Joint Realtime and Nonrealtime Flows Packet Scheduling and Resource Block Allocation in Wireless OFDMA Networks

Alireza Sharifian, *Member, IEEE*, Rainer Schoenen, *Senior Member, IEEE*, and Halim Yanikomeroglu, *Senior Member, IEEE*

***Abstract*—In this paper, we consider joint realtime (RT) and nonrealtime (NRT) flows packet scheduling and resource block (RB) allocation in orthogonal frequency division multiple access (OFDMA) wireless networks. Radio RBs in the OFDMA plane are to be distributed among RT and NRT flows. In the conventional approach, RT and NRT flows are served sequentially. This sequential approach is inefficient because an RT flow may presumably have enough time until its delay deadline while its channel is in deep fade. In this situation, the transmission of NRT flows with higher level of efficiency can be performed. Intuitively speaking, the conventional sequential approach is too conservative, which can be reengineered. We propose a novel joint RT and NRT flows disutility-based packet scheduling and RB allocation in a common pool of RBs. The proposed joint approach enlarges the effective capacity of the associated wireless system when compared with the separated pool of RBs. The joint approach is particularly relevant for improving voice over LTE (VoLTE). We use mean bit-rate, mean queue-length, and instantaneous queuing delay information, in addition to channel information, to match the demand and supply. Furthermore, we develop a novel model for input–output bit-rate behavior of the mixture of RT and NRT flows. This model sheds light on the identification of different load regions and understanding of the system in an intuitive manner. Our approach and methodology can be extended for broader quality-of-service (QoS) requirements and for the utility of future applications. Simulation results show that the proposed framework is able to unify the serving mechanism of the RT and NRT flows and is able to achieve higher admissible bit-rate when handling mixed RT and NRT flows, compared with various baselines.**

***Index Terms*—Joint approach, multiuser diversity, nonrealtime (NRT) flows, packet scheduling, quality of service (QoS), queuing delay, realtime (RT) flows, resource allocation.**

Manuscript received November 18, 2013; revised April 5, 2014 and September 11, 2014; accepted March 22, 2015. Date of publication April 23, 2015; date of current version April 14, 2016. This work was supported in part by Huawei Technologies Canada and in part by the Ontario Ministry of Economic Development and Innovations Ontario Research Fund-Research Excellence Program. The review of this paper was coordinated by Dr. M. Dianati.

A. Sharifian was with Carleton University, Ottawa, ON K1S 5B6, Canada. He is now with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: alirezash@ece.utoronto.ca).

R. Schoenen was with Carleton University, Ottawa, ON K1S 5B6, Canada. He is now with Hamburg University of Applied Sciences (HAW), 20099 Hamburg, Germany (e-mail: rainer.schoenen@haw-hamburg.de).

H. Yanikomeroglu is with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: halim@sce.carleton.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2015.2425883

I. Introduction

New Services, Importance of QoS, Overprovisioning, and Capacity Crunch: One of the main requirements for the next generation of wireless networks is that it should cost effectively provide guaranteed quality of service (QoS), particularly in terms of delay and bit-rate requirement, with ubiquitous high bit-rate coverage, when and where required [1]. Wireless networks are part of a highly complex heterogeneous interactive system, where consumers share limited radio resources for a broad range of services such as Voice over IP (VoIP), telemedicine, online games, industrial/home automation, wearable connected devices, Hulu, Netflix, and Chrome OS. The flows for these vastly different services require highly different QoS. Traditionally, QoS in cellular communications has been implemented with *overprovisioning* or through costly higher layer mechanisms and overheads. Overprovisioning results in a network design for its peak load, which makes the system highly inefficient. In this setup, when the network becomes congested (load approaching the capacity), conventional rate limiters or bandwidth throttling is used, which causes user dissatisfaction [2], [3]. Tomorrow's networks with more frequent congestion problems will not have the luxury of overprovisioning or using various forms of excessive overhead. Advanced access technologies, such as the Long-Term Evolution (LTE), are purely scheduled system based on orthogonal frequency division multiple access (OFDMA), which creates the opportunity to dynamically and efficiently exploit various types of diversity and to schedule for diverse requirements, instead of overprovisioning. The main question, then, is how to perform the packet scheduling and resource allocation efficiently to treat different flows with heterogeneous demands and heterogeneous wireless link qualities.

Conventional Sequential Approach: The packet-switched connections can generally be divided into realtime (RT) and nonrealtime (NRT) flows. Conventionally, the packet scheduling for RT flows is designed based on the earliest deadline first (EDF) [4]. On the other hand, some versions of the proportional fairness (PF) [5]–[7], or generalized PF (GPF) [8]–[10], algorithm are used for NRT flows. Note that the PF algorithm is queue blind; thus, it cannot be used properly for RT flows, and it is not stable with respect to queues [11].

Inefficiency of the Sequential Approach: In the conventional approach, resource allocation for RT and NRT flows is executed sequentially [12]–[14]. In other words, RT flows are served first, and if any resources are still available, NRT flows are served subsequently. The static priority separation, or sequential

approach, is inefficient. The reason for this inefficiency is that an RT flow may presumably have enough time until its deadline while its channel is in deep fade. In this situation, the transmission of other flows, and possibly an NRT flow with good channel condition, can take place. Moreover, NRT flows are not completely insensitive to delay, particularly to the mean delay. Therefore, joint RT and NRT flows packet scheduling and RB allocation, which achieves higher level of multiuser diversity, becomes important. The joint RT and NRT flows packet scheduling and RB allocation exploits both intra- and interclass opportunism across flows. Heterogeneous QoS requirements in time and among flows, different load conditions, limited capacity in comparison to demands, and more frequent congestions make the joint RT and NRT flows packet scheduling and RB allocation more significant.

