A Matching Based Coexistence Mechanism between eMBB and uRLLC in 5G Wireless Networks

Anupam Kumar Bairagi
Department of Computer Science and Engineering, Kyung Hee University, Yongin 17104, South Korea
anupam@khu.ac.kr

Md. Shirajum Munir
Department of Computer Science and Engineering, Kyung Hee University, Yongin 17104, South Korea
munir@khu.ac.kr

Madyan Alsenwi
Department of Computer Science and Engineering, Kyung Hee University, Yongin 17104, South Korea and
malsenwi@khu.ac.kr

Nguyen H. Tran
School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia
nguyen.tran@sydney.edu.au

Choong Seon Hong
Department of Computer Science and Engineering, Kyung Hee University, Yongin 17104, South Korea and
cshong@khu.ac.kr

ABSTRACT

Ultra-reliable low-latency communication (uRLLC) and enhanced mobile broadband (eMBB) are two major service classes in the emerging 5G mobile network. uRLLC applications demand a stringent latency and reliability whereas eMBB services necessitate utmost data rates. The coexistence of uRLLC and eMBB services on the same radio resource leads to a challenging scheduling problem because of the trade-off among latency, reliability and spectral efficiency. In this paper, a puncturing scheme based coexistence approach between uRLLC and eMBB traffic is proposed for the upcoming 5G mobile networks. Specifically, an optimization problem is formulated with the objective of maximizing the minimum expected achieved rate of eMBB users in the long run basis while meeting the uRLLC requirements. To solve this co-scheduling optimization problem, we decomposed it into two sub-problems with the same objective of the original problem: 1) resource allocation problem for eMBB users, and 2) resource allocation problem for uRLLC users. A heuristic algorithm is used for solving the first sub-problem, whereas the one-sided matching game is used for solving the second. Simulation results show the advantages of the proposed approach over other baseline methods in terms of the minimum achieved rate and fairness among the eMBB users.

CCS CONCEPTS

• **Networks → Mobile networks**;

KEYWORDS

5G; eMBB; uRLLC; Coexistence; Matching Game

SAC '19, April 8–12, 2019, Limassol, Cyprus

ACM ISBN 978-1-4503-5933-7/19/04…$15.00
https://doi.org/10.1145/3297280.3297513

ACM Reference Format:
Anupam Kumar Bairagi, Md. Shirajum Munir, Madyan Alsenwi, Nguyen H. Tran, and Choong Seon Hong. 2019. A Matching Based Coexistence Mechanism between eMBB and uRLLC in 5G Wireless Networks. In *The 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19), April 8–12, 2019, Limassol, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3297280.3297513

1 INTRODUCTION

With the explosive trends of mobile traffic [12], the wireless industries are also experiencing diverse set of emerging applications and services i.e. high-resolution video streaming, virtual reality (VR), augmented reality (AR), autonomous cars, smart cities and factories, smart grids, remote medical diagnosis, unmanned aerial vehicles (UAV), artificial intelligence (AI) based personal assistants, sensing, metering, monitoring etc. The mobile application market is expected to grow in a cumulative average growth rate (CAGR) of 29.1% in the estimated period 2015 − 2020 [14]. The requirements of these applications and services are different in terms of energy efficiency, latency, reliability, data rate etc. For handling these diversified requirements, International Telecommunication Union (ITU) has already listed 5G services into three prime categories: ultra-reliable and low latency communication (*uRLLC*), massive machine-type communication (mMTC), and enhanced mobile broadband (*eMBB*) [15]. uRLLC traffic requires very high reliability (99.999%) with extremely low delay (0.25 ∼ 0.30 ms/packet), whereas mMTC needs high connection density with better energy efficiency, and eMBB expects gigabit per second level data rates [1].

Generally, eMBB users produce the major portion of wireless traffic. However, the uRLLC traffic is sporadic in nature and hence, need to be served instantaneously. To resolve this issue, the simplest way is to reserve some time/frequency resources for uRLLC transmission, and it is preferable from latency/reliability viewpoints. However, this may cause under-utilization of radio resources and hence, requires a dynamic multiplexing of different traffics. Thus, 3GPP has proposed superposition/puncturing framework [1] and short transmission time interval (short-TTI)/puncturing based technique [2] for dynamic multiplexing of eMBB and uRLLC traffics in 5G cellular system. Short-TTI based approach is easy to implement, but reduce spectral efficiency due to high control channel overhead.

On the other hand, puncturing based technique lowers the control channel overhead, but requires efficient mechanism for detecting and recovering the punctured event and data, respectively. To meet the latency requirement of uRLLC traffic, two time units namely slot (1 ms) and mini-slot (0.125 ms) are identified in 5G new radio (NR) proposal. The scheduling for eMBB traffic is made at the beginning of a slot and remains fixed during the slot. Thus, the uRLLC traffics are overlapped onto scheduled eMBB transmission at each mini-slot boundary if they use the same physical resources.

