# Resource Scheduling for Mixed Traffic Types with Scalable TTI in Dynamic TDD Systems

Qi Liao\*, Paolo Baracca\*, David Lopez-Perez<sup>†</sup> and Lorenzo Galati Giordano<sup>†</sup>

\*Nokia Bell Labs, Stuttgart, Germany

<sup>†</sup>Nokia Bell Labs, Dublin, Ireland

Email: {qi.liao, paolo.baracca, david.lopez-perez, lorenzo.galati\_giordano}@nokia-bell-labs.com

Abstract—This paper analyses the performance benefits of a user-centric scheduling approach, exploiting the flexibility of both dynamic time division duplex (TDD) and a variable transmission time interval (TTI), where the downlink to uplink ratio and TTI duration can be adapted to the traffic load. The formulation of the joint optimisation problem takes into consideration the individual requirements of each single user in terms of sustainable latency and desired throughput, thus implementing a real user-centric scheduling approach. Moreover, the developed solution is evaluated in a scenario with mixed traffic types, mobile broadband (MBB) and mission critical communications (MCC), showing remarkable performance enhancement of the proposed scheme over baseline dynamic TDD schemes with a fixed TTI in terms of both achievable throughput of the MBB users and guaranteed latency for the MCC users.

Keywords: 5G, dynamic TDD, scalable TTI, user-centric scheduler, mixed traffic types.

#### I. INTRODUCTION

One of the main drivers of next generation mobile networks is the future heterogeneity of services and applications [1]. Besides the support for mobile broadband (MBB), the fifth generation (5G) systems should also manage machine type of communications (MTC), which are mostly characterised by small packet transmissions, and have very different requirements than MBB traffic. For example, two representative use cases of MTC are massive machine communications (MMC) and mission critical communications (MCC) [2]. While MMC applies to scenarios with a huge number of lowcost devices transmitting only sporadically with quite relaxed latency requirements but strict energy constraints, MCC refers to scenarios where nodes have low to moderate throughput demands but stringent latency and reliability demands [3]. This traffic diversity calls for a thorough reappraisal of contemporary wireless network technologies, and the change from a base station (BS)-centric to a user equipment (UE)-centric networking paradigm, where only the necessary resources, no more, no less, are allocated to each UE to satisfy its throughput, latency and energy requirements.

Recent studies have shown that time division duplex (TDD) represents a more flexible option than frequency division duplex (FDD) system to manage this traffic heterogeneity and realise such UE-centric networking [4]. In this line, the third generation partnership project (3GPP) has already introduced a number of semi-static TDD configurations in the

last LTE releases that can be dynamically selected by the BS to deal with traffic burstiness [5]. Moreover, the 3GPP has also commenced to consider the support for shorter and variable time transmission intervals (TTIs), which will reduce latency at the physical and medium access control layers, and further help to provide a tailored amount of resources to UEs, while avoiding any waste [2].

Dynamic TDD allows a BS to dynamically adapt the downlink (DL) to uplink (UL) ratio to the current traffic load. It has been shown in several works that BS clustering, a method that groups the cells to adopt the same DL to UL ratio when they are characterized by high interference [6] and have similar traffic profile and buffer size [7], can cope with DL-to-UL and UL-to-DL interference and provide good throughput for MBB users at a reasonable complexity. However, these works do not take into account different services classes. Motivated by the flexible frame structure in 5G, [8] proposed to fulfil the strict delay constraints for MCC users by dynamically selecting the TTI length in FDD system, using reverse calculation based on the delay budgets. Along similar lines, in [9] the authors apply the variable frame duration to achieve the ultra-low latency in millimeter-wave communication. Both works use the strategy that short TTI is selected first when MCC users exist in the system, followed by long TTI lengths when reaching steady state operation. However, such schemes always provide priority to the MCC users to fulfil their latency constraints, while sacrificing the throughput performance of the others.

In this paper, we aim at developing a true user-centric approach that provides a flexible tradeoff between mixed types of services to meet their specific requirements in both UL and DL for dynamic TDD systems. To the best of the authors' knowledge, this is the first work that develops a dynamic TDD framework for a scenario with mixed types of services (where UEs generate either MBB or MCC traffic in both UL and DL), considering scalable TTI lengths and individual UE requirements in terms of both throughput and latency. Our proposed user-centric approach decides for each scheduling time: a) the duplexing mode, i.e., either DL or UL, b) the TTI length, and c) the UEs to be served and the resources allocated in each TTI. Numerical results show that the proposed scheme significantly outperforms dynamic TDD schemes with a fixed TTI in terms of both throughput provided to the MBB UEs and latency guaranteed to the MCC UEs. When compared to the schemes that always admit high priority to MCC UEs, our proposed scheme can achieve comparably low latency for MCC UEs, while providing good throughput also for the MBB UEs.

The rest of the paper is organised as follows. In Section II, the system model is described together with the correspondent notation. In Section III, the problem formulation is introduced with an explanation of the different utility functions and the correspondent optimal solution. Finally, in Section IV, the numerical results are presented, followed by the concluding remarks drawn in Section V.