Need for Reengineering the Architecture of Data and Voice Services: Because of the preexistent voice services and the gradual emersion of data traffic in the cellular networks, voice services are designed separately from packet-switched data in the pre-4G cellular networks, such as GSM, UMTS, and CDMA2000 [15]. In these networks, only NRT flows are subject to scheduling over the shared channel, and voice services are served in traditional circuit-switched mode over the dedicated channel [15]. This static separation [16] sacrifices multiuser diversity. Today's wireless networks, such as LTE and the upcoming LTE Advanced, are moving toward packet-switching and IP-flat architecture to serve a broad range of applications with many diverse requirements [17]. Designing the flexible resource allocation framework, which considers the new heterogeneity, is crucial. In fact, providing properly engineered differentiable data flows is more cost effective than providing voice and data in separation. In the flat architecture, voice will be one of the differentiated packet-switched data flows that the packet scheduling and resource allocation algorithm is responsible to guarantee its QoS requirements. With voice as an RT differentiated data flow, networks will have reduced access delay, shorter wake-from-idle time, and will be able to offer not only regular voice calls but also differentiated higher quality (audio or video) calls, as well as enhanced rich communication services. There are three main ways of implementation for the new IP-flat architecture in regard to voice services, namely, voice over LTE (VoLTE), circuit-switched fall back (CSFB), and simultaneous voice and LTE (SVLTE) [18]. Among these three, VoLTE is the only one that really allows the delivery of voice as a data flow, within the LTE data bearer. CSFB and SVLTE are still dependent on the old pre-4G architecture and fall back to the legacy 2G or 3G circuit switching in voice calls. CSFB and SVLTE have several inefficiencies in comparison with VoLTE, such as longer call access delay, more expensive handsets and access points, and high power consumption on handset [18]. With regard to VoLTE, the joint RT and NRT flows scheduling and allocation is becoming more important in improving the efficiency of the wireless networks. Note that, although there are some partial solutions for the integration of RT and NRT flows in the application layer, the packet scheduling and RB allocation in the center of the medium access control (MAC) layer is the main component to be designed efficiently for guaranteeing differentiated QoS.

Bit-Rate Driven Utilities and Queue Awareness: Utility-based packet scheduling and resource allocation was the subject of a number of previous investigations aimed at different goals, based on the channel information and the queue information (see, for example, [19]–[23] and the references therein). Many algorithms based on bit-rate driven utilities [5], [24]–[39] have been studied in the literature. These bit-rate driven utilities are relevant to commonly called infinite-backlog scenario, where the queues are assumed to be always full, independent of service. Nevertheless, bit-rate driven utilities are blind to the requirements of flows, particularly to the RT ones, where the requirements are more important. Even for NRT flows, the queue blindness of bit-rate driven utilities makes the system potentially unstable [11], [40]. Therefore, queue awareness should be a key feature in designing the joint RT and NRT flows scheduling and allocation.

Related Work: Prior studies on packet scheduling and RB allocation of RT and NRT flows mainly focus on a single QoS element, such as head-of-the-line (HOL) delay [11], [41], mean delay [42], [43], bit-rate [31], [32], bit-rate with Sigmoid utility [25], or packet loss ratio [44]. Two heuristics, which are called exponential rule (EXP) and modified largest weighted delay first (MLWDF), have been also studied in [45]–[47] based on the HOL delay. Another study is a joint channel- and queue aware scheduling maximum delay utility (MDU) [42], [43], where the Little's delay is utilized. Since the averaging is a low-pass filter, RT flows suffer when the Little's delay (which is essentially a mean delay) is used. In other words, the decisions based on the mean delay are not sensitive enough to the HOL delay (or equivalently, instantaneous delay) requirements. Another approach based on time-utility function (TUF) has been proposed in [48]–[50], where the idea is to force the packet scheduler to transmit at near the delay deadlines. This approach cannot not fully exploit the multiuser diversity and can increase the number of dropped packets due to the passed delay deadlines. A utility-based adaptive approach with traffic prioritization is studied in [51] and [52], where, at the expense of 15% loss in throughput, it gains a decreased delay variance (or equivalently improved delay fairness) among wireless terminals (WTs).

Previously, we used the mean delay driven disutilities in [53] to introduce and analyze the minmax mean delay fairness notion. Here, in this paper, we use a framework inspired from [53] by incorporating HOL delay, queue-length information, and bit-rate information jointly in the disutilities and design the disutilities to advance a unified framework for joint RT and NRT flows scheduling and allocation.

Paper Contributions: The main contributions in this paper can be summarized into two main items.

- We propose a novel joint RT and NRT flows packet scheduling and RB allocation based on HOL delay, queue-length information, bit-rate information, and channel information. The proposed approach responds to heterogeneous delay requirements for RT flows and manages NRT flows effectively within a *common pool of RBs*, rather than the sequential resource allocation of RT and NRT flows. Although the framework is designed for wireless networks, we note that, whenever there is heterogeneity

