Joint Scheduling of Low-Latency and Best-Effort Flows in 5G Wireless Networks

T.R. Pijnappel
Eindhoven University of Technology
Eindhoven, The Netherlands

S.C. Borst
Eindhoven University of Technology
Eindhoven, The Netherlands

P.A. Whiting
Macquarie University
Sydney, Australia

***Abstract*—Emerging 5G networks will play a critical role in enabling a wide range of novel applications through Ultra-Reliable Low-Latency Communication (URLLC) capabilities besides offering even higher throughput to enhanced Mobile Broadband (eMBB) services. Supporting an increasingly heterogeneous set of performance requirements raises a strong need for innovative resource allocation mechanisms which achieve yet greater spectral efficiency and allow for highly agile transmissions, e.g. through puncturing of mini-slots. Motivated by these challenges, we introduce and analyze various joint scheduling schemes which aim to optimize throughput utility for eMBB flows while satisfying the delay requirements of URLLC flows and specifically accounting for channel variations over time and across frequencies. We show how the throughput dynamics of eMBB users can be described in the presence of puncturing of mini-slots for URLLC transmissions, and make several comparisons in terms of overall throughput performance and implementation complexity.**

***Index Terms*—5G, eMBB, URLLC, scheduling, puncturing, throughput utility optimization**

I. Introduction

Wireless service providers and equipment vendors have been accelerating their efforts towards large-scale deployment of 5G networks. These networks will introduce sophisticated physical-layer techniques (such as massive MIMO and mmWave communication) in order to boost capacity and improve throughput for enhanced Mobile Broadband (eMBB) services. Besides offering sheer capacity growth, 5G networks will also involve several MAC enhancements to provide superior connectivity for a wide range of novel applications enabling disruptive automation and intelligence (e.g. Industry 4.0, robotics, autonomous driving, tele-surgery, AR/VR) [1]–[3]. While these applications vary in nature, a crucial common feature is that they require extremely high reliability and low delay, broadly referred to as Ultra-Reliable Low-Latency Communication (URLLC), far exceeding the capabilities of current 4G LTE networks.

As alluded to above, supporting an increasingly diverse range of traffic categories such as eMBB and URLLC requires innovative MAC mechanisms and scheduling concepts. Specifically, 3GPP has proposed a puncturing framework [4], [5] where time is divided into slots with a duration ranging from one to two milliseconds. At the beginning of each slot decisions are made as to which of the available frequency-bands are assigned to each eMBB user (to transmit data). Since URLLC packets need to satisfy tight latency constraints, we cannot queue these until the next slot. Therefore, each eMBB slot is divided into mini-slots. Now upon arrival of URLLC packets, we can immediately schedule them for transmission during the next mini-slot by overriding/puncturing provisional eMBB transmissions. This is illustrated in Figure 1.

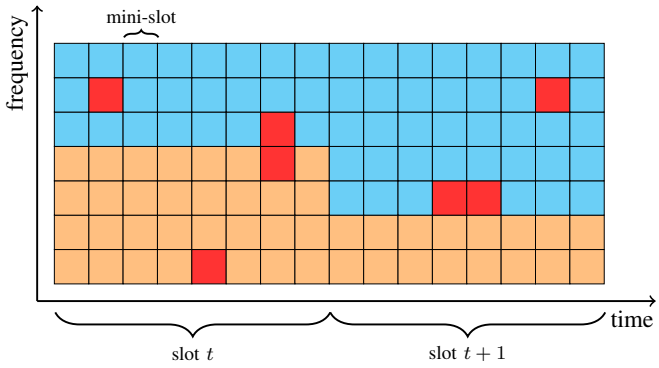


Fig. 1: Illustration of the assignment and puncturing of resource blocks. Blue and orange blocks are provisionally assigned to eMBB users 1 and 2 respectively at the beginning of that slot. Red blocks represent URLLC transmissions that puncture provisional eMBB transmissions during mini-slots.

The main challenge in this setting is to design an efficient joint scheduling method for eMBB and URLLC traffic. At the beginning of each slot we allocate resources to the eMBB users taking into account their channel states and a utility measure, with the goal to optimize this utility measure over a longer period of time. Meanwhile, at mini-slot level, we should decide which eMBB transmissions to puncture to satisfy the URLLC demands. Obviously, puncturing provisional eMBB transmissions during mini-slots affects the rates received by the eMBB users during that slot. This gives rise to a coupled problem over two time scales where we aim to optimize the joint scheduling of eMBB users on slots and the puncturing of the eMBB transmissions during mini-slots.

Motivated by the above challenges, we introduce and analyze in the present paper various schemes for allocating mini-slots to URLLC flows. While all these schemes aim to satisfy URLLC traffic demands fully and instantly (no queueing), they differ in the criteria used in deciding which specific mini-slots to puncture. Some schemes take into account the scheduling metrics for the various eMBB flows on the various frequencies, but follow different overall selection procedures,

ISBN 978-3-903176-29-4

while others do not consider the impact on eMBB flows at all and puncture in a blind way. The different levels of sophistication provide a trade-off between the implementation complexity and the degree of degradation in the throughput performance experienced by the eMBB flows.

In order to assess the latter performance implications, we extend the framework of [6] to allow for various forms of puncturing, and demonstrate how the throughput dynamics of the eMBB flows can be described in terms of a set of ordinary differential equations. In particular, we examine the limit point of these equations which provides an indication for the long-term throughput levels received by the eMBB flows. As a further performance benchmark, we also consider the optimal achievable throughput for the eMBB flows in the absence of any URLLC demands in an offline setting where full knowledge of the transmission rates is available in advance. This stringent optimality guarantee, in conjunction with a heuristic calculation of the minimum feasible throughput loss incurred from the puncturing, indicates that favorable performance can be achieved even with modest computational complexity.

To the best of our knowledge, the joint scheduling of eMBB and URLLC flows has received little attention in the literature so far, with the notable exception of [7]. The authors in [7] examined three different 'loss' models for capturing the impact of puncturing on the rates of eMBB flows, and the setup that we adopt somewhat resembles their 'linear' model. However, the spectral resources are essentially treated as a homogeneous continuum in [7], implying that only the amount of puncturing matters and that consideration of specific frequencies in allocating mini-slots does not play a significant role. In contrast, we explicitly account for the discrete nature of resource blocks (e.g. subbands) and frequency-selective fading properties.

The remainder of the paper is organized as follows. In Section II we present a detailed model description and introduce the basic setup for the scheduler operations. In Section III we demonstrate how the throughput dynamics of the best-effort flows can be described in terms of a set of ordinary differential equations. Section IV discusses the numerical and simulation experiments that we have conducted to illustrate the analytical results. In Section V we explore offline performance benchmarks. In Section VI we offer some brief concluding remarks and suggestions for further research.