#### II. SYSTEM MODEL

We will use the following notation throughout the paper. Bold capital letters and bold lowercase letters denote matrices and vectors, respectively, while  $\mathcal{X}$  denotes a set, and  $\mathbf{x} \succeq 0$ implies that  $x_i \ge 0$  for all components *i*. Moreover, the set of non-negative real numbers and the set of positive real numbers are denoted by  $\mathbb{R}_+$  and  $\mathbb{R}_{++}$ , respectively. The cardinality of set  $\mathcal{X}$  is denoted by  $|\mathcal{X}|$ , and the Cartesian product of two finite sets  $\mathcal{X}$  and  $\mathcal{Y}$  is denoted by  $\mathcal{X} \times \mathcal{Y}$ .

We consider a single cell scenario where we assume that the system operates in a dynamic TDD mode, and that the average received co-channel inter-cell interference can be estimated. Under this setup, let the TTI be discrete, and indexed by  $n \in \{0, 1, \ldots\}$ , where the length of the TTI  $\Delta(n)$  is scalable as shown in Fig. 1 and can be chosen from a finite set of M lengths  $\mathcal{T} = \{\Delta_i : i = 1, \ldots, M\}$ . In addition, let  $t_n$  denote the time point at which the *n*th TTI begins (which is also the time point at which the (n-1)th TTI finishes if n > 0).

At the *n*th TTI, the set of existing services (including both uplink and downlink)<sup>1</sup> is denoted by  $S(n) := \{s : s = 1, \ldots, S(n)\}$ . Now and hereafter, we omit the index *n* when we can do so without any ambiguity. The set *S* comprises the subsets  $S^{(d)}$ ,  $d \in D$ , where  $D := \{\text{UL}, \text{DL}\}$  denotes the duplexing mode of uplink or downlink. Since a transmission link can be either uplink or downlink in the dynamic TDD system, we separate the uplink and downlink services such that  $S^{(\text{UL})} \cup S^{(\text{DL})} = S$  and  $S^{(\text{UL})} \cap S^{(\text{DL})} = \emptyset$ , where  $|S^{(d)}| = S^{(d)}$  for  $d \in D$  denotes the cardinality of set  $S^{(d)}$ .

Our target is to construct a scheduler which selects at each TTI, a duplexing mode  $d \in \mathcal{D}$  and a TTI length  $\Delta \in \mathcal{T}$ , and schedules *one or more services* in the selected duplexing mode <sup>2</sup>. Under this setup, let  $\mathbf{p} := (p_1, \ldots, p_S)^T \in \mathcal{P} := [0, 1]^S$  denote the fraction of frequency resources assigned to services, which should satisfy the following resource constraint  $\sum_{s \in S(n)} p_s(n) \leq 1$ .

## A. Achievable Rate

We assume that the channel state remains the same within a TTI (regardless of the TTI lengths), and that the coherence

<sup>2</sup>Since we conduct our study in a single cell scenario in this paper, TTI length and duplexing mode are selected regardless the configuration of the neighboring cells. Frame alignment and inter-cell inter-mode interference (between uplink and downlink) will be considered in the follow-up work.



Fig. 1: Dynamic TDD frame structure with scalable TTI.

time of the channel lasts in general more than one TTI. Further, we assume that signal to interference plus noise ratio  $\gamma_s(n)$ can be estimated for each transmission link of service s and for each TTI n (perfect knowledge of transmission power, channel states, received interference and noise power).

In the following, we introduce the achievable rate considering the impact of the control signalling overhead. In fact, the TTI length is configured as a multiple of a number of orthogonal frequency division multiplexing (OFDM) symbols, and the signalling overhead is fixed. Thus, although some services can be scheduled with a short TTI to fulfil their strict latency requirement, this short TTI comes at the expense of higher control signalling cost and lower capacity for the services requiring higher data rates [2].

Taking the signalling overhead into account, the achievable data rate of service s at the nth TTI is given by

$$r_{s}(n) := r_{s} \left( p_{s}(n), d(n), \Delta(n) \right)$$
  
=  $\mathbb{1}_{\{s \in \mathbb{S}^{(d(n))}(n)\}} \psi(\Delta(n)) p_{s}(n) B \log \left( 1 + \gamma_{s}(n) \right)$  (1)

where  $\mathbb{1}_{\{A\}}$  is the indicator function, which equals to 1 if the event A occurs, and zero otherwise,  $S^{(d(n))}(n)$  is the set of services at duplexing mode d(n) at the *n*th TTI and *B* denotes the total bandwidth. Moreover,  $\psi(\cdot)$  is a function indicating the ratio of the effective time to transmit the payload. For example, if we assume that a fixed time duration  $\tau$  is required for the control signal transmission on the dedicated channel for any TTI length, where  $\tau$  is smaller than the shortest TTI length, we can then define  $\psi: T \to (0, 1]: x \mapsto (x - \tau)/x$ .

## B. Time-Varying Service Demand

Let  $t'_s$  denote the arriving time of service s, and  $n'_s$  denote the index of the TTI that contains the time instant  $t'_s$ , i.e.,  $t'_s \in [t_{n'_s}, t_{n'_s+1})$ . Each new incoming service s arrives with a traffic demand of  $\nu_s(n'_s)$  (in bits) and a latency constraint of  $l_s(n'_s)$  (in seconds). As the scheduler allocates resources to serve it, the remaining service demand decreases if the data transmission is successful. Note that service s cannot be served before the next TTI (with the index  $n'_s + 1$ ) begins.