TABLE I
LIST OF SYMBOLS

Symbol	Definition
$\Phi_{\mathrm{RT}}, \Phi_{\mathrm{NRT}}, \Phi$	Set of RT flows, set of NRT flows, set of all flows.
N, T	Number of sub-channels, number of time slots per sub-channel.
T_b, W_b	Time span (in second), frequency span (in Hertz) of each RB.
ϕ, j, k	Flow index, sub-channel index, frame index.
$\mathsf{f}\left(\mathsf{SINR}_{\phi}^{(j)}[k]\right)$	Function describing the AMC table (from SINR values).
$\mathsf{SINR}_{\phi}^{(j)}[k], b_{\phi}^{(j)}[k]$	SINR value, AMC value (on sub-channel j, for flow ϕ, in frame k).
$x_{\phi}^{(j)}[k]$	Optimization variable (for sub-channel j, flow ϕ, in frame k).
$r_{\phi}[k], \bar{r}_{\phi}[k]$	Frame bit-rate, mean bit-rate (for flow ϕ, frame k).
$q_{\phi}[k], \bar{q}_{\phi}[k]$	Frame queue-length, mean queue-length (for flow ϕ, frame k).
$D_{\phi}^{\mathrm{joint}}(\cdot)$	Disutility function (for flow ϕ).
$\mathcal{C}^{\mathrm{PHY}}$	Feasible set for the optimization variable.
$D_{\mathrm{net}}^{\mathrm{joint}}(\cdot)$	Total network disutility.
$\mathbf{d}^{\mathrm{HOL}}[k], \bar{\mathbf{q}}[k], \bar{\mathbf{r}}[k]$	HOL-delay vector, mean queue-length vector, mean bit-rate vector.
$d_{\phi}^{\mathrm{HOL}}[k]$	HOL-delay (for flow ϕ, in frame k).
$t_{\phi}^{\mathrm{HOL}}[k]$	Arrival time stamp of the HOL packet (for flow ϕ, in frame k).
$d_{\phi}^{\mathrm{HOL}^{\max}}$	Delay deadline of HOL-delay (for flow ϕ).
$\Delta d_{\phi}^{\mathrm{HOL}}[k]$	Difference of HOL-delay and its deadline (for flow ϕ).
$\frac{\partial D_{\phi}^{\mathrm{joint}}(\cdot)}{\partial x_{\phi}^{(j)}[k]}$	Disutility gradient (for flow ϕ, on sub-channel j and frame k).

among resources, including (but not limited to) multiuser diversity or any large-scale or small-scale signal variation, the proposed joint RT and NRT flows packet scheduling and RB allocation can offer higher performance, in comparison with the sequential approach. Furthermore, the developed framework enables putting different algorithms in the literature into perspective (see Section IV). Our approach is also a joint decision making in terms of both packet scheduling and RB allocation in one shot.

- We developed a novel model for input–output bit-rate behavior in packet scheduling and resource allocation of the mixture of RT and NRT flows. The model elaborates on different capacity definitions (for describing the system with the heterogeneous traffic), as well as their dependence on the input load. This model sheds light on identifying underload region, overload regions, the general trends of output bit-rate of RT & NRT flows, and understanding of the system in a simple and intuitive manner.

Paper Organization: This paper is organized into seven main parts: Introduction, background, and motivation are explained in Section I. The system model and definitions, for the joint RT and NRT flows packet scheduling and RB allocation, will be given in Section II. The formulation of the novel proposed joint RT and NRT flows packet scheduling and RB allocation will be described in Section III. The special cases of the proposed framework will be discussed in Section IV. The proposed novel algorithm will be explained in Section V. The high-level input–output system behavior study of the system will be discussed in Section VI, as a benchmark and as an explanation for the simulations (in Section VII). Section VII provides extensive simulation results, through two experiments, to evaluate our proposed algorithm and to compare it with the prior art.

List of Symbols: We summarize the symbols used throughout this paper, with their short definition, in Table I.

