Optimal resource preemption for aperiodic URLLC traffic in 5G Networks

Mira Morcos, Meriem Mhedhbi, Ana Galindo-Serrano, Salah Eddine Elayoubi

Optimal resource preemption for aperiodic URLLC traffic in 5G Networks

Mira Morcos[1], Meriem Mhedhbi[1], Ana Galindo-Serrano[1], and Salah Eddine Elayoubi[2]

[1]Orange Labs, Chatillon, France
[2]Laboratoire des Signaux et Systemes (L2S, CNRS), Centrale Supélec, Gif Sur Yvette, France

***Abstract*—In this paper, we study DownLink Preemption (DLP), a feature enabling dynamic scheduling for Ultra Reliable Low Latency Communications (URLLC) in the presence of ongoing enhanced Mobile Broad Band (eMBB) flows. We design a DLP technique with the objective of maximizing the admission of URLLC packets while minimizing the impact of preemption on the eMBB throughput. We propose two different approaches to solve the DLP problem. A first, the Offline Preemption Approach (OPA), is formulated as a multi-objective Integer Linear Program (ILP) and considers a perfect knowledge of traffic dynamics. We further propose the Impact-Aware Preemption Approach (IAPA), an algorithm solving the DLP problem and performing joint admission and preemption decisions on the fly. We conduct extensive simulations using a system level simulator in realistic 5G Network settings. Our numerical results demonstrate the efficiency of IAPA in guaranteeing a close-to-optimum performance.**

Keywords: 5G, URLLC, Downlink preemption, System level simulations.

I. Introduction

The design of the 5G New Radio (NR) targets the integration of critical Internet of Things (IoT) services on top of eMBB services without affecting their Quality of Service (QoS). In particular, Ultra Reliable Low Latency Communications (URLLC) is being specified as a novel type of services in the 3rd Generation Partnership Project (3GPP) standardization [1]. With its diverse domain of applications, URLLC offers a high potential market to mobile operators, such as industry 4.0 and power distribution.

Multiple features are being specified in the 3GPP standards to help the network meet with the URLLC requirements in terms of latency and reliability [2], [3]. These features affect both the Physical (PHY) and Medium Access Control (MAC) layers. At the PHY layer, flexibility in the frame structure configuration will help reduce the latency [4]. At the MAC layer we can distinguish two main features that will help in reducing the latency and providing finer scheduling granularity: 1) Mini-slot based scheduling which enables scheduling over a number of Orthogonal Frequency Division Multiplexing (OFDM) symbols that is equal to 2, 4 or 7 and 2) non-slot based scheduling which allows transmission over a fraction of a slot, starting at any OFDM symbol.

While exploiting the above-described PHY and MAC layer features, providing scheduling priority to the URLLC traffic is proven to reduce the achieved latency and also to ensure high reliability [4]. We can distinguish two main scheduling techniques that can privilege URLLC traffic: i) semi-persistent scheduling where URLLC packets are pre-scheduled using a reserved fraction of the bandwidth [5], [6] and ii) preemptive scheduling where both URLLC and eMBB are multiplexed on the same channel [7], [8]. Semi-persistent scheduling is shown to perform well with periodic traffic, yet it is less suited for sporadic traffic, due to its high impact on the spectrum efficiency and the eMBB throughput [9]. Preemptive scheduling makes a better candidate for sporadic traffic as it is more efficient in terms of spectrum usage and as its affects less the eMBB throughput.

Several studies have been conducted on the preemptive scheduling. In [10], the authors compare different preemption strategies to evaluate what eMBB resources are better to pre-empt; resources where URLLC packets can experience the best channel quality, or resources where eMBB packets experience good radio channel quality and thus can tolerate better the preemption. The authors in [11] propose a joint eMBB and URLLC scheduling technique satisfying instantaneous URLLC demands while maximizing utility for eMBB traffic. The impact on the eMBB is integrated in a utility function expressing the loss in eMBB throughput. In [12], the authors address the joint admission and scheduling problem for URLLC without evaluating the impact on the eMBB flow.