Recently, resource sharing has attracted a lots of attention for providing quality of experience (QoE) to the users in different areas. Authors of [7, 8, 10] propose for sharing of unlicensed spectrum between LTE and WiFi systems whereas the authors of [21] speak about sharing of resources between LTE-A and NB-IoT. In [5], the authors propose a solution approach for user association and resource allocation problem jointly in the downlink of the Fog network acknowledging the evergrowing demand of QoS provisions imposed by the uRLLC and eMBB services. The effective resource sharing between eMBB and uRLLC are discussed in the works [26] and [6]. In the work [26], the authors propose a dynamic puncturing of uRLLC traffic inside eMBB traffic for increasing the resource utilization. They introduce signal space diversity scheme with dynamic puncturing for improving the decoding performance of eMBB data through component interleaving and rotation modulation. The authors formulate a joint scheduling between eMBB and uRLLC for maximizing eMBB users' utility while satisfying aperiodic uRLLC demands in the paper [6]. Specifically, they introduce linear, convex and threshold models for measuring throughput loss of eMBB users in case of superposition/puncturing in [6]. In the paper [17], the authors investigate the performance of orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) for the multiplexing of eMBB and uRLLC users in the uplink using a cloud radio access network (C-RAN) architecture. They study the performance trade-offs between eMBB and uRLLC traffic by applying information-theoretic arguments. The authors explore the impact of uRLLC traffic on eMBB transmission for mobile front-haul in [27]. They also introduce some solutions for fulfilling the obligations of uRLLC services while minimizing the influence on eMBB services. In the paper [22], the authors try to serve the eMBB users with high throughput, while assisting uRLLC users within the stringent constraints. They propose restoration mechanisms for the impacted eMBB users and recommend efficient multiplexing of both traffics in case of punctured scheduling. The authors of [23] investigate the performance of orthogonal and non-orthogonal slicing of radio resources for provisioning eMBB, mMTC and uRLLC services of 5G. The heterogeneous requirements and characteristics of these services are incorporated into a communication-theoretic model in this paper. They unveil that the non-orthogonal slicing has significant performance gain over the orthogonal slicing in case of these service multiplexing. The work [20] focuses on the system level design for uRLLC traffic which is supported by $M/D/m/m$ queueing model. They reveal that static partitioning of bandwidth between eMBB and uRLLC is inefficient and hence, requires a dynamic multiplexing of the traffics in both time and frequency domains. In the paper [16], the authors present an up-to-date overview of physical layer challenges concerning uRLLC communications and solution approach in 5G NR. They emphasize on packet and frame structure, scheduling schemes, and reliability improvement techniques in [16]. They also touch the coexistence issue of uRLLC with eMBB in this work. The authors provide a crunchy insight for designing a low-latency and high-reliability 5G wireless network in [11]. Specifically, they explore various enablers of uRLLC, and also their inherent trade-offs in the cases of delay, reliability, packet size, network architecture, topology, and uncertainty.

Hence, most of the works in the field focus on system level design and analysis, and very few of them have concrete coexistence model between eMBB and uRLLC users. Therefore, we propose a coexistence mechanism between eMBB and uRLLC users for enabling 5G wireless networks in this paper. More specifically, the main contributions of this paper are as follows:

- We formulate an optimization problem to maximize the minimum expected rate of eMBB users over time.
- We decompose of the problem into two sub-problems: Resource allocation problem of eMBB users, and resource allocation problem of uRLLC users. The first sub-problem is solved using a heuristic algorithm and the second one is solved by utilizing matching game.
- We justify the quality of the proposed approach with extensive simulations.

The rest of the paper is organized as follows. In Section 2, we discuss the system model and problem formulation. The solution to this problem is discussed in Section 3. In Section 4, we present simulation results. Finally the paper is concluded in Section 5.

2 SYSTEM MODEL AND PROBLEM FORMULATION

Our deployment scenario consists of one next generation base station (gNB), a set of active eMBB users $\mathcal{E}$ and a set of uRLLC users $\mathcal{U}$ working in downlink mode, and shown in Fig. 1. gNB has a set of licensed resource blocks (RBs) $\mathcal{K}$ of uniform bandwidth B for supporting eMBB and uRLLC users. Each time slot, Δ, is divided into M mini-slots of length δ for providing low latency services. We consider T_s tradition LTE time slots for eMBB users and represented by $\mathcal{T} = \{1, 2, \cdots, T_s\}$. We assume that uRLLC traffics arrive at gNB in any mini-slot m of time slot t for uRLLC users follows Gaussian distribution, i.e., $U \sim \mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are the mean and variance of the distribution, referred to as uRLLC arrival rate, and payload size of each uRLLC user $u \in \mathcal{U}$ is $L_u^{m,t}$ (varying from 32 to 200 Bytes [3]).

At the beginning of each time slot t, gNB allocates the physical resources to the eMBB users. If gNB allocates a RB $k \in \mathcal{K}$ to eMBB user $e \in \mathcal{E}$ then the achievable rate of that user is as follows:

$$r_{e,k}^t = \Delta B \log_2(1 + \gamma_{e,k}^t) \tag{1}$$

where $\gamma_{e,k}^t$ is the signal-to-noise ratio (SNR) and $\gamma_{e,k}^t = \frac{P_e h_e^2}{N_0 B}$. P_e and h_e are transmission power for and gain of eMBB user $e \in \mathcal{E}$ from the gNB, respectively, and N_0 is the noise spectral density. Generally, eMBB users need more RBs for satisfying its' high-throughput requirements. Hence, the user $e \in \mathcal{E}$ can obtain

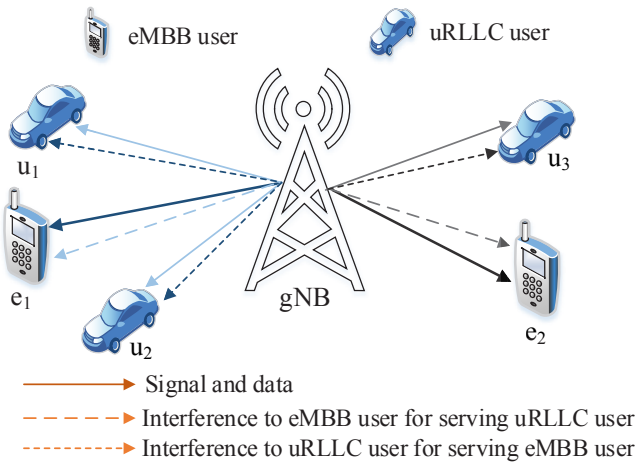

Figure 1: Illustration of the system model

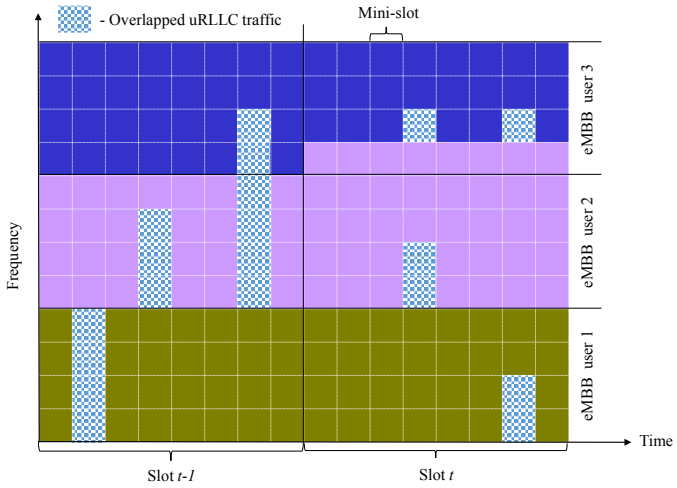

Figure 2: Illustration of multiplexing of eMBB and uRLLC traffic

the following rate at time slot t:

$$r_e^t = \sum_{k \in \mathcal{K}} \alpha_{e,k}^t r_{e,k}^t \tag{2}$$

where $\boldsymbol{\alpha}$ is the resource allocation vector of the gNB for it's eMBB users at time slot t, and $\alpha_{e,k}^t = 1$ if gNB allocates RB k for eMBB user $e \in \mathcal{E}$ at time slot t, $\alpha_{e,k}^t = 0$ otherwise.