TABLE I
(Continued.) LIST OF SYMBOLS

Symbol	Definition
κ	Channel awareness factor of RT flows.
$\mathcal{I}_{\phi}^{\mathrm{r}}(\bar{r}_{\phi}[k])$	Bit-rate importance function (for flow ϕ).
$\mathcal{I}_{\phi}^{\mathrm{d^{HOL}}}(d_{\phi}^{\mathrm{HOL}}[k])$	HOL-delay importance function (for flow ϕ).
ϑ	Scale factor.
ξ	NRT disutility gradient maximum.
$\mathcal{I}_{\phi}^{\mathrm{q}}(\bar{q}_{\phi}[k])$	Mean queue length importance function (for flow ϕ).
$\mathcal{I}_{\phi}^{\mathrm{r}}(\bar{r}_{\phi}[k])$	Mean bit-rate importance function (for flow ϕ)
π	Interpretable design parameter for parameter ξ.
α	GPF parameter.
μ_{ϕ}	EXP and MLWDF parameter (for flow ϕ).
δ_{ϕ}	Probability of exceeding the HOL-delay in ML-WDF (for flow ϕ).
τ_{ϕ}	Maximum delay threshold in MLWDF (for flow ϕ).
$\bar{d}^{\mathrm{HOL}}[k]$	Average of HOL-delay over flows (in frame k).
$\lvert\Phi_{\mathrm{RT}}\rvert, \lvert\Phi_{\mathrm{NRT}}\rvert$	Number of RT flows, number of NRT flows.
$j^*[k], \phi^*[k]$	Algorithm internal variables for selected sub-channel and for selected flow (in frame k).
η	EXP parameter controlling the delay fairness.
$\bar{d}_{\phi}[k]$	Mean-delay (until frame k).
$d_{\phi}^{\max}$	Maximum mean delay (for flow ϕ).
$\Phi_{\mathrm{BE}}, \Phi_{\mathrm{VoIP}}$	Set of BE flows, set of VoIP flows.
$\nu^{\mathrm{d}}, \nu^{\mathrm{d^{HOL}}}$	Mean-delay fairness parameter, HOL-delay fairness parameter.
$\iota_{\mathrm{RT}}, \iota_{\mathrm{NRT}}$	RT flows transmission interval, NRT flows transmission interval (in TUF).
l_{ϕ}	Length of interval for RT flows transmissions (in TUF, for flow ϕ).
$T^{(j)}$	Algorithm internal variable (for available slots per sub-channel j).
$\tilde{b}_{\phi}^{(j)}$	Algorithm internal variable for AMC values (on sub-channel j for flow ϕ).
Ω_{RT}	Capacity of system when output consists of RT-only flows.
$\Gamma_{\phi\in\Phi_{\mathrm{RT}}}^{(j)}[k]$	The selecting criterion for RT flows (flow ϕ, sub-channel j, frame k)
$\Gamma_{\phi\in\Phi_{\mathrm{NRT}}}^{(j)}[k]$	The selecting criterion for NRT flows (flow ϕ, sub-channel j, frame k).
Ω_{NRT}	Capacity of system when output consists of NRT-only flows.
Ω_{SLL}	Capacity of system when saturated with the lowest load.
$\Lambda_{\Sigma}^{\mathrm{in}}, \Lambda_{\mathrm{RT}}^{\mathrm{in}}, \Lambda_{\mathrm{NRT}}^{\mathrm{in}}$	Total input bit-rate, total RT flows input bit-rate, total NRT flows input bit-rate.
f_{RT}	Fraction of RT input.
$\Lambda_{\Sigma}^{\mathrm{out}}, \Lambda_{\mathrm{RT}}^{\mathrm{out}}, \Lambda_{\mathrm{NRT}}^{\mathrm{out}}$	Total output bit-rate, total RT flows output bit-rate, total NRT flows output bit-rate.
ρ	Normalized load coefficient.
ρ_{RTO}	Lowest normalized load coefficient which makes the output consists of RT-only.
Ω_{ρ}	Capacity of system for the normalized load coefficient ρ.
w_{ϕ}	Fraction of flow ϕ input.
$\mathfrak{U}^{\mathrm{non-frag}}, \mathfrak{U}^{\mathrm{frag}}$	Non-fragmented utilization, fragmented utilization.
$N_{\mathrm{used}}, N_{\mathrm{total}}$	Total number of used RBs, total number of RBs.
$B_n^{\mathrm{cap}}, B_n^{\mathrm{tx}}$	Capacity of n^{th} RB, number of transmitted bits on n^{th} RB.
$J(\cdot)$	Jain's index.

II. SYSTEM MODEL

A downlink scenario with an OFDMA air interface, which serves RT flows in set Φ_{RT} and NRT flows in set Φ_{NRT}, in a single cell, is considered here. The union of the flows is denoted by Φ, $\Phi = \Phi_{\mathrm{RT}} \cup \Phi_{\mathrm{NRT}}$.

A. OFDMA Frame

The total bandwidth is divided into N subchannels consisting of several OFDMA subcarriers. Each subchannel is further divided in time into T time slots. This way, the time–frequency plane, for each frame, is divided into NT RBs, each of which spans T_b seconds in time and W_b Hertzs in frequency. It is worth mentioning that we extend the OFDMA plane framework to have more than one single resource block (RB) on a specific subchannel, within a frame, over time. This generalization gives

the flexibility of including future technologies, where time division within a frame is possible. When this flexibility is not possible, $T = 1$ reduces the model to the conventional OFDMA plane of LTE. Time division within a frame, if possible, results in higher granularity and increases the efficiency in resource allocation.

The transmission frames are indexed by notation k, sequentially. We use the frame to refer to frame index throughout this paper. In frame k, the highest available spectral efficiency and corresponding adaptive modulation and coding (AMC) level, for a single RB on subchannel j for flow ϕ is

$$b_\phi^{(j)}[k] = \mathfrak{f}\left(\mathsf{SINR}_\phi^{(j)}[k]\right), \tag{1}$$

in bits per second per Hertz, where $\mathsf{SINR}_\phi^{(j)}[k]$ is the signal-to-the-interference-and-noise ratio (SINR) of RBs associated with flow ϕ on subchannel j in frame k, and $\mathfrak{f}(\cdot)$ represents the AMC table which depends on bit error rate, as well. In Section VII, we will use arrays of modulation levels, coding rates, and SINR thresholds, which define a specific $\mathfrak{f}(\cdot)$.

B. Frame Bit-Rate

Radio resources are assigned to the flows in terms of RBs; each RB carries data of only one flow at a time. The bit-rate of a flow is determined from the number of RBs it is allocated in the frame and the AMC level used in each RB. The bit-rate of the flow ϕ, in frame k, is

$$r_\phi[k] = W_b \sum_{j=1}^{N} b_\phi^{(j)}[k] x_\phi^{(j)}[k], \tag{2}$$

in bits per second, where $b_\phi^{(j)}[k]$ is the spectral efficiency of RBs on subchannel j for flow ϕ in frame k, W_b is the frequency span of RB as defined earlier, and $x_\phi^{(j)}[k]$ is the number of RBs allocated to flow ϕ on subchannel j in frame k.

III. Joint Realtime and Nonrealtime Flows Packet Scheduling and Resource Block Allocation Formulation

Here, we formulate the main joint RT and NRT flows packet scheduling and RB allocation based on disutility functions, after a preliminary discussion on delay driven disutility functions.