By the end of the *n*th TTI with  $n > n'_s$ , the remaining traffic demand of service s is given by

$$\nu_s(n) = \nu_s(n-1) - \Delta(n)r_s(n). \tag{2}$$

It is desired that the allocated resources just meet the traffic demand, but do not exceed it, otherwise, it is a waste of

<sup>&</sup>lt;sup>1</sup> Here a service refers to the communication between two nodes that occurs during the span of a single connection. For example, a service of web browsing (in downlink) starts when a user requests to connect to one URL and ends when all information from that URL is displayed.

resources. Thus, we define an additional constraint to avoid negative values of  $\nu_s(n)$ , expressed as

$$\nu_s(n-1) - \Delta(n)r_s(n) \ge 0. \tag{3}$$

The remaining latency constraint for  $n > n'_s + 1$ , i.e., the maximum remaining time to fulfil the traffic demand, is written as

$$l_s(n) = (l_s(n-1) - \Delta(n))^+.$$
 (4)

where  $(\cdot)^+ := \max\{\cdot, 0\}$ . For  $n = n'_s + 1$ , we have that  $l_s(n) = (l_s(n-1) - (t_{n'_s+1} - t'_s))^+$ , since service s cannot be served before the  $(n'_s + 1)$ th TTI begins. Note also that  $l_s(n) = 0$  implies that the latency constraint has expired.

Finally, a service is removed from the system, if the remaining traffic demand yields  $\nu_s(n) = 0$ .

#### **III. PROBLEM FORMULATION AND OPTIMAL SOLUTION**

By optimising the variables  $\{\mathbf{p}(n), d(n), \Delta(n)\}\$  at the *n*th TTI, our objective is to design a computationally efficient scheduling algorithm for mixed traffic types, which can work on a very short time scale and achieve a good tradeoff between heterogeneous performance metrics with respect to traffic demands and latency requirements.

#### A. Latency-Aware Cost Function

To penalise the undesirable long latency period for the MCC services characterised by strict low latency requirements, we introduce a cost function defined as follows.

$$J_{s}(n) := J_{s}(p_{s}(n), d(n), \Delta(n)) \\ = \frac{\frac{\nu_{s}(n)}{\overline{R}_{s}(n)} + (\Delta(n) - l_{s}(n-1))^{+}}{\max\{l_{s}(n), c_{\min}\}}$$
(5)

where  $\nu_s(n)$  depending on  $(p_s(n), d(n), \Delta(n))$  is given by (2),  $0 < c_{\min} \ll \Delta_{\min}$  is arbitrarily small to prevent divideby-zero errors, and  $\Delta_{\min} := \min_{\Delta} \mathcal{T}$  is the minimum TTI length. At the numerator, the term  $\nu_s(n)/\overline{R}_s(n)^3$  serves as the **best-case estimate of the time to serve the remaining traffic** by the end of the *n*th TTI, with  $\overline{R}_s(n)$  defined as the best-case serving rate after the *n*th TTI, while the term  $(\Delta(n) - l_s(n-1))^+$  is an **offset** to guarantee that assigning a TTI longer than the required latency constraint results in a high cost, even if the remaining traffic can be served within the chosen TTI (i.e.,  $\nu(n) = 0$ ). At the denominator, the term  $\max\{l_s(n), c_{\min}\}$  is the remaining latency constraint indicating the **maximum required time to serve the remaining traffic** by the end of the *n*th TTI. It is important to note that  $\overline{R}_s(n)$  can be estimated by averaging the maximum achievable rate  $B \log(1+\gamma_s(j))$  over the recent time period of successive TTIs  $j \in [n'_s, n]$ . To better describe the physical meaning of the cost function in (5), we provide a simple numerical example here below.

**Example 1.** Assume that at the initial time point, a MBB service (s = 1) and a MCC service (s = 2) arrive in uplink with the traffic demands  $\nu_1(0) = 10^3$  and  $\nu_2(0) = 2$  (in bits), and that the latency constraints are  $l_1(0) = 5$  and  $l_s(0) = 0.25 \cdot 10^{-3}$  (in seconds), respectively. Define  $c_{\min} = 10^{-6}$ . Assume that the achievable rate remains the same for the next  $10^{-3}$  seconds and we have  $r_1(1) = 10^4$  and  $r_2(1) = 8 \cdot 10^3$  (in bit/s) if the full bandwidth is allocated (i.e.,  $p_1 = 1, p_2 = 0$  or  $p_1 = 0, p_2 = 1$ ), respectively. We consider the following three cases of the configuration for the 1st TTI.

1) Configuration  $\Delta(1) = 10^{-3}$ , d(1) = UL,  $p_1(1) = 1$ ,  $p_2(1) = 0$ . We have  $J_1(1) = ((10^3 - 10)/10^4 + 0)/(5 - 10^{-3}) \approx 0.0198$ , while  $J_2(1) = (2/(8 \cdot 10^3) + 0.75 \cdot 10^{-3})/10^{-6} = 10^3$ . In this case, using a long TTI and allocating all resources to the MBB service cause high latency cost of service 2.

2) Configuration  $\Delta(1) = 10^{-3}$ , d(1) = UL,  $p_1(1) = 0.75$ ,  $p_2(1) = 0.25$ . We have  $J_1(1) = ((10^3 - 7.5)/10^4 + 0)/(5 - 10^{-3}) \approx 0.0199$ , while  $J_2(1) = (0 + 0.75 \cdot 10^{-3})/10^{-6} = 750$ . In this case, although for service 2 the remaining traffic is served within one TTI, its latency cost is still high due to the violation of the latency constraint.