uRLLC traffic can reach at any time (i.e. mini-slot) within time slot t and needs to be served immediately. Due to the latency requirements and diversity of uRLLC traffic, each one should be served within a single mini-slot period. Typically, the uRLLC traffic-length is quite small and hence, Shannon's asymptotic analysis cannot be directly applied [17]. If uRLLC traffic overlapped with eMBB traffic, then the achievable rate in RB $k \in \mathcal{K}$ for uRLLC user can be well approximated by using [25] and shown as follows:

$$r_{u,k}^{m,t} = \delta[B \log_2(1 + \gamma_u^{m,t}) - \sqrt{\frac{V_u}{N_u^b}} Q^{-1}(\varepsilon_u^d)] \tag{3}$$

where $\gamma_u^{m,t}$ is the signal to interference plus noise ratio (SINR) for uRLLC user $u \in \mathcal{U}$ at mini-slot m of slot t, and $\gamma_u^{m,t} = \frac{h_u^2 P_u}{N_0 B + h_u^2 P_e}$ with $h_u^2 P_e$ represents interference from serving uMBB user $e \in \mathcal{E}$. Here, V represents the channel dispersion and given by $V_u = \frac{h_u^2 P_u}{N_0 B + h_u^2 (P_u + P_e)}$, N_u^b is the blocklength of uRLLC traffic, and Q is the complementary Gaussian cumulative distribution function with probability of decoding error ε_u^d. However, this interference of eMBB user makes the reliability issue of uRLLC traffic vulnerable, and hence superposition is not a preferred candidate [27]. Therefore, our priority is on puncturing technique for serving uRLLC users with high reliability. gNB allocates zero power for eMBB transmission in the punctured mini-slot and hence, uRLLC traffic is not affected by any eMBB interference. Thus, $\gamma_u^{m,t} = \frac{h_u^2 P_u}{N_0 B}$ and $V = \frac{h_u^2 P_u}{N_0 B + h_u^2 P_u}$. If an uRLLC user needs multiple RBs to support its request, then the achievable rate of that user is as follows:

$$r_u^{m,t} = \sum_{k \in \mathcal{K}} \beta_{e,k}^{m,t} r_{u,k}^{m,t} \tag{4}$$

where $\boldsymbol{\beta}$ is the resource allocation vector of the gNB for it's uRLLC users at mini-slot m of time slot t, and $\beta_{e,k}^{m,t} = 1$ if gNB allocates RB k for uRLLC user $u \in \mathcal{U}$, $\beta_{e,k}^{m,t} = 0$ otherwise.

The serving probability of all the uRLLC users in any mini-slot m is almost sure, and hence,

$$P(\sum_{u \in \mathcal{U}} \phi_u^{m,t} < U) \leq \epsilon, \forall m, t \tag{5}$$

Here, the vector $\boldsymbol{\phi}$ represents the current serving uRLLC users, and $\phi_u^{m,t} = 1$ if gNB serves for uRLLC user $u \in \mathcal{U}$ in mini-slot m of time slot t, $\phi_u^{m,t} = 0$ otherwise. The payload $L_u^{m,t}$ of uRLLC user $u \in \mathcal{U}$ needs to be transmitted within the stipulated time period δ if it is served and hence the following condition is satisfied.

$$\phi_u^{m,t} L_u^{m,t} \leq \delta r_u^{m,t}, \forall u, m, t \tag{6}$$

Therefore, (5) and (6) jointly protect the reliability and latency issues of uRLLC users. Moreover, if gNB punctures uRLLC traffic within eMBB user $e \in \mathcal{E}$ at time slot t, then it surely loses some throughput at this slot. For calculating the throughput-losses occurred to eMBB users, we use linear model as proposed in [6]. Thus, the amount of throughput-losses for eMBB user $e \in \mathcal{E}$ is as follows:

$$r_{e,loss}^t = \sum_{k \in \mathcal{K}} r_{e,k}^t \sum_{m \in \mathcal{M}} \sum_{u \in \mathcal{U}} \mathbb{I}(\alpha_{e,k}^t = \beta_{u,k}^{m,t}). \tag{7}$$

Hence, the actual rate achieved by eMBB user $e \in \mathcal{E}$ at time slot t is as follows:

$$r_{e,actual}^t = r_e^t - r_{e,loss}^t. \tag{8}$$

Now our goal is to maximize the achieved rate of every eMBB user over the time period while satisfying almost every uRLLC request within their stringent latency constraint. For this purpose, we use *Max-Min* fairness policy which provides stable service quality, increases spectral efficiency and makes eMBB users happier in the

wireless network. Hence, we formulate an optimization problem as follows:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\min_{e\in\mathcal{E}}(\mathbb{E}(\sum_{t=1}^{|\mathcal{T}|} r_{e,actual}^{t}))) \tag{9}$$

$$\text{s.t.} \quad P(\sum_{u\in\mathcal{U}} \phi_u^{m,t} < U) \leq \epsilon, \forall m, t \tag{9a}$$

$$\phi_u^{m,t} L_u^{m,t} \leq \delta r_u^{m,t}, \forall u, m, t \tag{9b}$$

$$\sum_{e\in\mathcal{E}} \alpha_{e,k}^{t} \leq 1, \forall k, t \tag{9c}$$

$$\sum_{u\in\mathcal{U}} \beta_{u,k}^{m,t} \leq 1, \forall k, m, t \tag{9d}$$

$$\sum_{k\in\mathcal{K}} \alpha_{e,k}^{t} \geq 1, \forall e, t \tag{9e}$$

$$\sum_{k\in\mathcal{K}} \phi_u^{m,t} \beta_{u,k}^{m,t} \geq 1, \forall u, m, t \tag{9f}$$

$$\sum_{e\in\mathcal{E}} \sum_{k\in\mathcal{K}} \alpha_{e,k}^{t} \leq |\mathcal{K}|, \forall t \tag{9g}$$

$$\sum_{u\in\mathcal{U}} \sum_{k\in\mathcal{K}} \phi_u^{m,t} \beta_{u,k}^{m,t} \leq |\mathcal{K}|, \forall m, t \tag{9h}$$

$$\alpha_{e,k}^{t}, \beta_{u,k}^{m,t}, \phi_u^{m,t} \in \{0,1\}, \forall e, u, k, m, t. \tag{9i}$$