II. Model description & resource allocation policies

A. System setup and scheduler operations

As mentioned earlier, we will consider a system where time is divided into slots of 1-2 milliseconds. In order to meet the low-latency constraints of URLLC packets, each time slot is further split into M equally sized mini-slots. The available radio spectrum is divided into a set of orthogonal frequency-bands denoted by $\mathcal{F}$. Each of the frequency-bands can only be used by one user at a time, and we do not explicitly consider multi-user MIMO transmissions.

The system supports a fixed population of eMBB users indexed by the set $\mathcal{E}$ and a fixed population of URLLC users labeled by the set $\mathcal{U}$. The eMBB users are assumed to have infinite amounts of data to be served. For the URLLC users we assume that the demands are independent from the other users, arrive according to some stochastic process and have a size $D_{j,t,m}$ bits where $j \in \mathcal{U}$ denotes the URLLC user, t the slot and m the mini-slot.

For each time slot we assume that estimates of the feasible transmission rates of all users on all frequency-bands are available to the scheduler. We will denote these rate estimates by $r_{i,f,t}$ (in bits/slot) where i denotes the user, f the frequency-band and t the corresponding slot.

Now the challenge is to determine good scheduling decisions on a slot-by-slot basis that maximize the throughput of the eMBB users, but at the same time 1) share the resources fairly among the eMBB users and 2) satisfy the demands of the URLLC users. For this we will first explain how the system keeps track of the throughputs of the eMBB users. In Section II-C we will elaborate how the demands of URLLC users are satisfied. In Section II-D we formulate the joint scheduling problem for which we introduce the scheduling policy of the eMBB users in Section II-E, and propose three puncturing policies in Section II-F.

B. Throughput tracking for eMBB users

To be able to track the evolution of the throughputs over time, we introduce decision variables $I^{\varepsilon}_{i,f,t}$ which equal 1 if frequency-band f is assigned to eMBB user i during slot t, and zero otherwise. For puncturing we introduce decision variables $J^{\varepsilon}_{j,f,t,m}$ which equal 1 if URLLC user j overwrites the provisional transmission on frequency-band f during mini-slot m of slot t, and zero otherwise. Now define the discounted throughput of user i up to time t (which for large t and small ε approximates the long-term average throughput $\theta_{i,t}$) as

$$\theta^{\varepsilon}_{i,t} = (1-\varepsilon)^t \theta^{\varepsilon}_{i,0} + \varepsilon \sum_{\tau=1}^{t} (1-\varepsilon)^{t-\tau} Y^{\varepsilon}_{i,\tau-1}, \qquad (1)$$

with

$$Y^{\varepsilon}_{i,\tau-1} = \sum_{f\in\mathcal{F}} r_{i,f,\tau} I^{\varepsilon}_{i,f,\tau} \left(1 - \frac{1}{M}\sum_{m=1}^{M}\sum_{j\in\mathcal{U}} J^{\varepsilon}_{j,f,\tau,m}\right), \quad (2)$$

where $\theta^{\varepsilon}_{i,0}$ is some initial value and ε is the discount factor (or smoothing parameter). Alternatively we can also write Equation (1) as

$$\theta^{\varepsilon}_{i,t+1} = \theta^{\varepsilon}_{i,t} + \varepsilon \left(Y^{\varepsilon}_{i,t} - \theta^{\varepsilon}_{i,t}\right), \qquad (3)$$

which is a useful recursion relation for implementation purposes in online scheduling algorithms. The term $\sum_{f\in\mathcal{F}} r_{i,f,\tau} I^{\varepsilon}_{i,f,\tau}$ in Equation (2) denotes the total throughput that eMBB user i would obtain during slot τ in the absence of puncturing. However, if some frequency-bands get punctured, we correct for this by multiplying with $1 - \frac{1}{M}\sum_{m=1}^{M}\sum_{j\in\mathcal{U}} J^{\varepsilon}_{j,f,\tau,m}$, which is the fraction of mini-slots that are not punctured.

C. Demand satisfaction for URLLC users

To make sure that all URLLC demands are instantaneously met, the set of constraints

$$\sum_{f\in\mathcal{F}} \frac{r_{j,f,t}}{M} J^{\varepsilon}_{j,f,t,m} \geq D_{j,t,m} \quad \forall_{j\in\mathcal{U}} \tag{4}$$

must be satisfied for all mini-slots m of all slots t. The factor $1/M$ is included in these constraints because $r_{j,f,t}$ is expressed in bits/slot instead of bits/mini-slot.

Since the URLLC demands are random in nature, it may not be possible to satisfy the constraints in Equation (4). In these circumstances, we assume that unsatisfied demands are lost and not backlogged for later mini-slots. However, in these scenarios, we can still assign all frequency-bands to the URLLC users in such a way that as many URLLC demands as possible are fulfilled. In the present paper we implicitly assume that a suitable admission control mechanism is in place to provide high levels of reliability to URLLC flows and ensure that the constraints in Equation (4) can be satisfied with probability close to one. Thus, we will not explicitly account for possible infeasibility of these constraints in the formulation of optimization problems.

D. Joint scheduling problem

To maximize the throughput of the eMBB users subject to fairness constraints, we will consider a Proportional Fair utility function, i.e. $U(\boldsymbol{\theta}^{\varepsilon}_t) = \sum_{i\in\mathcal{E}} \log(\theta^{\varepsilon}_{i,t})$ with $\boldsymbol{\theta}^{\varepsilon}_t = (\theta^{\varepsilon}_{1,t}, \ldots, \theta^{\varepsilon}_{|\mathcal{E}|,t})$, which is at the heart of most schedulers implemented in 4G/5G networks [8]–[10]. We aim to assign frequency-bands to various users so as to maximize the gain (minimize the loss) in utility during each slot, i.e. to maximize $U(\boldsymbol{\theta}^{\varepsilon}_t) - U(\boldsymbol{\theta}^{\varepsilon}_{t-1})$ see also for instance [11], [12]. Using a first-order Taylor expansion, we derive

$$U(\boldsymbol{\theta}^{\varepsilon}_t) - U(\boldsymbol{\theta}^{\varepsilon}_{t-1}) = \varepsilon \sum_{i\in\mathcal{E}} \frac{Y^{\varepsilon}_{i,t-1} - \theta^{\varepsilon}_{i,t-1}}{\theta^{\varepsilon}_{i,t-1}} + \mathcal{O}(\varepsilon^2). \tag{5}$$