3) Configuration  $\Delta(1) = 0.25 \cdot 10^{-3}$ , d(1) = UL,  $p_1(1) = 0$ ,  $p_2(1) = 1$ . We have  $J_1(1) = (10^3/10^4 + 0) / (5 - 0.25 \cdot 10^{-3}) \approx 0.02$  while  $J_2(1) = 0$ . Compared to the first configuration,  $J_1(1)$  increases only slightly due to the loose latency constraint of the MBB service, while  $J_2(1)$  reduces to zero because the strict latency requirement of the MCC service is fulfilled. Thus, the third configuration leads to a lower joint latency cost  $J_1(1) + J_2(1)$ .

### B. Utility-Based Resource Allocation Problem

In this section, we formulate a utility-based optimisation problem, where the objective metric with the throughput utility added and the latency-aware cost function subtracted is maximised. Omitting the TTI index n for convenience, the objective function to maximise is defined as

$$U(\mathbf{p}, d, \Delta)$$
  
:=  $\sum_{s \in \mathbb{S}} \left( \alpha_s u_s \left( r_s(d, \Delta, p_s) \right) - \beta_s J_s(d, \Delta, p_s) \right)$  (6)

where the concave utility function  $u_s : \mathbb{R}_+ \to \mathbb{R}_+ : x \mapsto \log(1+x)$  is applied to achieve throughput fairness among the services, and  $\alpha_s \in \mathbb{R}_+$  and  $\beta_s \in \mathbb{R}_+$  for  $s \in S$  are the service-specific weight factors to give different priorities to services with different rate or latency requirements, respectively. Such service-specific weight factors are fed back from

<sup>&</sup>lt;sup>3</sup> The term  $\nu_s(n)/\overline{R}_s(n)$  in (5) is designed as the best-case estimate of the time to serve the remaining traffic. The myopic optimisation scheme cannot estimate the remaining delay because it depends on the scheduling decision in the future. However, we can offer a best-case estimate for the remaining serving time under the optimal condition that a service is scheduled with full resource allocation. This relaxes the penalty for the MBB services because there is usually some room to fulfill the demand in the upcoming TTIs if  $l_s(n-1)$  is large enough, while it raises the cost of MCC services if a long TTI is used, leading to a small value of  $l_s(n) = (l_s(n-1) - \Delta(n))^+$ .

the UE to the BS, and they are the mean to achieve a usercentric networking. For example, a higher  $\alpha_s$  can be defined for the MBB services with high throughput demands, or a higher  $\beta_s$  can be defined for the MCC services with strict latency requirements. The optimisation of the values of  $\alpha_s$ and  $\beta_s$  for each service is outside the scope of this paper.

The optimisation problem for each TTI over a set of variables  $(\mathbf{p}, d, \Delta)$  is written as

**Problem 1**[Joint optimisation problem]

$$\max_{\mathbf{p},d,\Delta} \quad U(\mathbf{p},d,\Delta) \tag{7a}$$

s.t. 
$$(1), (2), (3), (4), (5)$$
 and  $(6)$  (7b)

$$d \in \mathcal{D}, \Delta \in \mathcal{T} \tag{7c}$$

$$\mathbf{p} \succeq 0, \sum_{s \in \mathbb{S}^{(d)}} p_s \le 1, \ \sum_{s \in \mathbb{S} \setminus \mathbb{S}^{(d)}} p_s = 0 \qquad (7d)$$

where (7d) guarantees that uplink services and downlink services cannot be served at the same TTI in the same cell.

Knowing that  $|\mathcal{D}| = 2$  and assuming that we only have limited choice of TTI lengths with  $|\mathcal{T}| = M$ , in order to simplify the joint optimisation problem over the variables  $(\mathbf{p}, d, \Delta)$ , we can optimise  $\mathbf{p}$  in Problem 1 for each combination of fixed parameters  $(d', \Delta')$ , and find the optimum solution to the following subproblem.

**Problem 2**[Subproblem with fixed  $(d', \Delta')$ ]

$$\mathbf{p}'(\Delta', d') = \operatorname*{arg\,max}_{\mathbf{p}} U(\mathbf{p}, d', \Delta')$$
s.t. (1), (2), (3), (4), (5), (6) and (7d)
(8)

where  $(d, \Delta)$  in the constraints are replaced by  $(d', \Delta')$ .

Then, the optimum solution to the general Problem 1, denoted by  $(d^*, \Delta^*, \mathbf{p}^*)$ , can be obtained by searching for the maximum utility U over the solutions to Problem 2 with respect to every combination of  $(d, \Delta)$ , i.e.,

$$(d^{\star}, \Delta^{\star}) = \operatorname*{arg\,max}_{d' \in \mathcal{D}, \Delta' \in \mathcal{T}} U\left(\mathbf{p}'(d', \Delta'), d', \Delta'\right)$$
(9)

$$p^{\star} = p'(d^{\star}, \Delta^{\star}) \tag{10}$$

where  $U(\mathbf{p}'(d', \Delta'), d', \Delta')$  in (9) is derived from the optimum solution to Problem 2, and  $p'(d^*, \Delta^*)$  is the optimum solution of  $\mathbf{p}$  with respect to the optimum parameters  $(d^*, \Delta^*)$ found by (9).