Here, the constraints (9a) and (9b) preserve the reliability and latency requirements of the uRLLC users respectively. The orthogonality of RBs for uMBB and uRLLC users are endured by the constraints (9c) and (9d) respectively. Each served eMBB and uRLLC user poss at least one RBs and that are represented by the constraints (9e) and (9f) respectively. The constraints (9g) and (9h) present the limitation of gNB RBs. Every element of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ can be either 0 or 1 and shown in the constraint (9i) accordingly. The optimization in (9) is a Combinatorial Programming (CP) problem with chance constraint, which is NP-hard for its nature and cannot be solved in real time.

3 SOLUTION APPROACH BY DECOMPOSITION OF PROBLEM (9)

We assume that gNB has enough traffic for eMBB users, and hence, uMBB users are scheduled on all available RBs at the beginning of a time slot and fixed over the slot. When uRLLC traffic arrives at the gNB (in any mini-slot of current slot), the scheduler aims at immediately scheduling such traffic with the mini-slot. Thus, the uRLLC traffic is scheduled on radio resources currently allocated to eMBB users and an overlapping of uRLLC traffic happens as shown in Figure 2. Due to the hard latency bindings of uRLLC traffic, we choose puncturing mechanism to be applied for the overlapped uRLLC traffic in this paper. Generally, uRLLC traffic has small payload size and thus, requires a fraction of all RBs for such traffic. However, the question is of selecting the RBs, which is currently occupied by eMBB users, are the best to be punctured keeping the objective of problem (9) in mind. Moreover, due to the unpredictable nature and latency requirements of uRLLC traffic, we cannot decide on the fly. Now we want to decompose the problem in (9) into two sub-problems so that each of them can be solved with suitable technique. The resource allocation problem of the eMBB users are shown as follows:

$$\max_{\boldsymbol{\alpha}}(\min_{e\in\mathcal{E}}(\mathbb{E}(\sum_{t=1}^{|\mathcal{T}|} r_{e,actual}^{t}))) \tag{10}$$

$$\text{s.t.} \quad \sum_{e\in\mathcal{E}} \alpha_{e,k}^{t} \leq 1, \forall k, t \tag{10a}$$

$$\sum_{k\in\mathcal{K}} \alpha_{e,k}^{t} \geq 1, \forall e, t \tag{10b}$$

$$\sum_{e\in\mathcal{E}} \sum_{k\in\mathcal{K}} \alpha_{e,k}^{t} \leq |\mathcal{K}|, \forall t \tag{10c}$$

$$\alpha_{e,k}^{t} \in \{0,1\}, \forall e, k, t. \tag{10d}$$

Secondly, the resource allocation problem of uRLLC users considering the same objective of (10) with fixed $\alpha^t, \forall t$ are shown as follows:

$$\max_{\boldsymbol{\beta}}(\min_{e\in\mathcal{E}}(\mathbb{E}(\sum_{t=1}^{|\mathcal{T}|} r_{e,actual}^{t}))) \tag{11}$$

$$\text{s.t.} \quad P(\sum_{u\in\mathcal{U}} \phi_u^{m,t} < U) \leq \epsilon, \forall m, t \tag{11a}$$

$$\phi_u^{m,t} L_u^{m,t} \leq \delta r_u^{m,t}, \forall u, m, t \tag{11b}$$

$$\sum_{u\in\mathcal{U}} \beta_{u,k}^{m,t} \leq 1, \forall k, m, t \tag{11c}$$

$$\sum_{k\in\mathcal{K}} \phi_u^{m,t} \beta_{u,k}^{m,t} \geq 1, \forall u, m, t \tag{11d}$$

$$\sum_{u\in\mathcal{U}} \sum_{k\in\mathcal{K}} \phi_u^{m,t} \beta_{u,k}^{m,t} \leq |\mathcal{K}|, \forall m, t \tag{11e}$$

$$\beta_{u,k}^{m,t}, \phi_u^{m,t} \in \{0,1\}, \forall u, k, m, t. \tag{11f}$$

3.1 Heuristic Algorithm based Solution for Sub-Problem (10)

Since the sub-problem (10) is still NP-hard, and it is computationally expensive to find global optimal solution. Moreover, we have many eMBB users in reality, time dependency of (10), and a small amount of time to solve such resource allocation problem. Hence, we redefine the sub-problem (10) for every time slot t in (12) so that it can be solved with a heuristic algorithm in a faster and efficient fashion by sacrificing optimality.

$$\min_{\boldsymbol{\alpha^t}} \sum_{e\in\mathcal{E}} W_e^t(\boldsymbol{\alpha^t}), \forall t, \tag{12}$$

$$\text{s.t.} \quad (10a), (10b), (10c), (10d) \tag{12a}$$

$$W_e^t(\boldsymbol{\alpha^t}) = \Big| \frac{1}{t|\mathcal{E}|} \sum_{e'\in\mathcal{E}} \Big(\sum_{t'=1}^{t-1} r_{e',actual}^{t'} + r_e^t \Big) - \frac{1}{t} \Big(\sum_{t'=1}^{t-1} r_{e,actual}^{t'} + r_e^t \Big) \Big| \tag{12b}$$

Now, for solving (12), we propose an heuristic algorithm as shown in Algorithm 1 for each time slot $t \in \mathcal{T}$. In the first time slot, the Algorithm 1 allocates RBs to the eMBB users equally. However, the Algorithm 1 considers the impact of RBs allocation of uRLLC users till $t-1$ time slots when it allocates for t^{th} time slot for eMBB

Algorithm 1: Heuristic Algorithm at time slot t for Solving (12)