Delay Driven Disutility Functions: Bit-rate driven utility functions have been proposed first in [5], which inspired many other studies (see, for example, [6], [10], [26], [31], [32], and [54]). However, there exist other QoS measures, such as delay, which are independent of bit-rate. It is for the same underlying reason that, to meet the packet delay deadlines of RT flows, it is not sufficient to only guarantee a minimum mean bit-rate to those flows [22]. Therefore, recently, delay has been taken into account as an input to the disutilities [43], [53], [55]. In this paper, we use disutility functions with respect to the HOL delay, queue-length information, and bit-rate information. RT flows have sensitivities based on HOL delay, whereas NRT flows sense the mean delay mainly. The concept of using HOL delay, in addition to other information inside disutility functions, creates the opportunity to use the framework for QoS classes that may emerge in the future. This concept will be further elaborated in Section IV-H.

A. Formulation

The network objective is to minimize the total disutility $D_{\text{net}}^{\text{joint}}(\mathbf{d}^{\text{HOL}}[k], \bar{\mathbf{q}}[k], \bar{\mathbf{r}}[k])$, which depends on HOL delay vector $\mathbf{d}^{\text{HOL}}[k]$, mean queue-length vector $\bar{\mathbf{q}}[k]$, and mean bit-rate vector $\bar{\mathbf{r}}[k]$. The corresponding optimization problem can be casted as

$$\min_{x_\phi^{(j)}[k] \in \mathcal{C}^{\text{PHY}}} \sum_{\phi=1}^{|\Phi|} D_\phi^{\text{joint}}\left(d_\phi^{\text{HOL}}[k], \bar{q}_\phi[k], \bar{r}_\phi[k]\right), \tag{3}$$

where $D_\phi^{\text{joint}}(d_\phi^{\text{HOL}}[k], \bar{q}_\phi[k], \bar{r}_\phi[k])$ describes the combined disutility with respect to the HOL delay (denoted by $d_\phi^{\text{HOL}}[k]$), mean queue-length (denoted by $\bar{q}_\phi[k]$), and mean bit-rate (denoted by $\bar{r}_\phi[k]$). The HOL delay, for flow ϕ, is defined as the delay experienced by the packet at the HOL of the associated queue. Formally, this is the difference of the current frame index k and the arrival time-stamp frame index of HOL packet $t_\phi^{\text{HOL}}[k]$, i.e.,

$$d_\phi^{\text{HOL}}[k] = k - t_\phi^{\text{HOL}}[k]. \tag{4}$$

Mean queue-length $\bar{q}_\phi[k]$ is defined, based on frame queue-length $q_\phi[k]$, recursively, as

$$\bar{q}_\phi[k] = \left(1 - \frac{1}{k}\right)\bar{q}_\phi[k-1] + \frac{1}{k} q_\phi[k]. \tag{5}$$

Likewise, mean bit-rate $\bar{r}_\phi[k]$ is defined, based on frame bit-rate $r_\phi[k]$, recursively, as

$$\bar{r}_\phi[k] = \left(1 - \frac{1}{k}\right)\bar{r}_\phi[k-1] + \frac{1}{k} r_\phi[k]. \tag{6}$$

As defined earlier, the number of frequency subchannels is denoted by N, and the number of time slots per frequency subchannel is denoted by T. The total number of flows is denoted by $|\Phi|$. The optimization constraints are induced by the physical (PHY) layer limitation of RBs in a frame and the fact that scheduling and allocation does not map an RB to more than one flow. Accordingly, the feasible set for the optimization variable is

$$\mathcal{C}^{\text{PHY}} = \left\{ x_\phi^{(j)}[k] \,\middle|\, \forall j : \sum_{\phi=1}^{|\Phi|} x_\phi^{(j)}[k] \le T, \quad \forall \phi, j : x_\phi^{(j)}[k] \in \{0, \ldots, T\} \right\}. \tag{7}$$

Note that the optimization variable $x_\phi^{(j)}[k]$ indicates how many slots in subchannel j are assigned to flow ϕ, in frame k, which is an integer number in $[0, T]$. In addition, since each frequency subchannel has T time slots, the total assignment to any frequency subchannel should be less than T. The channel information is embedded in the optimization. The dependence of the optimization to the channel information will show itself when we use the gradient of the network disutility function to develop the algorithm in Section III-B. The disutility functions are nondecreasing in their delay argument, nondecreasing in

their queue-length argument, and are nonincreasing in their bit-rate argument. To the best of our knowledge, the formulation (3) is novel in the sense that it incorporates the HOL delay, mean queue-length, and mean bit-rate information.

In the following, we demonstrate how the framework is the generalization of the sequential approach (static separation), present the ways of choosing disutility functions for RT and NRT flows in the proposed joint approach, and show the perspective with respect to the relevant literature. Later, in Section IV, we will list candidates of packet scheduling and RB allocation for RT and NRT from the literature, their main properties, and the underlying reason of their properties in their structure. Since we use gradient-based algorithm, we directly design the gradient of the disutility functions in Section III-B.

Example of the Sequential Approach: Conventionally, packet scheduling or RB allocation of RT flows ($\phi \in \Phi_{\mathrm{RT}}$) and NRT flows ($\phi \in \Phi_{\mathrm{NRT}}$) is executed based on two *sequential* algorithms, where Φ_{RT} and Φ_{NRT} are the set of RT flows and the set of NRT flows, respectively. The sequential approaches result in complete separation of RBs into two sets for RT and NRT flows. In other words, RBs are assigned to RT flows based on an RT packet scheduling and RB allocation algorithm, and if any RBs remain, the NRT flows are served. Traditionally, EDF is used for RT flows based on their HOL delay, as well as their delay deadlines. EDF works based on the HOL delay margin (denoted by $\Delta d_\phi^{\mathrm{HOL}}[k]$), which is the difference of the flow's current HOL delay and its maximum threshold, i.e.,