## C. Solution to the Optimisation Problem

Due to the limited number of the combinations of  $\mathcal{D}$  and  $\mathcal{T}$  with  $|\mathcal{D} \times \mathcal{T}| = 2M$ , we can find (9) and (10) by exhaustive search with the computational complexity  $\mathcal{O}(2M)$ . Then, the remaining task is to solve Problem 2.

With fixed d', (7d) implies that  $p_s = 0$  for all  $s \in S \setminus S^{(d')}$ . To optimise  $p_s$  for  $s \in S^{(d')}$ , let us write the equivalent problem of Problem 2 here below.

For convenience's sake, we denote the vector of fraction of resources assigned to the services with duplexing mode d' by  $\mathbf{x} := (p_s : s \in \mathcal{S}^{(d')})^T \in [0, 1]^{S^{(d')}}$ . The achievable rate in

(1) and the cost function in (5) are simply linear and affine functions of  $x_s$ , respectively.

$$r_s(x_s) := r_s(x_s(n)) = A_s x_s \tag{11a}$$

$$J_s(x_s) := J_s(x_s(n)) = B_s + C_s x_s$$
 (11b)

where

$$A_s := A_s(n) = \psi(\Delta') B \log(1 + \gamma_s(n))$$

$$(11c)$$

$$\psi(n-1)$$

$$B_s := B_s(n) = \frac{\frac{P_s(n-1)}{\overline{R}_s(n)} + (\Delta' - l_s(n-1))^+}{\max\{l_s(n), c_{\min}\}}$$
(11d)

$$C_s := C_s(n) = -\frac{\Delta' A_s}{\overline{R}_s(n) \max\{l_s(n), c_{\min}\}}$$
(11e)

Moreover, we define the constraint set

$$\mathfrak{X} := \{ \mathbf{x} \succeq 0 : x_s \le D_s, s \in \mathcal{S}^{(d')} \}$$
(12a)

with 
$$D_s := D_s(n) = \frac{\nu_s(n-1)}{\Delta' A_s(n)}$$
 (12b)

where (12b) is derived from constraint (3).

In addition, by taking only  $s \in S^{(d')}$  into account, the objective function in (6) can be simplified as

$$V(\mathbf{x}) = \sum_{s \in \mathcal{S}^{(d')}} \left( \alpha_s u_s \left( r_s(x_s) \right) - \beta_s J_s(x_s) \right).$$
(13)

Using (11), (12) and (13), the optimisation problem equivalent to Problem 2 is written in below.

Problem 3[Primal problem]

$$V^{\star} := \max_{\mathbf{x} \in \mathcal{X}} V(\mathbf{x}), \text{ s.t. } \sum_{s=1}^{S^{(d')}} x_s \le X_{\max}.$$
(14)

where the constraint set  $\mathcal{X}$  is given in (12), and  $X_{\max} := \min\{1, \sum_s D_s\}$  is obtained from the individual resource constraint  $x_s \leq D_s$  (such that the allocated resource is just sufficient to finish the remaining traffic demand and no resource is wasted) and the sum resource constraint  $\sum_s x_s \leq 1$ .

It is easy to verify that  $V(\mathbf{x})$  is concave in  $\mathbf{x}$ , because *i*)  $u_s \circ r_s$ , as the composition of a concave and monotone nondecreasing function and a concave function, is concave [10, pag. 84]; *ii*)  $J_s$ , as affine function, is both convex and concave and the sum of concave functions is concave [10, pag. 79]. Note that the constraint set defined by  $\mathcal{X}$  and  $\sum_s x_s \leq X_{\text{max}}$ is also convex. Thus, Problem 3 is a convex optimisation problem, which can be solved by looking at the dual formulation where there is no duality gap.

Define the Lagrangian  $L(\mathbf{x}, \lambda)$  for the primal problem in (14) and the *dual function* respectively by

$$L(\mathbf{x}, \lambda) = V(\mathbf{x}) + \lambda (X_{\max} - \sum_{s} x_{s})$$
(15)

$$L(\lambda) := \max_{\mathbf{x} \in \mathcal{X}} L(\mathbf{x}, \lambda).$$
(16)

The *dual problem* is then given in below. **Problem 4**[Dual problem]

$$L^{\star} := \min_{\lambda \ge 0} L(\lambda) \tag{17}$$

Applying the Lagrangian optimality, the feasible fraction of allocated frequencies  $\mathbf{x}^*$  that maximises the Lagrangian  $L(\mathbf{x}, \lambda)$  over  $\mathbf{x} \in \mathcal{X}$  is given by

$$x_s^*(\lambda) = \begin{cases} D_s, \text{ if } \lambda \le -\beta_s C_s \\ \left(\min\left\{\frac{\alpha_s}{\lambda + \beta_s C_s} - \frac{1}{A_s}, D_s\right\}\right)^+, \text{ o.w.} \end{cases}$$
(18)