In order to maximize (5), we need to maximize $\sum_{i\in\mathcal{E}} Y^{\varepsilon}_{i,t-1}/\theta^{\varepsilon}_{i,t-1}$, but at the same time satisfy all URLLC demands during slot t. Thus the optimal assignment for slot t is given by the following optimization problem

$$\max \quad \sum_{i\in\mathcal{E}} \sum_{f\in\mathcal{F}} \frac{r_{i,f,t} I^{\varepsilon}_{i,f,t} \left(1 - \frac{1}{M} \sum_{m=1}^{M} \sum_{j\in\mathcal{U}} J^{\varepsilon}_{j,f,t,m}\right)}{\theta^{\varepsilon}_{i,t-1}} \tag{6}$$

$$\text{subject to } I^{\varepsilon}_{i,f,t} \in \{0,1\} \quad \forall_{i\in\mathcal{E}}\ \forall_{f\in\mathcal{F}}, \tag{7}$$

$$J^{\varepsilon}_{j,f,t,m} \in \{0,1\} \quad \forall_{j\in\mathcal{U}}\ \forall_{f\in\mathcal{F}}\ \forall_{m\in\{1,\ldots,M\}}, \tag{8}$$

$$\sum_{i\in\mathcal{E}} I^{\varepsilon}_{i,f,t} \leq 1 \quad \forall_{f\in\mathcal{F}}, \tag{9}$$

$$\sum_{j\in\mathcal{U}} J^{\varepsilon}_{j,f,t,m} \leq 1 \quad \forall_{f\in\mathcal{F}}\ \forall_{m\in\{1,\ldots,M\}}, \tag{10}$$

$$\sum_{f\in\mathcal{F}} \frac{r_{j,f,t}}{M} J^{\varepsilon}_{j,f,t,m} \geq D_{j,t,m} \quad \forall_{j\in\mathcal{U}}\ \forall_{m\in\{1,\ldots,M\}}. \tag{11}$$

The objective function (6) serves to maximize $\sum_{i\in\mathcal{E}} Y^{\varepsilon}_{i,t-1}/\theta^{\varepsilon}_{i,t-1}$. Constraints (7) and (8) ensure that the decision variables $I^{\varepsilon}_{i,f,t}$ and $J^{\varepsilon}_{j,f,t,m}$ are indeed binary. Constraints (9) and (10) ensure that a frequency-band can be assigned to at most one eMBB or URLLC user during a slot and mini-slot respectively. To enforce the satisfaction of the URLLC demands we have Constraint (11). Note that scenarios where URLLC users puncture frequency-bands provisionally assigned to eMBB users are incorporated in the objective function (6).

E. Scheduling policy for eMBB users

As the random demands of URLLC users are only known at most a mini-slot in advance, we do not know the values $D_{j,t,m}$ in Constraint (11) at the time scheduling decisions for eMBB users have to be made, so that it is impossible to solve the problem given by (6)-(11). Therefore we schedule the eMBB users as if there are no URLLC demands, i.e. $D_{j,t,m} = 0$ for all $j \in \mathcal{U}$ and $m \in \{1, \ldots, M\}$. In this case, the optimal solution of (6)-(11) is given by

$$I^{\varepsilon}_{i,f,t} = \begin{cases} 1 & \text{if } i \in \arg\max\limits_{i\in\mathcal{E}} \dfrac{r_{i,f,t}}{\theta^{\varepsilon}_{i,t-1}}, \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

ties being broken arbitrarily (and $J^{\varepsilon}_{j,f,t,m} = 0$ for all j, f, t, m). In a scenario with one frequency, this assignment policy is exactly the Proportional Fair scheduling rule [6], [8]–[10].

F. Puncturing policies for URLLC users

We still need a method to decide which frequency-bands get punctured to satisfy URLLC demands. For that purpose, define the *score* of frequency-band f as $\max_{i\in\mathcal{E}} r_{i,f,t}/\theta^{\varepsilon}_{i,t-1}$, which equals M times the decrease of the objective function (6) when a frequency-band f is punctured during a mini-slot of slot t. In this paper we will consider the following puncturing policies:

Advanced Score-Based (ASB) policy: In order to maximize (6), given the values of $I^{\varepsilon}_{i,f,t}$ (eMBB scheduling decisions) we should choose the $J^{\varepsilon}_{j,f,t,m}$ (puncturing actions) so as to solve the following optimization problem in mini-slot m of slot t

$$\min \quad \sum_{f\in\mathcal{F}} \sum_{j\in\mathcal{U}} J^{\varepsilon}_{j,f,t,m} \max_{i\in\mathcal{E}} \frac{r_{i,f,t}}{\theta^{\varepsilon}_{i,t-1}} \tag{13}$$

$$\text{subject to } J^{\varepsilon}_{j,f,t,m} \in \{0,1\} \quad \forall_{j\in\mathcal{U}}\ \forall_{f\in\mathcal{F}}, \tag{14}$$

$$\sum_{j\in\mathcal{U}} J^{\varepsilon}_{j,f,t,m} \leq 1 \quad \forall_{f\in\mathcal{F}}, \tag{15}$$

$$\sum_{f\in\mathcal{F}} \frac{r_{j,f,t}}{M} J^{\varepsilon}_{j,f,t,m} \geq D_{j,t,m} \quad \forall_{j\in\mathcal{U}}, \tag{16}$$

which minimizes the aggregate score of the frequency-bands sacrificed for puncturing. We will refer to a policy that punctures based on (13)-(16) as Advanced Score-Based (ASB) policy. However, given the limited amount of time (less than a mini-slot) available to decide, it might not be feasible to find the optimal solution to this problem in real time.

Simplified Score-Based (SSB) policy: For a lower-complexity policy, which we will refer to as the Simplified Score-Based (SSB) policy, we denote the remaining demand of URLLC user j during mini-slot m of slot t by $\widehat{D}_{j,t,m}$. Then scheduling the URLLC demands for mini-slot m is done by executing the following steps:

1) Determine the frequency-band (that is not yet punctured) with the lowest score - say frequency f^*;
2) Assign frequency f^* to the URLLC user with the highest value of $\min\{r_{j,f,t}/M,\ \widehat{D}_{j,t,m}\}$;
3) If user j is chosen, then reduce $\widehat{D}_{j,t,m}$ by $\min\{r_{j,f,t}/M,\ \widehat{D}_{j,t,m}\}$;
4) If there is still a URLLC user with positive remaining demand and there is at least one frequency-band that is not punctured, then return to step 1. Otherwise terminate.

Random Puncturing (RP) policy: For comparison we will also consider a random puncturing (RP) policy. This policy executes the following steps during each mini-slot:

1) Check that there is at least one URLLC demand for the next mini-slot, if not terminate;
2) Select a frequency-band (that is not yet punctured) uniformly at random;
3) Assign the selected frequency-band to a URLLC user whose demand is not satisfied;
4) Check whether all URLLC demands are satisfied or all frequency-bands are punctured, if not return to step 2. Otherwise terminate.