Input: $\mathcal{E}, \mathcal{K}, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\beta}^{t-1}$
Result: $\boldsymbol{\alpha}^t$

if $t == 1$ **then**
  Compute $N_{RB} = \frac{|\mathcal{K}|}{|\mathcal{E}|}$ **for** *each* $e \in \mathcal{E}$ **do**
    for *each* $k = 1 \cdots N_{RB}$ **do**
      Update $\alpha^t_{e,(e-1)*N_{RB}+k} = 1$
    end
  end
end
else
  Compute $r^{t-1}_{e,actual}$ for all $e \in \mathcal{E}$ by using (8) Update $loc = 0$ Compute $Avg_{\mathcal{E}} = \frac{1}{|\mathcal{E}|\cdot(t-1)} \sum_{e\in\mathcal{E}} \sum_{t'=1}^{t-1} r^{t'}_{e,actual}$ Compute $Avg_e = \frac{1}{(t-1)} \sum_{t'=1}^{t-1} r^{t'}_{e,actual}, \forall e \in \mathcal{E}$ **for** *each* $e \in \mathcal{E}$ **do**
    Calculate $N^e_{RB} = \frac{2*Avg_{\mathcal{E}} - Avg_e}{\sum_{e'\in\mathcal{E}}(2*Avg_{\mathcal{E}} - Avg_e)} \cdot |\mathcal{K}|$ **for** *each* $k = 1 \cdots N^e_{RB}$ **do**
      Update $\alpha^t_{e,loc+k} = 1$
    end
    Set $loc = loc + N^t_{RB}$
  end
end

users. That means, the Algorithm 1 envisages the expected actual achieved rates of the eMBB users till $t-1$ time slots.

3.2 One-Sided Matching based Solution for Sub-Problem (11)

We can see from (11) that the sub-problem is difficult to solve using the typical optimization solvers due to the appearance of chance constraint and also combinatorial nature of $\boldsymbol{\beta}$. To solve the sub-problem (11), we need to convert the chance constraint (11a) into deterministic one. For that, let $g(\boldsymbol{\phi}, U) = \sum_{u\in\mathcal{U}} \phi^{m,t}_u - U$, $U \in \mathbb{R}$ and $U \sim \mathcal{N}(\mu, \sigma^2), \forall m, t$. Hence,

$$Pr\{g(\phi, U) \le 0\} = Pr\Big\{ \sum_{u\in\mathcal{U}} \phi^{m,t}_u - U \le 0 \Big\} \tag{13}$$

$$= Pr\Big\{ \sum_{u\in\mathcal{U}} \phi^{m,t}_u \le U \Big\} \tag{13a}$$

$$= 1 - Pr\Big\{ \sum_{u\in\mathcal{U}} \phi^{m,t}_u \ge U \Big\} \tag{13b}$$

$$= 1 - Pr\Big\{ \frac{U-\mu}{\sigma} \le \frac{\sum_{u\in\mathcal{U}} \phi^{m,t}_u - \mu}{\sigma} \Big\} \tag{13c}$$

$$= 1 - F_U\Big(\sum_{u\in\mathcal{U}} \phi^{m,t}_u \Big) \tag{13d}$$

Here, F_U is the cumulative distribution function (CDF) of random variable U. Thus, from constraint (11a), we can rewrite as follows:

$$Pr\{g(\phi, U) \le 0\} \ge \epsilon \tag{14}$$

$$1 - F_U\Big(\sum_{u\in\mathcal{U}} \phi^{m,t}_u \Big) \le \epsilon \tag{14a}$$

$$F_U\Big(\sum_{u\in\mathcal{U}} \phi^{m,t}_u \Big) \ge 1 - \epsilon \tag{14b}$$

$$\sum_{u\in\mathcal{U}} \phi^{m,t}_u \ge F_U^{-1}\Big(1 - \epsilon\Big) \tag{14c}$$

$$\sum_{u\in\mathcal{U}} \phi^{m,t}_u - F_U^{-1}(1-\epsilon) \ge 0 \tag{14d}$$

Here, (14d) is the deterministic equivalent of (11a). Now, rewrite the sub-problem (11) for each of the mini-slot m of time slot $t \in \mathcal{T}$ as follows:

$$\min_{\boldsymbol{\beta}^t} \quad \sum_{e\in\mathcal{E}} V^t_e(\boldsymbol{\alpha}^t, \boldsymbol{\beta}^t), \forall t \tag{15}$$

$$\text{s.t.} \quad \sum_{u\in\mathcal{U}} \phi^{m,t}_u - F_U^{-1}(1-\epsilon) \ge 0, \forall m \tag{15a}$$

$$(11b), (11c), (11d), (11e), (11f), \forall u, m \tag{15b}$$

$$V^t_e(\boldsymbol{\alpha}^t, \boldsymbol{\beta}^t) = \Big| \frac{1}{|\mathcal{E}|} \sum_{e'\in\mathcal{E}} r^t_{e',loss} - r^t_{e,loss} \Big|, \forall e \tag{15c}$$

The problem shown in (15) is still difficult to solve in real time. However, for the fixed value of ϵ, the reliability constraint (15a) holds by serving a particular number of uRLLC users $U' \le U$. Let the set of these uRLLC users in a particular mini-slot m is $\mathcal{U}' = \{1, 2, ..., U'\}$ and $\phi^{m,t}_u = 1, \forall u \in \mathcal{U}'$. Fixing δ to the upper-bound in (11b), we can find the required RBs for each uRLLC user $u \in \mathcal{U}'$ in mini-slot m. Now the problem (15) turns into a resource assignment problem in each mini-slot m, which can be solved efficiently by using the matching framework, which is successfully used in resource allocation problems [9, 18, 19], namely the *house allocation problem* (HAP) [28], [4] considering the objective of optimization (11). HAP provides a suboptimal solution with less complexity. The house allocation problem is a one-sided matching represented by a tuple $(\mathcal{A}, \mathcal{H}, \mathcal{P})$, where $\mathcal{A}$ is the set of agents, $\mathcal{H}$ is comprised of a set of houses, and $\mathcal{P}$ represents the preferences of agents over the houses. In our context, $\mathcal{U}'$ are the agents, $\mathcal{E}$ correspond to the houses, and $\mathcal{U}'$ have preferences over $\mathcal{E}$. The gNB is assumed to have full knowledge on their eMBB and uRLLC users. In this matching model, if gNB allocates RB(s) to uRLLC user $u \in \mathcal{U}'$ that is already hold by eMBB user $e \in \mathcal{E}$, then uRLLC user u is said to be matched with eMBB user e and form a matching pair (u, e). Thus, a matching is an assignment of uRLLC users in $\mathcal{U}'$ to eMBB users $\mathcal{E}$, which can be defined as follows:

Definition 3.1. A matching Ω between $\mathcal{U}'$ and $\mathcal{E}$ is a mapping from the set $\mathcal{U}' \cup \mathcal{E}$ to the set of all subsets of $\mathcal{U}' \cup \mathcal{E}$ such that for every $u \in \mathcal{U}'$ and every $e \in \mathcal{E}$: (i) $\Omega(u) \in \mathcal{E}$ and $\Omega(e) \in \mathcal{U}'$, (ii) $|\Omega(u)| = 1$, (iii) $|\Omega(e)| \ge 0$, (iv) $u \in \Omega(e)$ if and only if $e \in \Omega(u)$, and (v) u gets at least q_u RBs from the matching Ω.