We develop in this paper a joint optimization framework for both URLLC and eMBB. We provide an efficient preemption policy to schedule aperiodic URLLC traffic, guaranteeing their requirements in terms of latency while reducing the impact on eMBB services. We specifically address the problem of: 1) what eMBB packets are best to interrupt? 2) how many Resource Blocks (RBs) are to preempt from a victim eMBB packet? We formulate the DLP as a problem solving the joint scheduling decision for URLLC packets and preemption decision for eMBB packets and taking into account link adaptation and interference constraints. We propose two different approaches to solve the DLP problem. We first propose the Offline Preemption Approach (OPA) formulated as an ILP with the multiple objectives of maximizing admission control over URLLC packets and minimizing the impact on the eMBB throughput. OPA takes into consideration the future arrivals of URLLC packets and leads to the optimal preemption solution. However, relying on the prediction of future arrivals is not realistic, and we propose the Impact Aware Preemption Approach (IAPA), based on an approximation algorithm performing scheduling

and preemption decisions on the fly. IAPA leads to a sub-optimal solution with respect to the one obtained with OPA. We compare the different preemption approaches by conducting an extensive simulation campaign, using a system level simulator in realistic network scenarios. Numerical results demonstrate the efficiency of IAPA in guaranteeing a close to optimum performance with respect to the one obtained with OPA.

The remainder of this paper is organized as follows. Section II introduces the system model and presents the preemption scenario. Section III shows the different problem formulations. Section IV illustrates and analyzes numerical results obtained from simulations. Finally Section V draws useful conclusions and recommendations for designing URLLC systems.

II. System model

In this section, we present the system model. We first describe the radio frame configuration and the traffic model. Finally, we present the preemption scenario and the assumptions we consider in the formulation of the DLP problem.

A. Frame configuration and latency model

We follow the NR frame configuration as specified in the 3GPP standards: we make use of the FDD band at 700 Mhz with a Sub-carrier spacing (SCS) of 15 kHz. As illustrated in Figure 1, a Transmission Time Interval (TTI) occupies $N = 14$ time units and the short TTI occupies $N_{short} = 2$ time units. A time unit has the duration of an OFDM symbol (OS). The transmission bandwidth B is divided into R RBs with 12 Resource Elements (REs) per RB.

We evaluate the delay induced at the PHY and MAC layers user plane latency which is the contribution of the radio network to the time from when the source sends a packet to when the destination receives and decodes it correctly. In other words, this is the one way duration it takes to successfully deliver a packet.

B. Traffic model

We model eMBB flows as a full buffer traffic and schedule eMBB packets according to Round Robin (RR) policy so that the selected packets are served continuously over the $N = 14$ time units. We assume that the bandwidth is divided into R RBs and are fully occupied by G eMBB packets and denote by $S_g = \{1, \ldots, G\}$ the corresponding set. A given packet g occupies R_g RBs (before the preemption). Let $S_g^r = \{1, \ldots, R_g\}$ be the set of RBs allocated to eMBB packet g.

The URLLC traffic arrives to the gNodeB (gNB) buffer following a Poisson traffic model and consists of short size packets of $L = 96$ information bits. We denote by $S_u = \{1, \ldots, U\}$ the set of URLLC packets ready to be transmitted in a given TTI. URLLC packet u is characterized by a 3-tuple: (1) t_u the time unit at which the packet is ready to be transmitted 2) d_u the amount of RBs required to schedule the packet, and (3) δ_u the minimum required Signal to Interference Noise Ratio (SINR). Parameters δ_u and d_u are linked to the Modulation and Coding Scheme (MCS) assigned to URLLC packet u; The MCS for a given URLLC packet and the number of RB required are calculated based on link quality measurements.