$$\Delta d_\phi^{\mathrm{HOL}}[k] \triangleq d_\phi^{\mathrm{HOL}}[k] - d_\phi^{\mathrm{HOL}^{\max}}, \tag{8}$$

where $d_\phi^{\mathrm{HOL}^{\max}}$ is the HOL delay deadline of flow ϕ. A flow ϕ is in the safe region when $\Delta d_\phi^{\mathrm{HOL}}[k] < 0$. Accordingly, the gradient of disutilities of RT flows, in EDF, can be interpreted as

$$\frac{\partial D_\phi^{\mathrm{joint}}\left(d_\phi^{\mathrm{HOL}}[k], \bar{r}_\phi[k], \bar{q}_\phi[k]\right)}{\partial x_\phi^{(j)}[k]} = \Delta d_\phi^{\mathrm{HOL}}[k], \tag{9}$$

for $\phi \in \Phi_{\mathrm{RT}}$. It is worth mentioning that EDF packet scheduling is channel blind. This can be noticed from the fact that the left-hand side of (9) is dependent on both subchannel index j and flow index ϕ, whereas the right-hand side is not dependent on j. In other words, a flow is scheduled based on EDF, irrespective of its subchannel condition, even if their subchannels are in deep fade. A channel aware version of EDF can be defined by its disutility gradient as

$$\frac{\partial D_\phi^{\mathrm{joint}}\left(d_\phi^{\mathrm{HOL}}[k], \bar{r}_\phi[k], \bar{q}_\phi[k]\right)}{\partial x_\phi^{(j)}[k]} = b_\phi^{(j)}[k]\, \Delta d_\phi^{\mathrm{HOL}}[k], \tag{10}$$

for $\phi \in \Phi_{\mathrm{RT}}$. The gradient of the disutility, for any frame index k, forms a two-dimensional array, where its maximum value plays an important role in decision making.

Having finished RT flows packet scheduling and RB allocation, if any RBs are still available, NRT flows are served based on PF, which is determined as

$$\frac{\partial D_\phi^{\mathrm{joint}}\left(d_\phi^{\mathrm{HOL}}[k], \bar{r}_\phi[k], \bar{q}_\phi[k]\right)}{\partial x_\phi^{(j)}[k]} = \frac{b_\phi^{(j)}[k]}{\bar{r}_\phi[k]}, \tag{11}$$

for $\phi \in \Phi_{\mathrm{NRT}}$, where $\partial D_\phi^{\mathrm{joint}}\left(\bar{r}_\phi[k], \bar{q}_\phi[k], d_\phi^{\mathrm{HOL}}[k]\right)/\partial x_\phi^{(j)}[k]$ is the gradient of the corresponding disutility function, in this special case. Note that we use the notation $D_\phi^{\mathrm{joint}}(\cdot)$ as the disutility, for both the sequential approach and for the joint approach. The design of the disutility, or its gradient, with respect to the QoS measurements and QoS requirements, distinguishes the joint approach from the sequential approach.

B. Proposed Joint Approach

As discussed earlier, the complete separation of RB sets for RT and NRT flows results in system inefficiency. RT flows can presumably have sufficient time with respect to their delay deadlines while their channels are in deep fade. In this situation, NRT flows with good channels can be transmitted. We propose a generalized framework for disutility functions of RT and NRT flows, enabling RB allocation from a common pool of RBs. This framework not only enables the joint RT and NRT flows packet scheduling and RB allocation, but also paves the way for future unified designs in higher number of flow types.

Here, we propose the general structure of the gradient of the disutility for RT and NRT flows. Different sets of information are relevant to RT and NRT packet scheduling and RB allocation. While HOL delay is the most relevant information in decision making for RT flows, long-term information such as mean bit-rate is relevant for NRT flows.

Design of the Gradient of the Disutility for the RT Flows: For RT flows, we start defining the gradient of disutility function as

$$\frac{\partial D_\phi^{\mathrm{joint}}\left(d_\phi^{\mathrm{HOL}}[k], \bar{r}_\phi[k], \bar{q}_\phi[k]\right)}{\partial x_\phi^{(j)}[k]} = \frac{\left(b_\phi^{(j)}[k]\right)^{\kappa}}{F_\phi^{\bar{\mathrm{r}}}\left(\bar{r}_\phi[k]\right)} F_\phi^{\mathrm{d^{HOL}}}\left(d_\phi^{\mathrm{HOL}}[k]\right), \qquad \text{if } \phi \in \Phi_{\mathrm{RT}}, \tag{12}$$

where κ is the channel awareness exponent of RT flows, and $F_\phi^{\bar{\mathrm{r}}}(\bar{r}_\phi[k])$ & $F_\phi^{\mathrm{d^{HOL}}}(d_\phi^{\mathrm{HOL}}[k])$ are nondecreasing functions, $\mathbb{R} \longrightarrow \mathbb{R}^+$, which represents the component of the gradient of the disutility with respect to the mean bit-rate and HOL delay, respectively. Since the channel and queuing processes are asynchronous, a flow at its peak channel state may not have packets to transmit, whereas another flow may have serious queuing delay. Therefore, for RT flows, the algorithm has to be able to trade off between channel efficiency and delay. This is enabled through the parameter κ. The channel awareness exponent of RT flows κ can be used to trade off the RT output bit-rates with fulfilling HOL delay requirement. Parameter κ can be also adjusted for preferring the RT flows in cell edge, instead of the NRT flows with high channel quality. We use a simple choice of $\kappa = 1$, in the simulation section. The image of function $F_\phi^{\mathrm{d^{HOL}}}(d_\phi^{\mathrm{HOL}}[k])$ should be positive ($\mathbb{R}^+$) because it is multiplied with channel awareness factor. Function $F_\phi^{\bar{\mathrm{r}}}(\bar{r}_\phi[k])$ enforces the bit-rate fairness of RT flows. Note that, when $\Delta d_\phi^{\mathrm{HOL}}[k]$ becomes positive,