III. Throughput dynamics

In this section we demonstrate how the throughput dynamics of the eMBB users in the presence of puncturing for URLLC transmissions can be characterized in terms of a set of differential equations. First we introduce some notation. Let ξ_t^ε denote the relevant history prior to time slot t, i.e. the feasible transmission rates, URLLC demands and scheduling/puncturing decisions during slots $\tau < t$. Now we can define the expected throughput vector during slot t as $\boldsymbol{g}_t^\varepsilon(\boldsymbol{\theta}_t^\varepsilon, \xi_t^\varepsilon) = \mathbb{E}\left[\boldsymbol{Y}_t^\varepsilon \mid \xi_t^\varepsilon\right]$. Furthermore we assume that there exists a continuous function $\bar{\boldsymbol{g}}(\cdot)$ such that for each fixed $\boldsymbol{\theta}$ we have

$$\lim_{k,n,\varepsilon} \frac{1}{k} \sum_{t=n}^{n+k-1} \mathbb{E}\left[\boldsymbol{g}_t^\varepsilon(\boldsymbol{\theta}, \xi_t^\varepsilon) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}) \mid \xi_t^\varepsilon\right] = 0 \qquad (17)$$

in probability, where $\lim_{k,n,\varepsilon}$ means that the limit is taken as $k \to \infty$, $n \to \infty$ and $\varepsilon \downarrow 0$ simultaneously in any way at all. The existence of the $\bar{\boldsymbol{g}}(\cdot)$ function does not require the URLLC demands and feasible transmission rates to be i.i.d. over time and is also ensured by mild conditions on the decay of time correlations. The component $\bar{g}_i$ of the function $\bar{\boldsymbol{g}}(x)$ can be interpreted as the expected long-term average throughput of user i when the Proportional Fair scheduling decisions are based on the observed discounted throughput of x. Lastly we define $\boldsymbol{\theta}^\varepsilon(t) = \boldsymbol{\theta}_n^\varepsilon$ for $t \in [n\varepsilon, n\varepsilon + \varepsilon)$ as the continuous-time interpolation of the process $\boldsymbol{\theta}_n^\varepsilon$.

Theorem 1. *Under suitable technical assumptions (listed and further discussed in the appendix), and recursion* (3)*, the process* $\{\theta^\varepsilon(t), k = 1, 2, \ldots\}$ *converges weakly to*

$$\boldsymbol{\theta}(t) = \boldsymbol{\theta}(0) + \int_0^t \left[\bar{\boldsymbol{g}}(\boldsymbol{\theta}(s)) - \boldsymbol{\theta}(s)\right]\ ds \qquad (18)$$

as $\varepsilon \to 0$.

For the proof of this theorem we refer to [13, Chapter 8]. The assumptions involved in Theorem 1 (as well as Theorems 2 and 3 below) are mostly technical in nature, and fairly mild from a practical viewpoint. It is worth observing that the structure of Equation (18) is similar to that in the absence of any puncturing, and that the impact of the URLLC transmissions on the throughput dynamics of the eMBB users is fully encapsulated by the function $\bar{\boldsymbol{g}}(\cdot)$.

Theorem 1 implies that for small values $\varepsilon > 0$, the values $\boldsymbol{\theta}_t^\varepsilon = (\theta_{i,t}^\varepsilon,\ i \in \mathcal{E})$ will be close to $\boldsymbol{\theta}(t)$. In particular, for small values $\varepsilon > 0$ and large t the value of $\boldsymbol{\theta}_t^\varepsilon$ should be close to a limit point (a point that the process visits infinitely often) of $\boldsymbol{\theta}(t)$. The next theorem shows that the latter limit point is unique and independent of the initial condition.

Theorem 2. *Under suitable technical assumptions (again listed in the appendix) and recursion* (3)*, the limit point* $\bar{\boldsymbol{\theta}}$ *of* (18) *is unique, and does not depend on the initial condition* $\boldsymbol{\theta}^\varepsilon(0)$*. In other words the process* $\boldsymbol{\theta}^\varepsilon(t)$ *converges to* $\bar{\boldsymbol{\theta}}$ *as* $\varepsilon \downarrow 0$ *and* $t \to \infty$.

The proof of this theorem and Theorem 3 (see below), can be found in the appendix.

Theorems 1 and 2, in conjunction with the fact that $\boldsymbol{\theta}^\varepsilon(t)$ is the continuous-time interpolation of the process $\boldsymbol{\theta}_n^\varepsilon$, imply that $\boldsymbol{\theta}_n^\varepsilon \approx \bar{\boldsymbol{\theta}}$ for small values of $\varepsilon > 0$ and large n. The next theorem further establishes that the ASB puncturing policy is optimal within a certain class of policies $\mathcal{A}$ in the sense that it maximizes the throughput utility enjoyed by the eMBB users corresponding to the limit point $\bar{\boldsymbol{\theta}}$. Specifically, the class $\mathcal{A}$ includes all policies that 1) make scheduling decisions without any advance knowledge of the URLLC demands and upcoming puncturing actions, 2) satisfy the full URLLC demands whenever possible and 3) puncture frequency-bands only based on the feasible transmission rates and the observed discounted throughput at the beginning of the slot. In particular, policies in $\mathcal{A}$ cannot use information about puncturing decisions or URLLC demands that already occurred during the same slot. We chose to exclude the latter options, because of the added computational complexity.

Theorem 3. *Under suitable technical assumptions (again listed in the appendix) and recursion* (3)*, scheduling the eMBB users according to Equation* (12) *and puncturing using the ASB policy yields a limiting throughput* $\bar{\boldsymbol{\theta}}$ *such that there is no policy in* $\mathcal{A}$ *that yields limiting throughput* $\widetilde{\boldsymbol{\theta}} \neq \bar{\boldsymbol{\theta}}$ *with* $U(\widetilde{\boldsymbol{\theta}}) \geq U(\bar{\boldsymbol{\theta}})$.

As mentioned above, the impact of puncturing on the throughput dynamics of the eMBB users is entirely captured

by the function $\bar{\boldsymbol{g}}(\cdot)$. This function depends on the statistical properties of the channel rate processes $r_{i,f,t}$ in an intricate way, and it is in general difficult to determine in an explicit form. We now show that $\bar{\boldsymbol{g}}(\cdot)$ can be expressed in closed form however in a scenario with Rayleigh fading (Jakes model [14]). For this scenario, let $s_{i,f}$ denote the average transmission rate that user i can receive using frequency-band f. A lengthy derivation yields that the function $\bar{g}_i(\cdot)$ takes the form

$$\bar{g}_i(\boldsymbol{\theta}) = \sum_{f\in\mathcal{F}} \mathbb{P}\left(H_f\right) s_{i,f} \sum_{J\in Q_i} (-1)^{|J|}\, h_i(\boldsymbol{s}, f, \boldsymbol{\theta}, J), \quad (19)$$

where

$$h_i(\boldsymbol{s}, f, \boldsymbol{\theta}, J) = \left(\frac{\theta_i}{\theta_i + s_{i,f}\sum_{k\in J}\frac{\theta_k}{s_{k,f}}}\right)^2, \quad (20)$$

$\boldsymbol{s} = \{s_{i,f} \mid i \in \mathcal{E},\ f \in \mathcal{F}\}$ and Q_i is the collection of all subsets of $\mathcal{E}\backslash\{i\}$. The factor $\mathbb{P}(H_f)$ represents the probability that frequency-band f is *not* punctured and accounts for the loss in throughput experienced by eMBB users due to punctured transmissions. This probability can be determined based on the puncturing rule, the channel rate processes and the demand processes of the URLLC users, but this involves a lengthy calculation which is omitted because of page limitations.