Figure 1 illustrates a subframe consisting one TTI with N OFDM symbols. A short TTI (represented in orange) can be of size 2 symbols. A given RB k can be allocated to URLLC packet u if and only if the SINR of RB k if allocated to packet u $sinr_u^k$, verifies $sinr_u^k \geq \delta_u$.

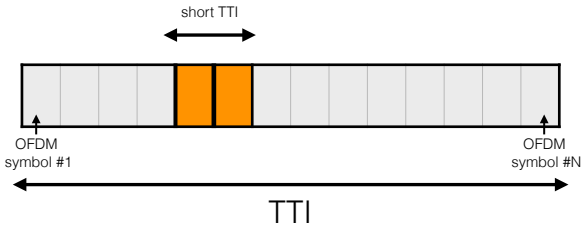

Figure 1: a Transmission Time Interval (TTI) divided into N time units. A short TTI can have the length of 2 OFDM symbols

C. Preemption

Recall that we schedule eMBB packets over $N = 14$ OFDM symbols and URLLC packets over $N_{short} = 2$ OFDM symbols. We denote by $S_n = \{1, \ldots, N\}$ the set of time units within a TTI. When a URLLC packet u is ready to be transmitted, the gNB will schedule the packet by puncturing resources from eMBB packets. An example of a preemption scenario is illustrated in Figure 2, where eMBB packets A and B occupy the total bandwidth. At the arrival of URLLC packet 1 at OFDM symbol 3, preemption happens by puncturing RBs from both packets A and B during the short TTI duration.

In this paper, for the sake of simplicity we refer to the RB at a given OS by *short RB*, as illustrated in Figure 2. We also assume that 1) preempting from a given eMBB packet is done by puncturing a given number of the RBs on certain time units, i.e., by puncturing short RBs; Referring to Figure 2, RBs 2 and 3 are punctured on OS 3,4,9 and 10 while RBs 1 is punctured on OS 9 and 10. (2) to schedule a given URLLC packet, it is possible to preempt from different packets; as shown in Figure 2, URLLC packet 1 is scheduled by preempting from both packets A and B. It is worth noting that, in case all RBs are preempted, at given OS, the URLLC packet will be delayed to the following one.

III. Problem definition and formulation

In this section we formulate the DLP problem using two approaches. We start with OPA, the offline preemption approach formulated using an ILP and leading to the optimal preemption solution. OPA cannot be applied in realistic scenarios, since it assumes the knowledge of future URLLC packets arrivals. To this end, we also propose the Impact Aware Preemption Approach (IAPA), a heuristic based on an approximation algorithm and leading to a sub-optimal solution compared to the one obtained with OPA.

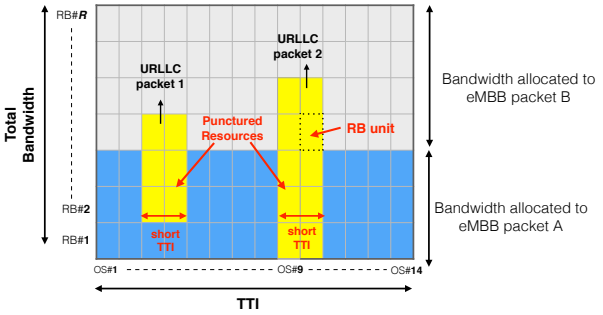

Figure 2: A preemption scenario example

Table I: Parameters description

Parameter	Description
N, S_n	Number of OFDM symbols in a slot and the corresponding set
U, S_u	Number of URLLC packets that need to be scheduled and, and the corresponding set
G, S_g	Number of eMBB packets and the corresponding set
R, S_g^r	Number of RBs allocated to packet g, and the corresponding set
d_u, t_u, δ_u	Number of RB that need to be preempted for URLLC packet u, arrival packets and $sinr$ threshold
N_{short}	Duration of the mini slot in terms of OFDM symbols
I_g^k	Impact on BLER when preempting k short RBs from packet g
$sinr_u^k$	SINR of RB k if allocated to URLLC packet u