the deadline has been passed, and the corresponding RT flows' packets will be discarded. Therefore

$$\Delta d_{\phi}^{\mathrm{HOL}}[k] \in \left[-d_{\phi}^{\mathrm{HOL}^{\max}}, 0\right], \tag{13}$$

when the packet deletion, due to passed deadline, is on. Moreover, note that the HOL delay in RT gradient should be amplified so that it can be compared with NRT gradient. Particularly, we use

$$F_{\phi}^{\mathrm{d^{HOL}}}\left(d_{\phi}^{\mathrm{HOL}}[k]\right) = \vartheta\, \mathrm{e}^{\Delta d_{\phi}^{\mathrm{HOL}}[k]}, \tag{14}$$

where ϑ is a scale factor. With choice of (14), RT gradient is in the interval of

$$F_{\phi}^{\mathrm{d^{HOL}}}\left(d_{\phi}^{\mathrm{HOL}}[k]\right) \in \left[\vartheta\, \mathrm{e}^{-d_{\phi}^{\mathrm{HOL}^{\max}}}, \vartheta\right]. \tag{15}$$

Design of the Gradient of the Disutility for the NRT Flows: For NRT flows, long-term information, namely, mean queue-length and mean bit-rate, is relevant. For NRT flows, we form the general structure of the gradient of the disutility function as

$$\frac{\partial D_{\phi}^{\mathrm{joint}}\left(d_{\phi}^{\mathrm{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\right)}{\partial x_{\phi}^{(j)}[k]} = \min\left[\xi, b_{\phi}^{(j)}[k] \frac{F_{\phi}^{\bar{\mathrm{q}}}\left(\bar{q}_{\phi}[k]\right)}{F_{\phi}^{\bar{\mathrm{r}}}\left(\bar{r}_{\phi}[k]\right)}\right], \qquad \text{if } \phi \in \Phi_{\mathrm{NRT}}, \tag{16}$$

where ξ is the instrument for sliding between the complete RT and NRT flows RB sets separation and the common pool of RBs, $\bar{q}_{\phi}[k]$ is the mean queue-length of flow ϕ until frame k, $\bar{r}_{\phi}[k]$ is the mean bit-rate of flow ϕ until frame k, $F_{\phi}^{\bar{\mathrm{q}}}(\bar{q}_{\phi}[k])$ is the mean queue-length importance function, and $F_{\phi}^{\bar{\mathrm{r}}}(\bar{r}_{\phi}[k])$ is the mean bit-rate importance function. It is worth mentioning that the framework is channel aware to have spectrally efficient transmissions. Based on this fact, the cell edge effect is compensated in the proposed framework for their poor links.

The general structures in (16) and (12) are inspired by keeping the desired pattern in the legacy packet scheduling and RB allocation designs and generalizing the structure to exploit existent degrees of freedom. The proposed approach is evolved further in [55] to incorporate the operators interest as well right into the packet scheduling and RB allocation frameworks.

Design of the Parameter ξ: One of the crucial facts in designing the joint RT and NRT flows packet scheduling and RB allocation is that the NRT flows' disutility gradient has to be bounded. Otherwise, when the RT packets pass their deadlines and are discarded, RT flows cannot compete with NRT flows, where the packets are kept in the queues for much longer time. Therefore, the structure in (16) essentially should be bounded. This is done through clipping by ξ, as in (16). On one hand, because of the discarding of RT packets after their delay deadline, the RT flows' disutility gradient value will decrease. On the other hand, NRT packets have been kept in their queues, contribute highly to their queue-length attributes, and result in increase in their gradient. We design parameter ξ based on the fact that when the RT packets reach π fraction $(0 < \pi < 1)$ of their deadline (equivalently pass $1-\pi$ fraction of their deadline), the RT disutility gradient, divided by its channel awareness term, should be strictly larger than their NRT counterparts as

$$\max_{\phi \in \Phi_{\mathrm{NRT}}} \frac{\partial D_{\phi}^{\mathrm{joint}}\left(d_{\phi}^{\mathrm{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\right)}{\partial x_{\phi}^{(j)}[k]} < \min_{\phi \in \Phi_{\mathrm{RT}}} \left.\frac{\partial D_{\phi}^{\mathrm{joint}}\left(d_{\phi}^{\mathrm{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\right)}{\partial x_{\phi}^{(j)}[k]}\right|_{k=\pi d_{\phi}^{\mathrm{HOL}^{\max}}}. \tag{17}$$

This ensures that, in this fraction, RT flows will be exclusively transmitted. Accordingly, and based on (14), the following rule is derived for designing ξ:

$$\xi = \vartheta\, \mathrm{e}^{-\pi \max_{\phi}\left(d_{\phi}^{\mathrm{HOL}^{\max}}\right)}. \tag{18}$$

Increasing π makes the algorithm prioritize RT flows over NRT flows with the cost of lower multiuser diversity. In other words, we divide adaptively the delay margin of RT flows into two regions: region 1, where NRT flows can compete with RT flows, based on their utility, and region 2, where RT flows are given strict higher priority. Nevertheless, note that these two regions do not necessarily separate the RT and NRT flows in time and any combination of RT and NRT flows transmissions can take place, unless when RT flows load does not allow. This adaptive separation enables the joint approach and exploits the hidden multiuser diversity in the conventional designs. Note that parameter π is a design parameter for ξ. The relationship between parameters π and ξ is one-to-one. However, π is interpretable in terms of when RT flows will be given strict priority.