The value of q_u is not predefined for each $u \in \mathcal{U}'$, rather gNB determines it dynamically based on latency requirement and payload

Algorithm 2: One-Sided Matching-Based Resource Allocation for uRLLC users in mini-slot m of time slot t

Input: $\mathcal{E}, \mathcal{U}^{'}, \boldsymbol{\alpha}^{t}$
Result: Ω

for *each* $u \in \mathcal{U}^{'}$ **do**
  gNB makes preference list $\mathcal{P}_u^{m,t}$ over $\mathcal{E}$ based on (16)
end
Update $\mathcal{E}_{sort}$ by sorting eMBB users of $\mathcal{E}$ depending on the expected achieved rate in descending order ;
for *each* $e \in \mathcal{E}_{sort}$ **do**
  Update $q_e = |\mathcal{U}^{'}|$ div $|\mathcal{E}|$
end
Update $q_{rem} = |\mathcal{U}^{'}|$ mod $|\mathcal{E}|$;
if $q_{rem} > 0$ **then**
  for $e = 1, 2, \cdots, q_{rem}$ **do**
    Update $q_e = q_e + 1$
  end
end
while *there is proposal from uRLLC users and quota available for eMBB users* **do**
  gNB, on behalf of each $u \in \mathcal{U}^{'}$, proposes to the first eMBB user of $\mathcal{P}_u^{m,t}$;
  for *each* $e \in \mathcal{E}$ **do**
    if *The number of proposals* $\leq q_e$ **then**
      eMBB user $e \in \mathcal{E}$ holds all the proposals temporarily and gNB prohibits further proposal from that uRLLC users
    end
    else
      eMBB user $e \in \mathcal{E}$ holds q_e proposals and gNB prohibits further proposal from that uRLLC users, rejects rest of the proposals and remove e from $\mathcal{P}_u^{m,t}$
    end
  end
end

size. Definition 3.1 states that an uRLLC user u can only be matched with one eMBB user from $\mathcal{E}$ while one eMBB user e can be matched with multiple uRLLC users of $\mathcal{U}^{'}$. For allocating RBs to the uRLLC users $\mathcal{U}^{'}$, each $u \in \mathcal{U}^{'}$ requires specifying its preferences over the eMBB users in mini-slot m of time slot $t \in \mathcal{T}$ as follows:

$$\mathcal{P}_u^{m,t} = \mathbb{E}\left(\sum_{t^{'}=1}^{t-1} r_{e,actual}^{t^{'}} + \sum_{m^{'}=1}^{m-1} r_{e,actual}^{m^{'}} \right), \forall e \in \mathcal{E} \tag{16}$$

Each uRLLC user $u \in \mathcal{U}^{'}$ can define its preference relation $\mathcal{P}_u^{m,t}$ over the set of users $\mathcal{E}$ in any mini-slot m of time slot $t \in \mathcal{T}$ such that for any two eMBB users $e, e' \in \mathcal{E}, e \neq e'$ and two matching

Table 1: Value of the principal simulation parameters

Symbol	Value	Symbol	Value
$\lvert\mathcal{E}\rvert$	10	$\lvert\mathcal{K}\rvert$	50
$\lvert\mathcal{T}\rvert$	1000	M	8
Δ	1ms	δ	0.125 ms
$P_e, \forall e$	21 dBm	N_0	−114 dBm
σ	$1, 2, \cdots, 10$		

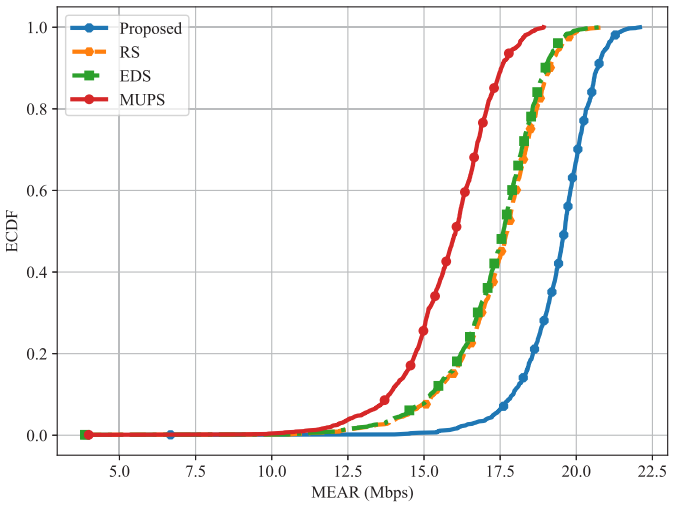

Figure 3: Comparison of MEAR of eMBB users with $\sigma = 5$ along with $L = 32$ bytes

$\Omega, \Omega^{'}, e \in \Omega(u), e' \in \Omega^{'}(u)$:

$$\begin{aligned}(e, \Omega) \succ_u (e', \Omega^{'}) \Leftrightarrow \mathbb{E}\left(\sum_{t^{'}=1}^{t-1} r_{e,actual}^{t^{'}} + \sum_{m^{'}=1}^{m-1} r_{e,actual}^{m^{'}} \right) \\ \succ_u \mathbb{E}\left(\sum_{t^{'}=1}^{t-1} r_{e',actual}^{t^{'}} + \sum_{m^{'}=1}^{m-1} r_{e',actual}^{m^{'}} \right).\end{aligned} \tag{17}$$

So, each uRLLC user $u \in \mathcal{U}^{'}$ has its preference list $\mathcal{P}_u^{m,t}$ depending on the expected actual achievable rate that it gets from (16) till the previous mini-slot by sorting in descending order. In the original house allocation problem, one agent is allocated to only one house but in our problem, multiple uRLLC users can be assigned to one eMBB user. Based on the preference $\mathcal{P}_u^{m,t}, \forall u \in \mathcal{U}^{'}$, a one-sided matching-based resource allocation process for the uRLLC users $\mathcal{U}^{'}$ is shown in Algorithm 2. The output Ω of Algorithm 2 can be transformed to a feasible allocation vector $\boldsymbol{\beta}^{t}$ of problem (15) for the gNB.