IV. Numerical results

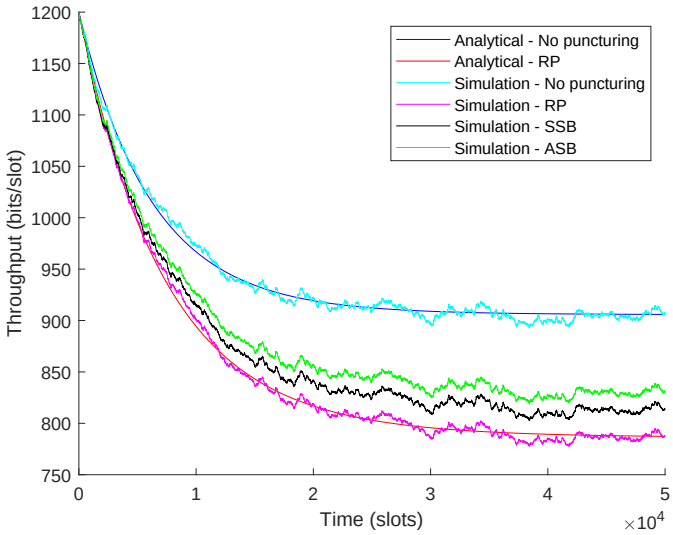
In order to corroborate Theorems 1 and 2, we consider a small scenario with three frequency-bands, two eMBB users, and one URLLC user. The users have average feasible transmission rates $s_{i,f}$ given in Table I.

Frequency-band	1	2	3
eMBB user 1	300	400	500
eMBB user 2	700	800	600
URLLC user	600	500	400

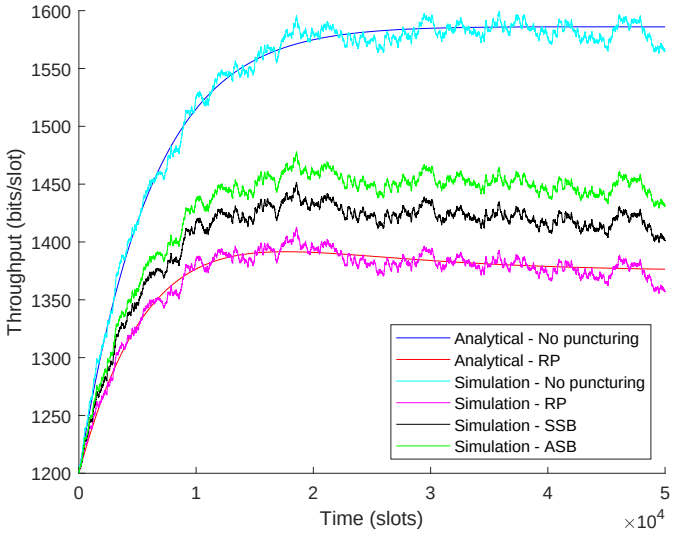
TABLE I: Average feasible transmission rates (in bits/slot) of the three users over each of the three frequency-bands.

As stated in Section II, we assume that both eMBB users have full buffers. We further suppose that the URLLC user generates traffic bursts with a constant size of 70 bits as a Bernoulli process, with probability 0.2 independently during each mini-slot.

We assume a slot duration of 1 millisecond, with 8 equally sized mini-slots. For the Rayleigh fading we use a maximum Doppler shift of 60Hz. For our simulation we simulate 50000 slots, with smoothing parameter $\varepsilon = 0.0001$ and choose the initial throughputs $\boldsymbol{\theta}_0 = (1200, 1200)$ bits/slot. Results of this simulation are shown in Figure 2, along with analytical results for cases with and without random puncturing obtained as discrete-time approximations of the ODE given by Equation (18), where $\bar{\boldsymbol{g}}(\cdot)$ is defined by Equation (19). As we can see in Figure 2, the simulation results match the discrete-time approximations of the ODEs quite well. We note that there is some fluctuation around the analytical curves reflecting the inherent randomness of the channel rate processes and the traffic demands of the URLLC users which through



(a) User 1



(b) User 2

Fig. 2: Sample paths of the throughput functions $\theta_{i,t}$, and discrete-time approximations for the scenario with and without random puncturing.

puncturing also impacts the throughputs of the eMBB users. Other results (omitted because of space limitations) indicate that the fluctuation diminishes for lower smoothing parameter values ε and higher fading frequencies. In all simulations that we considered, the convergence time of the smoothed throughputs is of the order $1/\varepsilon$ slots. Furthermore, we see that the SSB policy outperforms the RP policy as we would expect since it is more sophisticated. The SSB policy even performs nearly as well as the ASB policy while being much simpler to implement. Recall that the ASB policy is in fact asymptotically optimal within the class $\mathcal{A}$ as stated in Theorem 3.

V. Offline performance benchmarks

We already established that the ASB policy is asymptotically optimal within the class $\mathcal{A}$, and observed that the lower-complexity SSB policy performs nearly as well. In this section we continue to restrict to policies that satisfy the full URLLC demands whenever possible, but relax the other two restrictions on the policies in the class $\mathcal{A}$. Specifically, we widen the comparison to policies that may make scheduling decisions based on upcoming puncturing actions, are allowed full freedom in making puncturing decisions, and may have full knowledge of all feasible rates and URLLC demands in advance. This is conceptually similar in spirit to the offline performance benchmarks considered in [15]–[17] but pertains to a different problem.

We continue to consider the Proportional Fair utility function, thus aiming to maximize the aggregate logarithmic throughput utility. However, we now use the time-average throughput over a finite time horizon of T time slots, defined for user i as

$$\bar{R}_i = \frac{1}{T}\sum_{t=1}^{T}\sum_{f\in\mathcal{F}} r_{i,f,t} I^{\varepsilon}_{i,f,t}\left(1-\sum_{m=1}^{M}\sum_{j\in\mathcal{U}}\frac{J^{\varepsilon}_{j,f,t,m}}{M}\right). \quad (21)$$

Furthermore, we observe that maximizing the sum of the logarithms of the throughputs is equivalent to maximizing the product of the throughputs, and will compare various policies in terms of the ratio of $\sqrt[|\mathcal{E}|]{\prod_{i\in\mathcal{E}}\bar{R}_i}$ as a proxy for the relative difference in the achieved throughput on a per-user basis.

