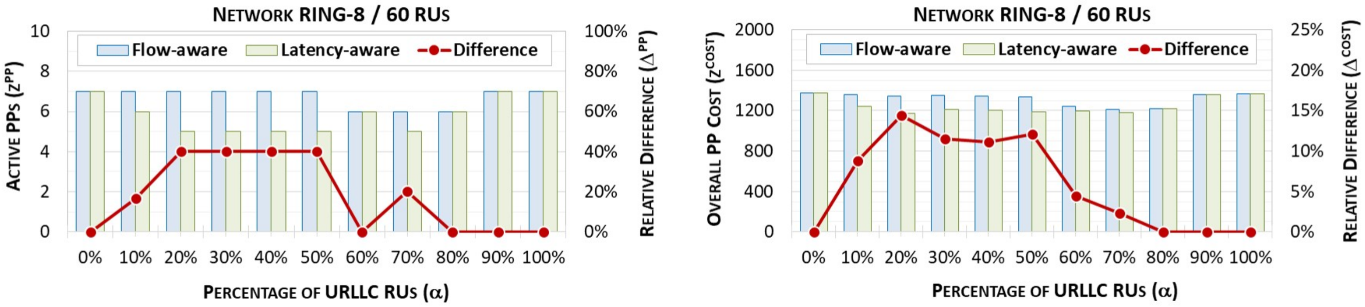

Fig. 6. Performance metrics in a function of the size of the URLLC slice (α) in network RING-8 with $R = 60$ RUs for $\gamma = 3$

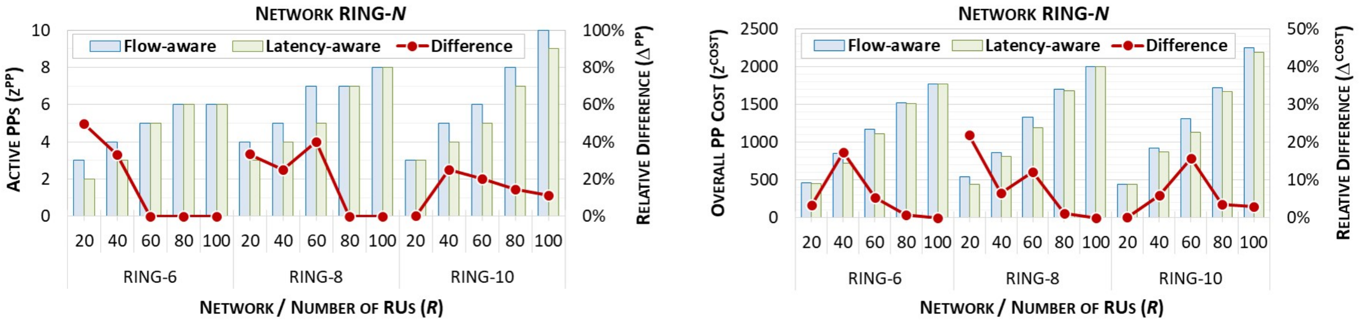

Fig. 7. Performance metrics in a function of the number of RUs (R) in networks RING-6, RING-8, and RING-10 assuming $\alpha = 50\%$ and $\gamma = 3$

activation of additional PPs is needed. The lower number of active PPs leads also to lower overall PP costs (z^{cost}), as shown in the right chart of Fig. 5, with the relative difference between both policies (Δ^{cost}) reaching up to $15\% - 20\%$ for $5 \leq \gamma \leq 9$.

In Fig. 6, we evaluate the impact of the size of the URLLC slice (expressed by parameter α) on the performance of network RING-8 with $R = 60$ and assuming $\gamma = 3$. We can see that the FA and LA policies offer the same performance for $\alpha = 0\%$ and $\alpha = 100\%$. It is obvious since in both cases only one type of slice exists in the network, respectively, either eMBB or URLLC, and there is no gain from assigning a higher priority to the low-latency URLLC flows. However, if both services are present, performance gains in terms of both z^{PP} and z^{cost} can be observed when using the LA policy. In particular, the relative differences between LA and FA reach up to $\Delta^{\text{PP}} = 40\%$ and $\Delta^{\text{cost}} \approx 10\% - 15\%$ for scenarios with $20\% \leq \alpha \leq 50\%$ of RUs belonging to the URLLC slice. For higher values of α, the amount of eMBB RUs and the volume of eMBB traffic decreases. As a result, the differences in performance diminish since the URLLC slice does not benefit from a higher priority of its flows.