A. *Optimal Puncturing Approach (OPA)*

We formulate the preemption problem using an ILP maximizing the admission control of URLLC packets and minimizing jointly the impact on the eMBB packets transmission and the URLLC packets latency. OPA performs admission control for URLLC packets and preemption decision for eMBB packet. For a given URLLC packet u, ready to be transmitted at a time unit $n \in S_n$, the scheduling decision is positive when there are d_u RBs to puncture. URLLC packet u, is dropped only when it exceeds a latency budget in the buffer. As for eMBB packets, preemption decision defines the packets impacted by the preemption and the number of short RBs they will be sacrificing.

We present in the following the decision variables we use in our approaches:

1) **Admission control for URLLC x_n^u:** is equal to 1 if URLLC packet u is admitted and to 0 otherwise. It is worth noting that URLLC packet u is admitted only if the system can preempt d_u RBs.
2) **RB allocation r_{un}^{gk}:** is equal to 1 if the $k-th$ RB, with $k \in S_g^r$, is preempted from eMBB packet g and allocated to URLLC packet u at time unit n, and to 0 otherwise. Having $sinr_u^k \geq \delta_u$ is a necessary condition to set r_{un}^{gk} to 1.
3) **Decision on the number of short RBs to puncture from a given eMBB packet y_g^r:** is equal to 1 if r RBs are preempted from eMBB packet g, and to 0 otherwise.

To assess the impact of preemption on eMBB packets, we take into consideration that the number of punctured short RBs affects directly the Block Error Rate (BLER) of packet g. I_g^k expresses the impact of packet g when preempting k short RBs. I_g^k is non-decreasing with respect to the number of short RBs punctured from eMBB packet g.

Table I details the parameters used in our optimization model. Our optimization model is formulated as follows:

$$\max \sum_{u \in S_u, n \in S_n} x_n^u \tag{1}$$

$$\min \sum_{g \in S_g, k \in S_g^r} I_g^k y_g^k + \sum_{u \in S_u, n \in S_n} (n - t_u) x_n^u \tag{2}$$

Subject to:

$$\sum_{g \in S_g, l \in S_g^r, n \in S_n, u \in S_u} r_{un}^{gl} \leq 1 \tag{3}$$

$$\sum_{n \in S_n} x_n^u \leq 1 \quad \forall u \in S_u \tag{4}$$

$$\sum_{k \in S_g^r} y_g^k \leq 1, \quad \forall g \in S_g \tag{5}$$

$$\sum_{l \in S_g^r, u \in S_u} r_{un}^{gl} \leq R_g \quad \forall g \in S_g, n \in S_n \tag{6}$$

$$\sum_{k \in S_g^r} k y_g^k \leq \sum_{l \in S_g^r, n \in S_n, u \in S_u} r_{un}^{gl}, \quad \forall g \in S_g \tag{7}$$

$$\sum_{\tau = min(n, N - N_{short})} x_{[n-\tau+1]}^u \leq \sum_{g \in S_g, k \in S_g^r : sinr_u^k \geq \delta_u} r_{un}^{gk}, \forall u \in S_u, n \in S_n \tag{8}$$

$$\sum_{g \in S_g, k \in S_g^r : sinr_u^k \geq \delta_u} r_{un}^{gk} = d_u \sum_{\tau = min(n, N - N_{short})} x_{[n-\tau+1]}^u, \forall u \in S_u, n \in S_n \tag{9}$$

Objective function (1) maximizes the admission of URLLC packets in the system. Objective function (2) minimizes jointly 1) the harm caused by the preemption on the BLER achieved by the eMBB packets and also 2) the delay of URLLC packets.