4 SIMULATION RESULTS

We verify the performance of the system based on minimum expected achieved rate (MEAR) and fairness among eMBB users. To

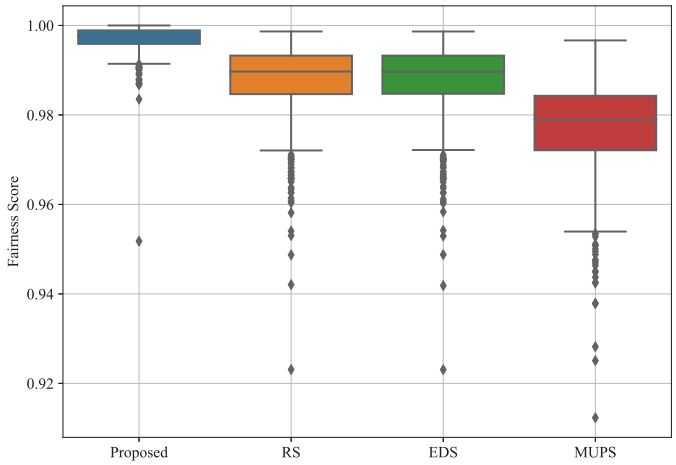

Figure 4: Comparison of fairness scores for $\sigma = 5$ and $L = 32$ bytes

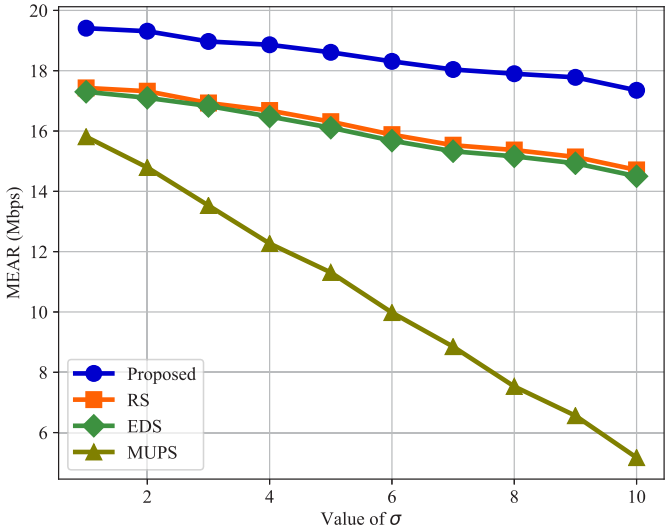

Figure 5: Comparison of average MEAR with varying value of σ and $L = 32$ bytes

measure the system fairness, we use Jain's fairness index [24]. gNB is located in the center of an area of radius 200 m and eMBB users are distributed randomly in the coverage area of the gNB. gNB operates on 10 MHz licensed band and we assume that each uRLLC user requires single RB for it's operation. Moreover, gNB uses free space propagation path-loss model with Rayleigh fading for eMBB users. The major simulation parameters for the gNB are shown in Table 1. We compare the performance of the proposed coexistence scheme with random scheduler (RS) (uRLLC users are served by picking RBs randomly from eMBB users), equally distributed scheduler (EDS) (uRLLC users are served by equally picking RBs from the eMBB users) and multi-user preemptive scheduler (MUPS)[13] mechanism after taking 1000 runs for all the methods.

Figure 3 shows empirical cumulative distribution function (ECDF) of the minimum achieved rate of eMBB users for 1000 runs among different methods with $\sigma = 5$. Figure 3 shows that the ECDF of the minimum achieved rate of eMBB users resulting from the proposed method is superior to those of the other baseline methods. Moreover,that the proposed, RS, EDS, and MUPS give MEAR value of at least 17.50 Mbps with probability 0.937, 0.565, 0.535, and 0.11, respectively.

In Figure 4, we present a comparison of fairness scores of eMBB users for different methods. Figure 4 shows that the fairness scores resulting from the proposed method for eMBB users is better than that of all other methods. Figure 4 shows that the medians of these fairness scores are 0.9977, 0.9897, 0.9897, and 0.9789 for the Proposed, RS, EDS, and MUPS methods, respectively. Moreover, these scores of the Proposed method is 0.80%, 0.80%, and 1.88% fairer than the RS, EDS, and MUPS methods, respectively.

In Figure 5, we present a comparison of the MEAR of eMBB users on average for different methods with increasing number of uRLLC request (varying by σ value). Figure 5 shows that the MEAR of eMBB users resulting from the proposed method outperforms that of all other methods for all cases. We also reveal from the Figure 5 that the MEAR decreases with an increasing value of σ due to the demand of additional RBs of extra uRLLC users. Specifically, the average MEAR value of the proposed method is 10.20%, 10.87%, and 18.55% higher than those of RS, EDS, and MUPS, respectively, for $\sigma = 1$, whereas these values are 15.22%, 16.43%, and 70.20% superior than those of the same corresponding methods, respectively, for $\sigma = 10$.

5 CONCLUSIONS

In this paper, we have proposed a novel method that allows uRLLC and eMBB users to coexist in the same radio resource for 5G wireless networks. We formulated the problem as maximizing the minimum expected achieved rate of the eMBB users while meeting the uRLLC requirements and solved it using a decomposition approach. We have solved the resource allocation sub-problem of eMBB users with a heuristic algorithm for each time slot, whereas the resource allocation problem of uRLLC users is solved via a one-sided matching game. Simulation results show that the proposed approach provides a better minimum expected achieved rate and fairness for eMBB users compared with the other methods.

ACKNOWLEDGMENTS

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2015-0-00567, Development of Access Technology Agnostic Next-Generation Networking Technology for Wired-Wireless Converged Networks) and supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2016R1D1A1B01015320). *Dr. CS Hong is the corresponding author.