Even in the offline setting it will be difficult to determine the optimal assignment for a scenario with puncturing. Therefore, we derive the optimal assignment for a scenario without URLLC users and then use that to construct an upper bound for the scenario where URLLC users are present.

A. Optimal assignment without URLLC users

When we do not consider any URLLC users in the offline setting, we obtain the following optimization problem:

$$\max \ \sum_{i\in\mathcal{E}}\log\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{f\in\mathcal{F}} r_{i,f,t} I^{\varepsilon}_{i,f,t}\right) \quad (22)$$

$$\text{subject to } \sum_{i\in\mathcal{E}} I^{\varepsilon}_{i,f,t}\le 1 \quad \forall_{f\in\mathcal{F}}\ \forall_{t\in\{1,\dots,T\}}, \quad (23)$$

$$I^{\varepsilon}_{i,f,t}\in\{0,1\} \quad \forall_{i\in\mathcal{E}}\ \forall_{f\in\mathcal{F}}\ \forall_{t\in\{1,\dots,T\}}. \quad (24)$$

To simplify this integer linear program, we consider the continuous relaxation, i.e. we replace Constraint (24) by

$$I^{\varepsilon}_{i,f,t}\ge 0 \quad \forall_{i\in\mathcal{E}}\ \forall_{f\in\mathcal{F}}\ \forall_{t\in\{1,\dots,T\}}. \quad (25)$$

The objective value of the relaxation provides an upper bound to our original problem. The next lemma shows that the optimal solution of the relaxed problem has a specific structure.

Lemma 1. *The optimal solution to the optimization problem with objective function* (22)*, and Constraints* (23) *and* (25) *satisfies the Constraints* (23) *with equality. Furthermore, there exists an optimal solution with at least* $|\mathcal{E}||\mathcal{F}|T-|\mathcal{F}|T-|\mathcal{E}|+1$ *variables* $I^{\varepsilon}_{i,f,t}=0$*, at least* $|\mathcal{F}|T-|\mathcal{E}|+1$ *variables* $I^{\varepsilon}_{i,f,t}=1$*, and at most* $2|\mathcal{E}|-2$ *variables* $I^{\varepsilon}_{i,f,t}\in(0,1)$.

The proof can be found in the appendix.

B. Heuristic upper bound accounting for URLLC users

While the scenario without puncturing provides an absolute performance benchmark, we need to account for the puncturing to obtain a more informative yardstick. To do this, we define F_f as the contribution of frequency f to the total throughput, so that

$$\mathbb{E}[F_f] = \sum_{i\in\mathcal{E}}\mathbb{E}\left[X_{i,f}\mathbb{1}_{\left\{\frac{X_{i,f}}{\theta_i}\ge\frac{X_{j,f}}{\theta_j},i\neq j\right\}}\right], \quad (26)$$

with $X_{i,f}$ representing the feasible transmission rate of user i on frequency-band f. Now note that at least one frequency-band gets punctured when there is a URLLC demand. But we can also determine the probability q that we must puncture at least two frequency-bands to satisfy the URLLC demand. So assuming that there is a URLLC demand with probability p, a heuristic lower bound for the total loss in eMBB throughput can be found by solving the following problem

$$\min \ L=\sum_{f\in\mathcal{F}} a_f\mathbb{E}[F_f] \quad (27)$$

$$\text{subject to } \sum_{f\in\mathcal{F}} a_f = p(1+q), \quad (28)$$

$$a_f\in[0,p] \quad \forall_{f\in\mathcal{F}}, \quad (29)$$

where the first constraint ensures that we puncture a fraction $p(1+q)$ of the resources, which is a lower bound for the amount of resources that we expect to puncture. The second constraint follows from the observation that we expect to puncture a fraction p of the mini-slots. Using this heuristic lower bound for L, we can determine a heuristic upper bound for the performance measure by solving the following problem

$$\max \ \sqrt[|\mathcal{E}|]{\prod_{i\in\mathcal{E}} R_i/\widetilde{R}_i} \quad (30)$$

$$\text{subject to } \sum_{i\in\mathcal{E}}\widetilde{R}_i-\sum_{i\in\mathcal{E}}R_i\ge L, \quad (31)$$

$$R_i\in[0,\widetilde{R}_i] \quad \forall_{i\in\mathcal{E}}, \quad (32)$$

where the $\widetilde{R}_i$ denote the rates corresponding to the upper bound in the scenario without puncturing. The first constraint ensures that we incur at least the minimum loss in total throughput, and the second constraint ensures that none of the users will receive a higher average rate than in an offline setting without puncturing.

Since we use expected values for this heuristic upper bound, it could happen that performance measure of one of the puncturing policies attains a higher value than this upper bound. However, as we also use lower bounds to determine the expected total loss in throughput L, the performance measure is unlikely to exceed the upper bound. This also reflects that the heuristic upper bound is in general not attainable, even by the optimal puncturing policy.

C. Performance comparison and discussion

To illustrate the performance of the various policies, we use the same example as before (average rates given in Table I, $\varepsilon = 0.0001$, $\boldsymbol{\theta}_0 = (1200, 1200)$ and $T = 50000$ slots), but now we assume that the URLLC user has a demand of 70 bits during a mini-slot with probability $p \in [0, 1/2]$. To obtain reasonably accurate results, we average the results of 100 simulations for each value of p. Results of these simulations can be found in Figure 3.

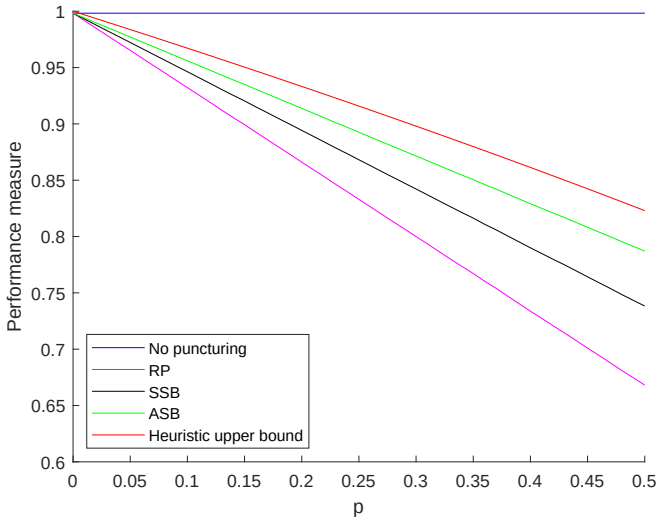


Fig. 3: Average performance measures $\sqrt[|\mathcal{E}|]{\prod_{i\in\mathcal{E}} \bar{R}_i/\widetilde{R}_i}$ of the online scenarios as function of p, and a heuristic upper bound for the puncturing scenarios.