Constraint (3) ensures that a given RB is preempted at most once, at a given time unit n. Constraints (4) ensure that a given URLLC packet u is scheduled to start at one time unit at most. Constraint (6) ensures that the number of preempted RBs from a given eMBB packet at given time unit n, is not exceeded.

Expressions (5) and (7) represent the number of RBs punctured from a given eMBB packet at given time unit n. Finally, constraints (8) and (9) guarantees that a given URLLC packet u, if admitted, will be scheduled by puncturing d_u resources during N_{short} *consecutive* time units.

B. Impact-Aware Preemption Approach (IAPA)

We propose hereafter IAPA, an approach based on an approximation algorithm that can solve the DLP problem on the fly, without any knowledge about the future URLLC packets arrivals. As output, IAPA performs the admission decision of URLLC packets and the preemption decision of eMBB packets and leads to a sub-optimal solution, compared to the one obtained with OPA. IAPA is run when a given URLLC packet is ready to be transmitted and identifies eMBB packets to puncture and the number of short RBs to preempt from the victim packets.

We define in the following the notation we will use to describe IAPA:

- $SuitRBSet^u$: is the set of resources blocks that are suitable to be allocated to packet u at time units $t \in \{n, \dots, n + N_{short}\}$. A given RB k is suitable to be allocated to packet u if the RB is not preempted at any time units in $t \in \{n, \dots, n + N_{short}\}$ and if its SINR is higher than the threshold, i.e., if $sinr_k^u \geq \delta^u$.
- $Imp_g(y)$: represents the impact on the BLER of packet g induced by the preemption. This parameter depends from the number of short RBs y that are punctured from packet g.
- y_n^g: is the number of short RBs that are preempted at all time units preceding time unit n, i.e., at all time units $\in \{1, \dots, n\}$.
- $ImpSet_n$: is the set of the harm of all packets achieved by the preemption occurring on time units $\in \{1 \dots n\}$; $ImpSet_n = \{Imp_g(y = y_n^g)\}$.

Algorithm 1 is run whenever a URLLC packet is ready to be transmitted.

IAPA reads as follows: at time unit n, for a given URLLC packet u characterized by the tuple Θ_u=$\{d_u$,t_u,$\delta_u\}$, the set of suitable RBs $SuitRBSet^u$ is initialized . If the there is enough resources in $SuitRBSet^u$ to schedule packet u, then the packet is admitted. If not, the packet is delayed to the next time unit $(n + 1)$, only if the delay of packet u in the buffer does not exceed the latency budget, otherwise the packet is dropped. In case of admission, the step to follow is to decide what RBs to puncture at time units $n \in \{n, \dots, n + N_{short}\}$. This choice is done by choosing the packet that will be less affected when preempting from its RBs at this time unit. To achieve this, for each short RBs $\in \{1, \dots, d_u\}$ we do the following: we start by sorting the elements in $ImpSet_n$ in decreasing order. The packet to puncture from is the packet that will be less affected by the preemption and which index verifies: $argmin_g\{Imp_g(y = y_n^g)\}$. Now that the victim packet is identified, the algorithm selects the resource block to puncture. This resource block should verify $sinr_k^u \geq \delta^u$ and is thus selected from the set $SuitRBSet^u$.

Algorithm 1 Impact-Aware Preemption Approach (IAPA)

Data: Θ_u=$\{d_u$,t_u,$\delta_u\}$
Result: x_n^u, r_{un}^{gk}
initialization $SuitRBSet^u$
if $|SuitRBSet^u| \geq d_u$ **then**
  $x_n^u \leftarrow 1$
  for $l \leftarrow 1$ **to** d_u **do**
    Sort $ImpSet$
    int $P = argmin_g\{Imp_g(y = y_n^g)\}$
    int $k \in SuitRBSet^u \cap S_P^r$
    for $t \leftarrow n$ **to** N_{short} **do**
      $r_{ut}^{Pk} \leftarrow 1$
      Update y_t^g, $ImpSet_t$
    end
  end
else
  $x_n^u \leftarrow 0$ **end**

IV. Simulation Results

In this section, we evaluate numerically the preemption approaches we propose in this paper using a 5G NR system level simulator. We specifically quantify the achievable latency in DL scenario for the URLLC packets and also study the impact of the puncturing on the eMBB throughput.