In Fig. 7, we study the impact of network size on the performance of FA and LA. Namely, we consider network RING-N with different number of switching nodes (N) and RUs (R), assuming $\alpha = 50\%$ and $\gamma = 3$. In general, the higher number of RUs, the larger number of active PPs (z^{PP}) and higher PP cost (z^{cost}) can be observed. For $R \geq 80$ in MESH-6 and $R \geq 100$ in MESH-8, the networks are "saturated" with xHaul traffic and all PPs have to be used, which does not provide any significant gain from using the LA policy in these scenarios. However, for lower numbers of RUs, the differences between both policies appear and performance

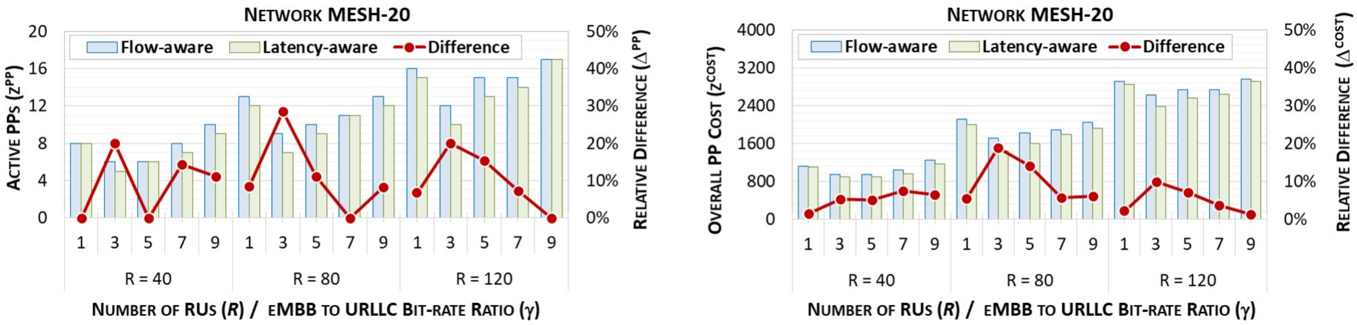

Fig. 8. Performance metrics in a function of the bit-rate ratio (γ) in network MESH-20 with different number of RUs (R) and assuming $\alpha = 50\%$

gains on the level of $25\% - 50\%$ for Δ^{PP} and $16\% - 22\%$ for Δ^{PP}, depending on particular network scenario, are achieved.

Finally, in Fig. 8, we present performance results in larger network MESH-20 with different number of RUs (R) and for different eMBB to URLLC bit-rate ratio values (γ), assuming the same size of both network slices ($\alpha = 50\%$). We can see that the highest differences in the FA and LA performance are achieved in a medium-size network (for $R = 80$), and they reach $\Delta^{\text{PP}} \approx 30\%$ and $\Delta^{\text{cost}} \approx 20\%$. In all cases, some gains in the use of PPs (z^{PP}) are observed for the LA policy, usually on the level of $1 - 2$ PPs saved. Also, similarly as in Fig. 5, the relative differences in performance increase up to some level, and afterwards tend to decrease with γ.

V. Concluding Remarks

We have focused on optimization of packet-switched 5G xHaul networks supporting convergent transport of the traffic flows related to network slices of different fronthaul latency requirements (e.g., such as eMBB and URLLC). To this end, we have analyzed the impact on network performance of two different xHaul traffic prioritization policies applied in packet switches, namely, flow-aware and latency-aware, in the xHaul network planning case study. By means of extensive numerical experiments in different network scenarios, we have shown that latency-aware prioritization of packets brings significant savings in terms of the number of active PP sites and the overall PP cost compared to the flow-aware approach. The gains depend highly on the xHaul bit-rate ratio of the services with different latency tolerance, the size of the lower-latency network slice (URLLC) with respect to the higher-latency slice (eMBB), as well as the network size. In particular, the savings in active PPs may reach up to $40\% - 60\%$ in ring scenarios and 30% in a mesh network, whereas the overall PP cost has been decreased by up to 20% for the cost values assumed in the analysis. In future work, we plan to focus on protection mechanisms in packet-switched xHaul networks.