The solution in (18) is derived by utilising that for any  $\lambda$ , the partial derivative  $\partial L(\mathbf{x}, \lambda)/\partial x_s = \alpha_s A_s/(1 + A_s x_s) - \beta_s C_s - \lambda$  is monotone decreasing over  $x_s \in [0, D_s]$ . If  $\partial L(\mathbf{x}, \lambda)/\partial x_s|_{x_s=D_s} \geq 0$ , the partial derivative is nonnegative at each point  $x_s \in [0, D_s]$ . Thus,  $x_s = D_s$  maximises  $L(\mathbf{x}, \lambda)$  over  $x_s \in [0, D_s]$  for any  $\lambda \leq \alpha_s A_s/(1 + A_s D_s) - \beta_s C_s$ . Along similar lines, we have that  $x_s^*(\lambda) = 0$ , if  $\partial L(\mathbf{x}, \lambda)/\partial x_s|_{x_s=0} \leq 0$ , while  $x_s^*(\lambda) = \alpha_s/(\lambda + \beta_s C_s) - 1/A_s$ if  $\partial L(\mathbf{x}, \lambda)/\partial x_s|_{x_s=0} \geq 0$  and  $\partial L(\mathbf{x}, \lambda)/\partial x_s|_{x_s=D_s} \leq 0$ . Moreover, for  $-\beta_s C_s < \lambda \leq \alpha_s A_s/(1 + A_s x_s) - \beta_s C_s$ , we have that  $\alpha_s/(\lambda + \beta_s C_s) - 1/A_s \geq D_s$ , and thus the second case in (18) also returns  $x_s^*(\lambda) = D_s$  for  $-\beta_s C_s < \lambda \leq \alpha_s A_s/(1 + A_s x_s) - \beta_s C_s$ . As a result, the solution can be written in the form of (18).

The complementary slackness provides  $\lambda^*(X_{\max} - \sum_s x_s^*) = 0$ . Thus,  $\sum_s x_s^* = X_{\max}$  if  $\lambda^* > 0$ .

The solution to (17) can be given by numerically minimising  $L(\lambda)$  over  $\lambda \ge 0$ . For this, we use a subgradient-based search, and update  $\lambda$  iteratively by

$$\lambda(t+1) = \left(\lambda(t) - \kappa(t) \left(X_{\max} - \sum_{s} x_{s}^{*}(\lambda(t))\right)\right)^{+}$$
(19)

where  $x_s^*(t)$  is updated by (18) and  $\kappa(t)$  is the step size at iteration t.

The algorithm of iteratively performing (19) and (18) converges, if  $\kappa(t)$  is chosen appropriately [11, Section 6.3.1]. There are many ways to select the step size. For constant step size, the subgradient algorithm is guaranteed to converge within some range of the optimal value [12]. Note that, from (18), if  $\lambda \geq \max_s(\alpha_s A_s - \beta_s C_s)$ , no service will be scheduled. Therefore, the optimal  $\lambda$  lies in the interval  $[0, \max_s(\alpha_s A_s - \beta_s C_s)]$ . In order to provide a good starting point for the fast convergence to the solution, we can numerically minimise the univariate function  $L(\lambda)$  over a finite set of uniformly distributed  $\lambda$  in  $[0, \max_s(\alpha_s A_s - \beta_s C_s)]$ , and choose the  $\lambda$  corresponding to the minimum value of  $L(\lambda)$  as the starting point.

Given the optimal  $\lambda^*$  and the corresponding  $x_s^*(\lambda^*)$  for all  $s \in S^{(d')}$ , the solution to the equivalent problem (8) can be obtained by collecting  $x_s^*(\lambda^*)$  for  $s \in S^{(d')}$  and  $p_s = 0$  for  $s \in S \setminus S^{(d')}$  in the joint vector  $\mathbf{p}'$ .

Fig. 2 shows that the subgradient-based searching algorithm based on (18) and (19) converges to the optimal solution  $x_1 =$  $0, x_2 = 1$  for the previous introduced Example 1. Small  $\beta$ s are chosen because the cost function has a much larger scale than the concave utility of rate in this example. We have  $\beta_2 > \beta_1$ because user 2 has higher latency requirement. Using  $\Delta =$ 



Fig. 2: Subgradient-based search over  $\{\mathbf{x} : x_s \in [0, D_s], \forall s; \sum_s x_s \leq X_{\max}\}$  for Example 1.  $\alpha_1 = 0.9, \alpha_2 = 0.85, \beta_1 = 0.1, \beta_2 = 0.15, \kappa = 1.$ 

0.25 ms achieves generally higher utility than using  $\Delta = 1$  ms because the latter violates the latency constraint and results in a higher cost. The best solution with respect to  $\Delta = 1$  ms is  $x_1 = 0.75, x_2 = 0.25$ .

#### **IV. NUMERICAL RESULTS**

In this section, we analyse the performance of the proposed user-centric scheduler, by jointly considering the delay of the MCC services and the throughput of the MBB services. In detail, we compare our proposed scheme employing the set  $\mathcal{T} = \{0.1 \text{ ms}, 0.2 \text{ ms}, \dots, 1 \text{ ms}\}$  of scalable TTI lengths against a baseline scheduler optimising the duplexing mode and the resource allocation between users to maximise the same utility (6) but with fixed TTI lengths. The tradeoff between the performance of MCC and MBB services can be flexibly tuned with parameters  $\alpha^{(MBB)}, \beta^{(MBB)}, \alpha^{(MCC)}, \beta^{(MCC)}$ . which denote the service-specific weight factors  $\alpha_s$  and  $\beta_s$  for MBB and MCC services, respectively. The joint selection of  $[\alpha_s, \beta_s]$  allows the scheduler to perform a truly user-centric allocation of the available resources. In the following results, reasonable values of  $\alpha_s$  and  $\beta_s$  have been selected for the two main service categories, MBB and MCC.