We observe that the throughputs of the eMBB users decrease with the amount of URLLC demands. The results further demonstrate that across the entire range the SSB policy distinctly outperforms the RP policy and comes reasonably close to the ASB policy, as already noted in the previous section for $p = 0.2$.

The performance gap between the ASB policy and the heuristic upper bound can be explained by the fact that we used a lower bound for the expected total loss in throughput. However, even compared to this heuristic upper bound, the performance gap for our example is less than 5%.

VI. Conclusions

In the present paper we introduced and analyzed joint scheduling schemes for 5G networks which involve puncturing of mini-slots to allow highly agile transmissions to URLLC flows. The proposed schemes take into account the scheduling metrics for the various eMMB flows so as to minimize the loss in throughput utility from puncturing while satisfying the delay requirements of URLLC flows. We showed that the throughput trajectories of the eMBB flows, when properly scaled, converge to the solution of a set of differential equations that account for the impact of the puncturing. In order to corroborate the analytical results, we demonstrated close agreement between the numerical solution of the differential equations and simulation experiments.

We leveraged these results to compare the throughput performance of various puncturing policies with varying degrees of implementation complexity. The Advanced Score-Based (ASB) policy offers the best throughput performance and is in fact asymptotically optimal under mild conditions, but is computationally demanding. The Simplified Score-Based (SSB) policy has lower computational complexity, and suffers some throughput degradation compared to the ASB policy, but still notably outperforms the Random Puncturing (RP) policy.

As a further performance benchmark, we also considered the optimal achievable throughput for the eMBB flows without any URLCC demands in an offline setting with full advance knowledge of the transmission rates. This absolute performance bound, combined with a simple approximation of the minimum feasible throughput loss incurred from the puncturing, suggested that the ASB policy is not far from optimal clairvoyant performance. By extension we conclude that the SSB policy provides a reasonable low-complexity heuristic. A challenging topic that remains for further research is to establish a rigorous and tight performance bound, for example by characterizing the optimal throughput in an offline setting in the presence of puncturing.

References

[1] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, and B. Vucetic. Ultra-reliable low latency cellular networks: Use cases, challenges and approaches. *IEEE Commun. Mag.* **56(12)**:119–125, December 2018.

[2] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim. Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects. *IEEE Wireless Commun.* **25(3)**:124–130, June 2018.

[3] C. Li, J. Jiang, W. Chen, T. Ji, and J. Smee. 5G ultra-reliable and low-latency systems design. In *Proc. European Conf. Netw. Commun. (EuCNC)*, pages 1–5, June 2017.

[4] 3GPP TSG RAN WG1 Meeting 87, November 2016.

[5] 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 *Proc. 86th IEEE VTC-Fall Conf.*, pages 1–6, Sep 2017.

[6] H.J. Kushner and P.A. Whiting. Convergence of proportional-fair sharing algorithms under general conditions. *IEEE Trans. Wireless Commun.* **3(4)**:1250–1259, July 2004.

[7] A. Anand, G. de Veciana, and S. Shakkottai. Joint scheduling of URLLC and eMBB traffic in 5G wireless networks. *Proc. IEEE INFOCOM 2018* pages 1970–1978, 2018.

[8] J.M. Holtzman Asymptotic analysis of Proportional Fair sharing. *Proc. IEEE PIRMC '01,* 33-37.

[9] A. Jalali, R. Padovani, R. Pankai Data throughput of CDMA HDR - a high efficiency data rate personal communication wireless system. *Proc. IEEE VTC '00 Spring,* 1854-1858.

[10] P. Viswanath, D.N.C. Tse, R. Laroia Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277-1294, June 2002.

[11] A.L. Stolyar. On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation. *Oper. Res.* **53(1)**:12–25, 2005.

[12] R. Agrawal and V. Subramanian. Optimality of certain channel-aware scheduling policies. In *Proc. 40th Annual Allerton Conf. Commun., Contr., Comput.* 2002

[13] H.J. Kushner and G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer New York, 1997.

[14] W.C. Jakes. *Microwave Mobile Communications*. IEEE Press classic reissue. IEEE Press, 1974.

[15] V. Joseph and G. de Veciana. Jointly optimizing multi-user rate adaptation for video transport over wireless systems: Mean-fairness-variability tradeoffs. In *Proc. IEEE INFOCOM 2012*, pages 567–575, March 2012.
[16] V. Joseph and G. de Veciana. Variability aware network utility maximization. *CoRR*, abs/1111.3728, 2011.
[17] V. Joseph, G. de Veciana, and A. Arapostathis. Resource allocation: Realizing mean-variability-fairness tradeoffs. In *Proc. 50th Annual Allerton Conf. Commun., Contr., Comput.*, pages 831–838, Oct 2012.

VII. Appendix

A. Technical assumptions Theorem 1

The technical assumptions for Theorem 1 are:

1. The rates $\{r_{i,f,n} : i \in \mathcal{U} \cup \mathcal{E}, f \in \mathcal{F}\}$ are bounded;
2. $\boldsymbol{g}_t^\varepsilon(\boldsymbol{\theta}_t^\varepsilon, \xi_t^\varepsilon)$ is continuous in $\boldsymbol{\theta}$ uniformly in t, ε and ξ_t^ε;
3. ξ_t^ε is defined on a compact sequence space.

To see that these assumptions are sufficient, we denote the assumptions in [13] by A1.x and show that all these assumptions are met. Assumption 1 covers A1.1, A1.8 and A1.10. Assumption 2 ensures that A1.2, A1.5 and A1.6 are met, but it also implies that A1.4 is met. Assumption 3 implies that A1.7 is met. A1.3 and A1.9 are implied by the definition of $\bar{\boldsymbol{g}}(\cdot)$.

B. Technical assumptions Theorems 2 and 3

The technical assumptions for Theorem 2 are:

1. The same technical assumptions as for Theorem 1;
2. $\dot{\boldsymbol{\theta}} = \bar{\boldsymbol{g}}(\boldsymbol{\theta}) - \boldsymbol{\theta}$ satisfies the K-condition (also known as Kamke-condition) and is Lipschitz continuous;
3. Subject to some constraints, the assignment of the eMBB users (the $I^\varepsilon_{i,f,t+1}$) is such that it maximizes
$$\sum_{i\in\mathcal{E}} \frac{\sum_{f\in\mathcal{F}} r_{i,f,t+1} I^\varepsilon_{i,f,t+1} - \theta^\varepsilon_{i,t}}{\theta^\varepsilon_{i,t}}, \tag{33}$$
and the puncturing of the URLLC users $J^\varepsilon_{j,f,t+1,m}$ does not depend on the $I^\varepsilon_{i,f,t+1}$ variables.