In our simulation we compare the different preemption approaches:

- The Optimal Preemption Approach (OPA): described in section III-A, where we implemented our optimization model using the CPLEX commercial solver.
- The Impact Aware Approach (IAPA) we proposed in section III-B
- Random Preemption Approach (RPA), a benchmark algorithm, puncturing eMBB packets in a random fashion.

A. System parameters and scenario description

a) Network settings: We adopt the Frequency Division Duplex (FDD) transmission mode with bandwidth 10 MHz. We simulate a 2 tiers Urban Macro (UMa) scenario, i.e. 7 tri-sectored gNBs with Inter-Site-Distance (ISD) of 500 m with 10 URLLC users and 5 eMBB users (in average). We assume that the URLLC packets are short of size 96 bits. The URLLC UEs adopt the 4-QAM modulation and the enhanced turbo coding schemes with coding rates varying between 0 and 0.8 [13]. However, the eMBB User Equipments (UEs) use 4, 16 and 64 QAM and the Long Term Evolution (LTE) turbo coding and rate matching schemes. Each TTI and short TTI consists of 14 and 2 OFDM symbols, respectively.

Table II illustrates system parameters we deployed in our scenarios.

b) Latency model: We adopt the same delay model used in [14], where the radio plane latency is expressed as follows:

$$T_{radio} = T_{Tx} + T_{Alg} + T_{OT} + T_{Rx}$$

Table II: System parameters description

Parameter	eMBB	URLLC
Environment	3GPP Urban Macro (UMa)	
Carrier	10 MHz carrier bandwidth at 700 MHz (FDD)	
PHY numerology	15 kHz subcarrier spacing configuration	
TTI sizes	0.143 ms (2-symbol mini-slot)	1 ms (2-slots of 7-symbols)
bler target	1%	10%
Number of users	5 users per cell	10 users per cell
Traffic model	full buffer	Poisson
Scheduling	Proportional fair	Punctured

where, T_{Tx} represents the gNB processing delay and is equal to 7 OS. T_{Alg} is the frame alignment delay and represents the time gap between the moment the packet is ready to be transmitted and the actual transmission time. T_{OT} corresponds to the over the air transmission delay and is assumed to be equal to the short TTI duration, i.e., to 2 OS. Finally T_{Rx} is the receiver processing time. In case of Hybrid Automatic Repeat Request schemes (HARQ) re-transmission and additional T_{HARQ} delay is included.

B. Results discussion

1) Impact of puncturing on the eMBB throughput: Figures 3 and 4 illustrate the impact of puncturing on the eMBB performance obtained with OPA, IAPA and RPA in terms of throughput. Figure 3 depicts the cumulative distribution function (cdf) of the throughput experienced by an eMBB user for low and high URLLC loads. The cdf for IAPA is close to the one obtained with OPA, especially for eMBB users with low throughput (bad radio conditions). Figure 4 shows the $50-$percentile eMBB throughput obtained with the three approaches for different URLLC load; IAPA does better than RPA in guaranteeing a throughput 5 % less than the one obtained with OPA and 20% higher than the one obtained with RPA, with high URLLC packets load. The difference between IAPA and RPA is due to the fact that IAPA takes into account the impact of preemption on the eMBB.