We focus our analysis on the isolated pico BS scenario defined in [5, Sect. 6.2] where the transmit power of both BSs and UEs are set to have an average signal to noise ratio at the cell edge of 10 dB. All the other simulation parameters mainly related to line-of-sight probability, path-loss and shadowing can be found in [5, Tab. 6.2-1]. Moreover, we assume that the BS serves 2 MBB UEs active in the DL, modelled as full buffer UEs, and 10 MCC UEs active in UL, which generate small packets of size 125 bytes [13], whose arrival rates are Poisson distributed with parameter  $\eta$ . Then, we further assume that the control signal transmission requires  $\tau = 0.05$  ms.

1) Adaptation of duplexing, TTI length and scheduled services over time. In Fig. 3, we report the number of bits transmitted over time to show how our proposed algorithm is able to adapt to the MCC packet arrival, well selecting both the duplexing (UL or DL) and the TTI length.

By setting a high value for  $\beta^{(MCC)}$ , e.g.,  $\beta^{(MCC)} = 3$ , we increase the priority of MCC services and more resources are



(b) Adaptation of duplexing with  $\eta = 500$  packet/s.

Fig. 3: Dynamic duplexing and TTI length adapted to packet arrival with  $\alpha^{(\text{MBB})} = 1$ ,  $\beta^{(\text{MBB})} = 0.2$ ,  $\alpha^{(\text{MCC})} = 0.01$ ,  $\beta^{(\text{MCC})} = 3$ , and  $c_{\min} = 10^{-6}$ .

allocated to them in UL. As a consequence, when we increase the packet arrival rate from  $\eta = 100$  packet/s in Fig. 3a to  $\eta = 500$  packet/s in Fig. 3b, we observe that more UL TTIs are scheduled in order to fulfil the latency requirements of the MCC services, causing less resource allocation to MBB services in DL.

2) Predefining latency requirements for MCC services. For each MCC user/service-specific requirement, we can define a specific latency constraint (see also (4)). Fig. 4 shows the cumulative distribution function (CDF) of the delay of the MCC packets for different values of the latency constraint: these results show that the proposed resource allocation scheme manages to meet this latency constraint in more than 98% of the cases when it is larger than 1 ms. For a more strict latency constraint, e.g. 1 ms, the scheme meets it in almost 90% of the cases: this happens because a service cannot be scheduled before the beginning of the next TTI. For instance, in case a service arrives within the first 0.1 ms of the nth TTI and  $\Delta(n) = 1$  ms, it has to wait for more than 0.9 ms to be scheduled. Even if the shortest length  $\Delta(n+1) = 0.1$  ms is chosen for the next TTI, the latency constraint of 1 ms cannot be met.

3) Inappropriate selection of fixed TTI length leads to capacity loss. Assuming that high priority is given to the MCC services, the throughput of MBB services in DL mainly depends on the following two factors: *a*) the overhead cost  $\psi(\Delta)$ , and *b*) the probability that there exist MCC services



Fig. 4: CDF of delay of MCC packets depending on predefined latency constraints with  $\eta = 100$  packet/s,  $\alpha^{(\text{MBB})} = 1$ ,  $\beta^{(\text{MBB})} = 0.2$ ,  $\alpha^{(\text{MCC})} = 0.01$ ,  $\beta^{(\text{MCC})} = 3$ , and  $c_{\text{min}} = 10^{-6}$ .

within the duration of a TTI. This probability is denoted by  $P(\Delta) := \Pr\{N(t) > 0 | t \in [t', t' + \Delta]\},$  where N(t) is the number of MCC services at time t, and t' is an arbitrary time instant. It is important to note that a longer TTI length  $\Delta$  reduces the throughput loss due to the proportionally less overhead related to the control signal transmission. However, as the packet arrival rate  $\eta$  increases, larger  $\Delta$  also causes much higher  $P(\Delta)$ , thus longer time periods will be occupied by the MCC services than actually needed. For instance, if 10 MCC packets arrive uniformly on a time interval of 10 ms, then with  $\Delta = 1 \,\mathrm{ms}$ , all slots may be allocated to MCC services and leave the MBB services with no resources, while by configuring  $\Delta = 0.2 \,\mathrm{ms}$ , only 2 ms will be allocated to the 10 MCC packets, and the remaining 8 ms can be allocated to MBB services. In Fig. 5, we report the minimum MBB user rate for different values of  $\eta$  and show that fixed TTI length cannot cope with the tradeoff between  $\psi(\Delta)$  and  $P(\Delta)$ , while our proposed scheme with dynamic TTI adaptively finds a good tradeoff and provides an enhanced throughput performance to MBB services.