Assumptions 1 and 2 are the same as in [6], but we also need Assumption 3 to incorporate puncturing policies.

For Theorem 3 we drop the third assumption, as there is already a statement about the assignment policy in the formulation of the theorem.

C. Proof Theorems 2 and 3

Note that the evolution of the throughput variables $\theta^\varepsilon_{i,t}$ in Equation (3) is identical to the case without puncturing; the only difference lies in the additional puncturing variables $J^\varepsilon_{j,f,t,m}$ in the expression for $Y^\varepsilon_{i,t}$ in Equation (2).

For Theorems 2 and 3 and in view of the proof arguments in [6], it suffices to show that the $Y^\varepsilon_{i,t}$ variables satisfy a similar relationship as used there, namely $\sum_{i\in\mathcal{E}} Y^\varepsilon_{i,t}/\theta^\varepsilon_{i,t} \geq \sum_{i\in\mathcal{E}} \widetilde{Y}^\varepsilon_{i,t}/\theta^\varepsilon_{i,t}$ where the $\widetilde{Y}^\varepsilon_{i,t}$ variables correspond to a second assignment policy (denoted by $\widetilde{I}^\varepsilon_{i,f,t}$ and $\widetilde{J}^\varepsilon_{j,f,t,m}$) subject to the same set of constraints, but with limit point $\widetilde{\boldsymbol{\theta}}$.

For Theorem 2, we want to show that a specific assignment policy has only one limit point, so we can assume that the second assignment policy punctures in the same way, i.e. $\widetilde{J}^\varepsilon_{j,f,t,m} = J^\varepsilon_{j,f,t,m}$. Now similar arguments as in [6] give the desired inequality.

For Theorem 3, we notice that the assignment policy stated in the theorem maximizes $\sum_{i\in\mathcal{E}} Y^\varepsilon_{i,t}/\theta^\varepsilon_{i,t}$, which implies the desired inequality. ■

D. Proof Lemma 1

The KKT conditions, imply that any local optimum satisfies

$$0 = w_i r_{i,f,t} - \lambda_{f,t} + \mu_{i,f,t} \quad \forall_{i\in\mathcal{E}}\ \forall_{f\in\mathcal{F}}\ \forall_{t\in\{1,\ldots,T\}}, \tag{34}$$

$$0 = \lambda_{f,t}\left(\sum_{e\in\mathcal{E}} I^\varepsilon_{e,f,t} - 1\right) \quad \forall_{f\in\mathcal{F}}\ \forall_{t\in\{1,\ldots,T\}}, \tag{35}$$

$$0 = \mu_{i,f,t} I^\varepsilon_{i,f,t} \quad \forall_{i\in\mathcal{E}}\ \forall_{f\in\mathcal{F}}\ \forall_{t\in\{1,\ldots,T\}}, \tag{36}$$

where $\lambda_{f,t}$ and $\mu_{i,f,t}$ are non-negative real numbers that are not all zero, and $w_i = 1/\sum_{\tau=1}^{T}\sum_{k\in\mathcal{F}} r_{i,k,\tau} I^\varepsilon_{i,k,\tau}$. Since $r_{i,f,t} > 0$, $I^\varepsilon_{i,f,t} \geq 0$ and $\mu_{i,f,t} \geq 0$ for all i, f and t, Equation (34) gives that $\lambda_{f,t} \geq \max_{e\in\mathcal{E}} w_e r_{e,f,t} > 0$ for all f, t. Because of Constraint (35), it follows that Constraint (23) is met with equality for all f and t.

Suppose that we take $\lambda_{f,t} > \max_{e\in\mathcal{E}} w_e r_{e,f,t}$ for given f, t. Then Equation (34) gives that $\mu_{i,f,t} > 0$ for all i. This implies that $I_{i,f,t} = 0$ for all i because of Equation (36). But this means Equation (23) is not met with equality. Thus we conclude that we must have $\lambda_{f,t} = \max_{e\in\mathcal{E}} w_e r_{e,f,t}$.

Now Equation (34) gives $\mu_{i,f,t} = \max_{e\in\mathcal{E}}\{w_e r_{e,f,t}\} - w_i r_{i,f,t} \geq 0$ with equality for all $i \in \arg\max_{e\in\mathcal{E}} w_e r_{e,f,t}$ for all f, t. Thus we know that at least $|\mathcal{F}|T$ of the $\mu_{i,f,t} = 0$. Furthermore Equation (36) gives us that $I^\varepsilon_{i,f,t} = 0$ if $i \notin \arg\max_{e\in\mathcal{E}} w_e r_{e,f,t}$.

For an optimal assignment consider the bipartite 'assignment graph' which has two sets of vertices, on the left nodes corresponding to combinations of f and t, and on the right vertices corresponding to the users. Then let there be an edge between the vertices on the left corresponding to (f, t) and the vertices on the right corresponding to user i when $I^\varepsilon_{i,f,t} > 0$.

If the assignment graph is acyclic, it has at most $|\mathcal{F}|T+|\mathcal{E}|-1$ edges ($I^\varepsilon_{i,f,t} > 0$). This implies that $\sum_{i\in\mathcal{E}} \deg(v_i) \leq |\mathcal{F}|T+|\mathcal{E}|-1$, where $\deg(v)$ denotes the degree of the vertex v, and v_i denotes the vertex corresponding to user $i \in \mathcal{E}$. Constraint (23) is met with equality which implies that $\sum_{i\in\mathcal{E}} \deg(v_i) \geq |\mathcal{F}|T$. So we conclude that in the worst case there are at most $|\mathcal{E}|-1$ splits, and thus $2|\mathcal{E}|-2$ variables $I_{i,f,t} \in (0, 1)$. Since $\sum_{i\in\mathcal{E}} \deg(v_i) \leq |\mathcal{F}|T+|\mathcal{E}|-1$ and all $I_{i,f,t} \in [0, 1]$, we now know that at least $|\mathcal{F}|T - |\mathcal{E}| + 1$ variables $I_{i,f,t} = 1$. Since the assignment graph has at most $|\mathcal{F}|T + |\mathcal{E}| - 1$ edges, we know that at least $|\mathcal{E}||\mathcal{F}|T - |\mathcal{F}|T - |\mathcal{E}| + 1$ of the $I_{i,f,t} = 0$.

A careful but straightforward argument (omitted because of page constraints) shows that for each solution with an assignment graph containing a simple cycle, we can construct a solution that is no worse but has strictly fewer edges that need to be removed to obtain an acyclic assignment graph. This implies by induction that there always exists an optimal solution with an acyclic assignment graph, and thus completes the proof. ■