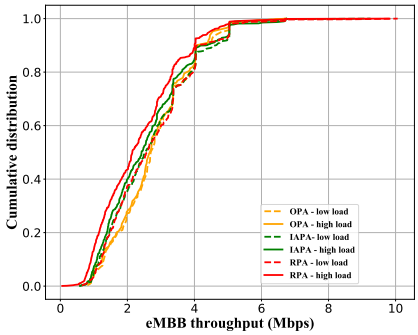

Figure 3: Cumulative distribution of eMBB packets throughput

Figure 5 plots the percentage of preempted eMBB packets decoded and the corresponding each range of re-transmission. With OPA, 97% of the preempted packets are decoded from the first attempt of transmission. The percentage obtained with IAPA is remarkably close to the one obtained with OPA, only 1% less, and 42% higher than RPA, where 37.3% of the preempted packets are not successfully decoded from the first attempt.

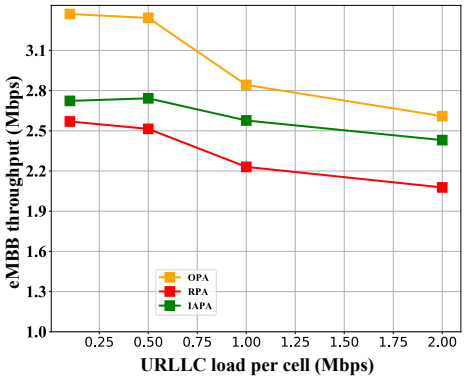

Figure 4: Average throughput of eMBB packets on the 50 percentile

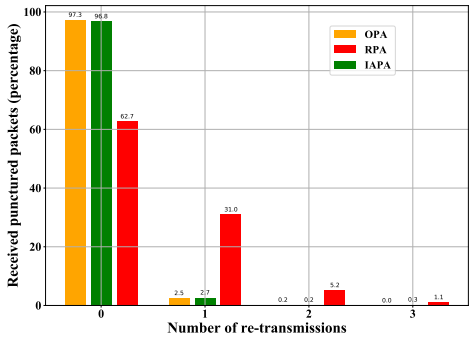

Figure 5: Number of retransmissions

2) URLLC packets achieved latency: Figure 6 illustrate the cdf of the received packets latency for OPA, RPA and IAPA. The 3 different approaches have a similar performance in terms of latency, except for very high percentiles. In particular, at 95 % the latency achieved is less than 0.6 ms. Figures 7 and 8 show the latency obtained at the 99% and $99,99\%$ percentiles, respectively. The performance of IAPA and RPA in terms of latency is similar, but OPA shows a better performance, especially when the URLLC packets load is high. This is due to the fact that OPA takes into account the future URLLC packet arrivals as well as the performance of eMBB packets when taking preemption decision.

If we now target a latency of 1 ms at $99,99\%$, only OPA is able to achieve the URLLC performance, while a 2 ms target is achievable for IAPA and RPA. This has two implications. Th first is that, for URLLC services whose latency target is not very tight, puncturing is a good solution as the eMBB users can be preserved, and IAPA achieves a good balance between URLLC and eMBB performances. For tight latency target of 1 ms, puncturing is not a good solution, as there is

no realistic algorithm that achieves the URLLC target while preserving eMBB, and resource reservation is a must.