4) Flexible tradeoff between delay of MCC services and throughput of MBB services by tuning parameters. From Fig. 5, we notice that by giving high priority to MCC services to fulfil their latency requirements, we sacrifice the throughput of MBB services as  $\eta$  increases. However, if we aim to provide an enhanced throughput performance to MBB services, while accepting a slight performance degradation on delay of MCC services, we can reduce  $\beta^{(MCC)}$  (or, similarly, increase  $\alpha^{(MBB)}$ ). Fig. 6 shows the CDF of delay of the MCC packets and the minimum MBB user rate for different values of  $\beta^{(MCC)}$ . First of all, we observe that although the baseline scheme with fixed  $\Delta = 0.2 \,\mathrm{ms}$  is able to guarantee a lower delay to the MCC services when compared to our scalable TTI proposal (see Fig. 6a), it also strongly loses in the MBB user rate (see Fig. 6b), mainly because of the higher impact of the control signalling overhead with very short TTI length. Moreover, by selecting a low  $\beta^{(MCC)} = 0.3$ , the system achieves an operating point with robust throughput of MBB services and acceptable slightly longer delays of MCC



Fig. 5: Minimum average MBB user throughput for different values of  $\eta$ ,  $\beta^{(MCC)} = 3$ .

services with respect to the case with  $\beta^{(MCC)} = 3$ . It is also worth mentioning that the proposed scheme with scalable TTI significantly outperforms the fixed  $\Delta = 1$  ms (configured TTI length in LTE) in terms of both delay and throughput.

Fig. 6 allows us also to quantify better the benefits of the proposed scheme with scalable TTI against the baseline with fixed TTI length. For example, by comparing our proposal against a scheme with fixed TTI of 0.2 ms and 1 ms respectively for  $\eta = 300$  packet/s and  $\beta^{(MCC)} = 0.3$ , we observe that both the scalable TTI and fixed  $\Delta = 0.2 \,\mathrm{ms}$ provide comparable performance of latency below 2 ms to all the MCC users, while the fixed  $\Delta = 1 \,\mathrm{ms}$  provides the same latency to only 90% of the MCC users. However, the scalable scheme provides 20% gain in the average MBB user throughput when compared to fixed  $\Delta = 0.2 \,\mathrm{ms}$  and 40% gain when compared to fixed  $\Delta = 1$  ms. Similarly, for a target MBB user throughput of about 12 Mbps, the baseline scheme with fixed  $\Delta = 0.2$  ms can support up to only  $\eta = 100$  packet/s for the MCC users, whereas our dynamic proposal can cope with more than  $\eta \geq 300$  packet/s, providing a significant gain in the number of MCC packets that can be served with latency below 2 ms while simultaneously supporting MBB user throughput of 12 Mbps.

#### V. CONCLUSIONS

In this paper, in order to cope with mixed traffic types, we have presented a new user-centric scheduling approach based on a dynamic TDD framework and with flexible TTI length configuration capabilities. In this framework, we have defined service-specific weight factors ( $\alpha_s$ ,  $\beta_s$ ) to better address the heterogeneous rate or latency requirements characterising each user. The optimisation variables of our scheduling scheme are the selection of UL or DL direction, the TTI length and the services to be scheduled. Extensive simulations show the remarkable performance gains of the proposed scheduling approach with respect to one with fixed TTI lengths in terms of both rate achieved by the MBB users and latency guaranteed to the MCC users.

#### REFERENCES

[1] NGMN, "NGMN 5G white paper," Next generation mobile networks (NGMN), A deliverable by the NGMN Alliance, Feb. 2015.





(b) Minimum average MBB user throughput for different values of  $\eta$ .

Fig. 6: Tradeoff between delay of MCC packets and throughput of MBB users obtained by tuning  $\beta^{(MCC)}$ ,  $l^{(0)}_{(MCC)} = 1$  ms.

- [2] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequencydivision duplex cases," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53–59, Mar. 2016.
- [3] G. Durisi, T. Koch, and P. Popovski, "Towards massive, ultrareliable, and low-latency wireless communication with short packets," http://arxiv.org/abs/1504.06526, Mar. 2016.
- [4] E. Lähetkangas, K. Pajukoski, J. Vihriälä, G. Berardinelli, M. Lauridsen, E. Tiirola, and P. Mogensen, "Achieving low latency and energy consumption by 5G TDD mode optimization," in *Proc. IEEE International Conference on Communications (ICC)*, Sydney (Australia), Jun. 2014.
- [5] 3GPP, "Further enhancements to LTE Time Division Duplex (TDD) for Downlink-Uplink (DL-UL) interference management and traffic adaptation (Release 11)," 3rd Generation Partnership Project (3GPP), TR 36.828, Jun. 2012.
- [6] M. Ding, D. Lopez-Perez, A. Vasilakos, and W. Chen, "Dynamic TDD transmissions in homogeneous small cell networks," in *Proc. IEEE International Conference on Communications (ICC)*, Sydney (Australia), Jun. 2014.
- [7] P. Baracca, "Traffic profile based clustering for dynamic TDD in dense mobile networks," in *Proc. IEEE Vehicular Technology Conference (VTC Fall)*, Montreal (Canada), Sep. 2016.
- [8] K. Pedersen, F. Frederiksen, G. Berardinelli, and P. Mogensen, "A flexible frame structure for 5G wide area," in *Proc. Vehicular Technology Conference (VTC Fall)*, Boston (MA), Sep. 2015.
- [9] T. Levanen, J. Pirskanen, and M. Valkama, "Radio interface design for ultra-low latency millimeter-wave communications in 5G era," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Austin (TX), Dec. 2014.
- [10] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [11] D. P. Bertsekas, Nonlinear programming. Athena scientific, 1999.
- [12] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," *lecture notes of EE3920, Stanford University, Autumn Quarter*, 2003.
- [13] "Scenarios, requirements and KPIs for 5G mobile and wireless system," METIS D1.1, Apr. 2013.