REFERENCES

[1] 3GPP. 2016. 3GPP TSG RAN WG1 Meeting 87. (November 2016).
[2] 3GPP. 2017. Downlink Multiplexing of eMBB and URLLC Transmission. *3GPP TSG RAN WG1 NR Ad-Hoc Meeting, R1-1700374* (January 2017).
[3] 3GPP. 2017. Study on New Radio Access Technology Physical Layer Aspects. *Document 3GPP RT 38.802v14.0.0* (March 2017).

[4] A. Abdulkadiroğlu and T. Sönmez. 1998. Random serial dictatorship and the core from random endowments in house allocation problems. *Econometrica* 66, 3 (May 1998), 689–701.

[5] Sarder Fakhrul Abedin, Md Golam Rabiul Alam, SM Ahsan Kazmi, Nguyen H Tran, Dusit Niyato, and Choong Seon Hong. 2018. Resource Allocation for Ultra-reliable and Enhanced Mobile Broadband IoT Applications in Fog Network. *IEEE Transactions on Communications* (2018).

[6] A. Anand, G. Veciana, and S. Shakkottai. 2018. Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks. In *Proceedings of IEEE International Conference on Computer Communications*.

[7] Anupam Kumar Bairagi, Sarder Fakhrul Abedin, Nguyen H Tran, Dusit Niyato, and Choong Seon Hong. 2018. QoE-Enabled Unlicensed Spectrum Sharing in 5G: A Game-Theoretic Approach. *IEEE Access* 6 (2018), 50538–50554.

[8] A. K. Bairagi, N. H. Tran, and C. S. Hong. 2018. A Multi-Game Approach for Effective Co-existence in Unlicensed Spectrum between LTE-U System and Wi-Fi Access Point. In *Proceedings of 2018 International Conference on Information Networking (ICOIN)*. Thailand, 380–385.

[9] Anupam Kumar Bairagi, Nguyen H Tran, Walid Saad, Zhu Han, and Choong Seon Hong. 2018. A Game-Theoretic Approach for Fair Coexistence between LTE-U and Wi-Fi Systems. *IEEE Transactions on Vehicular Technology* (2018).

[10] A. K. Bairagi, N. H. Tran, W. Saad, and C. S. Hong. 2018. Bargaining Game for Effective Coexistence between LTE-U and Wi-Fi Systems. In *Proceedings of NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*. Taiwan, 1–6.

[11] M. Bennis, M. Debbah, and H. V. Poor. [n. d.]. Ultra-Reliable and Low-Latency Wireless Communication: Tail, Risk and Scale. In *https://arxiv.org/abs/1801.01270*.

[12] Cisco. 2017. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016 - 2021. *White Paper* (June 2017).

[13] Ali Abdul-Mawgood Ali Ali Esswie and Klaus Pedersen. 2018. Multi-User Pre-emptive Scheduling For Critical Low Latency Communications in 5G Networks. In *IEEE International Symposium on Computers and Communications*.

[14] 5G Forum. 2016. 5G Service Roadmap 2022. *White Paper* (March 2016).

[15] ITU-R. 2015. IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond. *Recommendation M.2083-0* (September 2015).

[16] Hyoungju Ji, Sunho Park, Jeongho Yeo, Younsun Kim, Juho Lee, and Byonghyo Shim. 2018. Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects. *IEEE Wireless Communications* 25, 3 (2018), 124–130.

[17] R. Kassab, O. Simeone, and P. Popovski. [n. d.]. Coexistence of URLLC and eMBB services in the C-RAN Uplink: An Information-Theoretic Study. In *https://arxiv.org/abs/1804.06593*.

[18] SM Ahsan Kazmi, Nguyen H Tran, Walid Saad, Zhu Han, Tai Manh Ho, Thant Zin Oo, and Choong Seon Hong. 2017. Mode selection and resource allocation in device-to-device communications: A matching game approach. *IEEE Transactions on Mobile Computing* 11 (2017), 3126–3141.

[19] Tuan LeAnh, Nguyen H Tran, Walid Saad, Long Bao Le, Dusit Niyato, Tai Manh Ho, and Choong Seon Hong. 2017. Matching Theory for Distributed User Association and Resource Allocation in Cognitive Femtocell Networks. *IEEE Trans. Veh. Technol* 66 (2017), 8413–8428.

[20] C. Li, J. Jiang, W. Chen, T. Ji, and J. Smee. 2017. 5G ultra-reliable and low-latency systems design. In *Proceedings of 2017 European Conference on Networks and Communications (EuCNC)*. Oulu, 1–5.

[21] S. Liu, F. Yang, J. Song, and Z. Han. 2017. Block Sparse Bayesian Learning-Based NB-IoT Interference Elimination in LTE-Advanced Systems. *IEEE Transactions on Communications* 65, 10 (October 2017), 4559–4571.

[22] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad. 2017. Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband. In *Proceedings of 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. Toronto, ON, 1–6.

[23] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi. [n. d.]. 5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View. In *https://arxiv.org/abs/1804.05057*.

[24] D.M. Chiu R. Jain and W.R. Hawe. 1984. A quantitative measure of fairness and discrimination for resource allocation in shared computer system. *Eastern Research Laboratory, Digital Equipment Corporation* 38 (September 1984).

[25] J. Scarlett, V. Y. F. Tan, and G. Durisi. 2017. The dispersion of nearest-neighbor decoding for additive non-gaussian channels. *IEEE Transactions on Information Theory* 63, 1 (January 2017), 81–92.

[26] Z. Wu, F. Zhao, and X. Liu. 2017. Signal Space Diversity Aided Dynamic Multiplexing for eMBB and URLLC Traffics. In *Proceedings of 3rd IEEE International Conference on Computer and Communication*. 1396–1400.

[27] K. Ying, J. M. Kowalski, T. Nogami, Z. Yin, and J. Sheng. 2018. Coexistence of enhanced mobile broadband communications and ultra-reliable low-latency communications in mobile front-haul. *Proc.SPIE* 10559 (January 2018), 10559 – 10559 – 7.

[28] L. Zhou. 1990. On a conjecture by Gale about one-sided matching problems. *Journal of Economic Theory* 52, 1 (October 1990), 123–135.