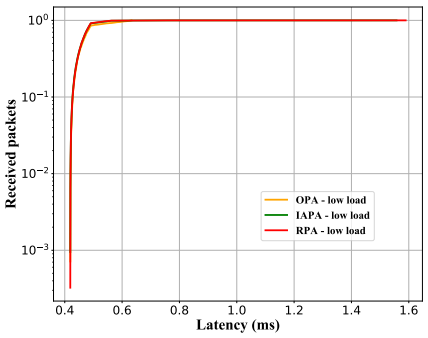

Figure 6: Cumulative distribution function of the URLLC packets achieved latency

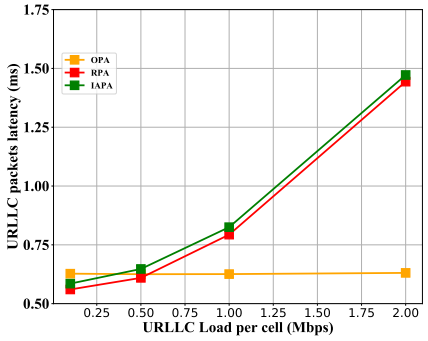

Figure 7: Latency of the URLLC packets on the 99-percentile

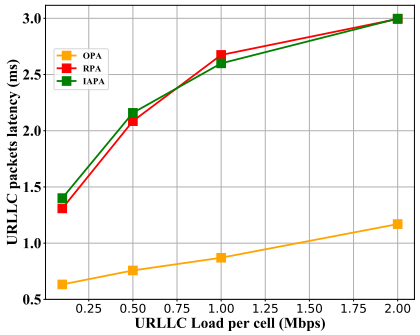

Figure 8: Latency of the URLLC packets on the $99,99$-percentile

V. Conclusion

We studied in this paper Donwlink preemption, to schedule URLLC packets on resources punctured from eMBB packets. We formulated the problem as a joint scheduling problem for URLLC packets and preemption problem for eMBB packets. We formulated the problem with the objective of maximizing the admission of URLLC packets and the impact on eMBB packets performance. We first proposed the Offline Preemption Approach (OPA), an technique based on a multi-objective optimization model maximizing the number of URLLC packets admitted while minimizing the impact on the eMBB packets performance. We also proposed the Impact aware preemption approach IAPA a technique based on an approximation algorithm.

We compared OPA and IAPA by implementing both approaches using a system level simulator in realistic network scenarios. Numerical results demonstrated the efficiency of IAPA that guarantees a 20% gain in terms of throughput with respect to a benchmark preemption policy, provided that the latency target of is not very tight. In scenarios with very tight URLLC constraint, preserving eMBB performance is possible for OPA, but difficult in practical online algorithms such as IAPA; puncturing alone is thus not sufficient and resource reservation for URLLC is to be considered.

References

[1] G. T. . V. (2017-09), "Study on New Radio Access Technology," September 2017.

[2] 3rd Generation Partnership Project, "Technical specification group services and system aspects, release 16 description, tr21.916," December 2019.

[3] 3rd Generation Partnership Project, "Technical specification group services and system aspects, release 15 description, tr21.915," March 2018.

[4] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, *et al.*, "Enabling technologies for ultra-reliable and low latency communications: from phy and mac layer perspectives," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2488–2524, 2019.

[5] G. T.-R. W. 86, "Semi-persistent scheduling for 5g new radio urllc," 2016.

[6] C. Li, J. Li, and W. Chen, "Adaptive ultra-reliable low-latency communications (urllc) semi-persistent scheduling," June 28 2018. US Patent App. 15/388,512.

[7] A. A. Esswie and K. I. Pedersen, "Null space based preemptive scheduling for joint urllc and embb traffic in 5g networks," in *2018 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, IEEE, 2018.

[8] S. R. Pandey, M. Alsenwi, Y. K. Tun, and C. S. Hong, "A downlink resource scheduling strategy for urllc traffic," in *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 1–6, IEEE, 2019.

[9] S. E. Elayoubi, P. Brown, M. Deghel, and A. Galindo-Serrano, "Radio resource allocation and retransmission schemes for urllc over 5g networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 4, pp. 896–904, 2019.

[10] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pp. 1–6, IEEE, 2017.

[11] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of urllc and embb traffic in 5g wireless networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 1970–1978, IEEE, 2018.

[12] A. Destounis and G. S. Paschos, "Complexity of urllc scheduling and efficient approximation schemes," in *2019 International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (Wiopt)*, IEEE, 2019.

[13] G. contribution R1-167414, "Enhanced Turbo Codes for NR: Performance Evaluation," August 2016.

[14] M. Mhedhbi, M. Morcos, A. Galindo-Serrano, and S. E. Elayoubi, "Performance evaluation of 5g radio configurations for industry 4.0," in *2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pp. 1–6, IEEE, 2019.