References

[1] 3GPP, “Study on new radio access technology: Radio access architecture and interfaces,” Tech. Rep. TR 38.801, v14.0.0, Sophia Antipolis, France, 2017.

[2] ——, “Architecture description (release 17),” Tech. Spec. TS 38.401, v17.0.0, Sophia Antipolis, France, 2022.

[3] Y. Xiao, J. Zhang, and Y. Ji, “Can fine-grained functional split benefit to the converged optical-wireless access networks in 5G and beyond?” *IEEE Trans. Netw. Serv. Manag.*, vol. 17, no. 3, pp. 1774–1787, 2020. [Online]. Available: https://doi.org/10.1109/TNSM.2020.2995844

[4] IEEE, “IEEE standard for packet-based fronthaul transport networks,” https://standards.ieee.org/project/1914_1.html, (accessed on 28 September 2020). [Online]. Available: https://standards.ieee.org/project/1914_1.html

[5] J. Ordonez-Lucena *et al.*, “Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges,” *IEEE Comm. Mag.*, vol. 55, no. 5, pp. 80–87, 2017. [Online]. Available: https://doi.org/10.1109/MCOM.2017.1600935

[6] S. Vassilaras *et al.*, “The algorithmic aspects of network slicing,” *IEEE Comm. Mag.*, vol. 55, no. 8, pp. 112–119, 2017. [Online]. Available: https://doi.org/10.1109/MCOM.2017.1600939

[7] IEEE, “802.1cm-2018 – IEEE standard for local and metropolitan area networks – time-sensitive networking for fronthaul,” Nov. 2018.

[8] “Common public radio interface: eCPRI V2.0 interface specification,” 10 May 2019.

[9] 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. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2923020

[10] A. Esmaeily, K. Kralevska, and T. Mahmoodi, *Slicing Scheduling for Supporting Critical Traffic in Beyond 5G*. IEEE, Jan. 2022.

[11] J. Yusupov, A. Ksentini, G. Marchetto, and R. Sisto, “Multi-objective function splitting and placement of network slices in 5g mobile networks,” in *Proc. of IEEE CSCN*, Paris, France, Oct. 2018. [Online]. Available: https://doi.org/10.1109/CSCN.2018.8581714

[12] S. Bhattacharjee *et al.*, “Network slicing for TSN-based transport networks,” *IEEE Access*, vol. 9, pp. 62 788–62 809, 2021. [Online]. Available: https://doi.org/10.1109/ACCESS.2021.3074802

[13] M. Klinkowski, “Optimization of latency-aware flow allocation in NGFI networks,” *Comp. Commun.*, vol. 161, pp. 344–359, 2020. [Online]. Available: https://doi.org/10.1016/j.comcom.2020.07.044

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

[15] M. A. Imran, S. A. R. Zaidi, and M. Z. Shakir, *Access, Fronthaul and Backhaul Networks for 5G & Beyond*. Institution of Engineering and Technology, 2017. [Online]. Available: https://doi.org/10.1109/MCOM.2017.1600735

[16] H. Yu, F. Musumeci, J. Zhang, Y. Xiao, M. Tornatore, and Y. Ji, “DU/CU placement for C-RAN over optical metro-aggregation networks,” in *Proc. of ONDM*, Athens, Greece, May 2019. [Online]. Available: https://doi.org/10.1007/978-3-030-38085-4_8

[17] ITU-T Technical Report, “Transport network support of IMT-2020/5G,” Oct. 2018.

[18] B. M. Khorsandi and C. Raffaelli, “BBU location algorithms for survivable 5G C-RAN over WDM,” *Comput. Netw.*, vol. 144, pp. 53–63, 2018. [Online]. Available: https://doi.org/10.1016/j.comnet.2018.07.026

[19] 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.

[20] IBM, “CPLEX optimizer,” http://www.ibm.com/, (accessed on 30 September 2022). [Online]. Available: http://www.ibm.com/