IV. Special Cases

A. Sequential EDF and GPF

The framework falls back into complete separation of RT and NRT, served by EDF and GPF [8]–[10], [37], sequentially with the following special choices of

$$\frac{\partial D_{\phi}^{\mathrm{joint}}\left(d_{\phi}^{\mathrm{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\right)}{\partial x_{\phi}^{(j)}[k]} = \Delta d_{\phi}^{\mathrm{HOL}}[k], \qquad \text{if } \phi \in \Phi_{\mathrm{RT}} \tag{19}$$

and

$$\frac{\partial D_{\phi}^{\mathrm{joint}}\left(d_{\phi}^{\mathrm{HOL}}[k], \bar{r}_{\phi}[k], \bar{q}_{\phi}[k]\right)}{\partial x_{\phi}^{(j)}[k]} = \min\left[\xi, \frac{b_{\phi}^{(j)}[k]}{\left(\bar{r}_{\phi}[k]\right)^{\alpha}}\right], \qquad \text{if } \phi \in \Phi_{\mathrm{NRT}}, \tag{20}$$

where $0 \leq \alpha$ is the parameter of GPF that influences the bit-rate fairness. Different types of bit-rate fairness can be achieved by changing α: Cases $\alpha \to 0$, $\alpha = 1$, $\alpha \to \infty$ correspond to sum bit-rate maximization (or max-SINR; first introduced in [56]), PF, and maxmin bit-rate fairness, respectively [8], [57]. The parameter α slides the allocation from no bit-rate fairness to the highest bit-rate fairness (maxmin bit-rate fairness), where it maximizes the minimum mean bit-rate. In other words, increasing α increases the lower percentiles (such as the fifth percentile) bit-rate at the cost of decrease in higher percentile

(such as the 95th percentile) of bit-rate [9], [58]. Parameter α in GPF can be used in a closed-loop control system to achieve a target level of fairness [59].

Now, for achieving the sequential execution of the algorithms for RT and NRT flows, the NRT disutility gradient should be clipped by the minimum value of the disutility in RT flows set. This results in selecting ξ as

$$\xi = \min_{\phi\in\Phi_{\rm RT},\ 1\le j\le N} \left[\frac{\partial D_\phi^{\rm joint}\left(d_\phi^{\rm HOL}[k], \bar{r}_\phi[k], \bar{q}_\phi[k]\right)}{\partial x_\phi^{(j)}[k]} \right] \quad (21)$$

which makes the RT and NRT flows disutility gradient disjoint, in their value.

Similar to Section IV-A, by substituting EDF with EXP, or MLWDF, the sequential versions of EXP-GPF, or MLWDF-GPF are constructed. We describe MLWDF-GPF and EXP-GPF in the following two subsections.

B. Sequential MLWDF and GPF

A popular approach for RT flows packet scheduling and RB allocation is the MLWDF [47], [60]. MLWDF can be considered as an advance algorithm in comparison with EDF. In each frame, a flow ϕ^* is selected repeatedly to be transmitted on an RB on subchannel j^* according to

$$(\phi^*, j^*) = \arg \max_{\substack{\phi\in\Phi_{\rm RT},\\ 1\le j\le N}} \mu_\phi \frac{b_\phi^{(j)}[k]}{\bar{r}_\phi[k]} d_\phi^{\rm HOL}[k], \quad \phi\in\Phi_{\rm RT}, \quad (22)$$

where μ_ϕ is suggested to be selected as

$$\mu_\phi = -\frac{\log\delta_\phi}{\tau_\phi}, \quad (23)$$

by large deviation optimality results of [61], $\bar{r}_\phi[k]$ is the mean bit-rate in frame k, τ_ϕ is the maximum allowable delay threshold, and δ_ϕ is a maximum probability of exceeding the delay threshold [11] as

$$\Pr\left(d_\phi^{\rm HOL}[k] > \tau_\phi\right) < \delta_\phi. \quad (24)$$

It is worth mentioning that, by fixing a close-to-one percentile for $1-\delta_\phi$ (or equivalently, small δ_ϕ), we can interpret τ_ϕ as the delay deadline, which we denoted earlier by $d_\phi^{\rm HOL^{max}}$. As an example, for $1-\delta_\phi$ equal to the 99th percentile, we have

$$\mu_\phi \approx \frac{2}{d_\phi^{\rm HOL^{max}}}. \quad (25)$$

By this interpretation, it can be noticed that the HOL delays are divided by their deadlines in the structure of MLWDF [see (22) and (25)]. It is in contrast to EDF, where the difference of HOL delays and their deadlines form the structure.

We note that, with selecting the disutility gradient component as

$$\frac{F_\phi^{\rm HOL}\left(d_\phi^{\rm HOL}[k]\right)}{F_\phi^{\bar{\rm r}}\left(\bar{r}_\phi[k]\right)} = \frac{\mu_\phi}{\bar{r}_\phi[k]} d_\phi^{\rm HOL}[k], \quad (26)$$

our framework for RT flows falls back to MLWDF. For a detailed analysis of MLWDF, see [62].

Similar to the previous section on EDF and GPF, the sequential MLWDF and GPF is constructed by serving the NRT flows by GPF, after RT flows are served by MLWDF.

C. Sequential EXP and GPF

Another mechanism, which is called EXP [45], has been also proposed for RT flows packet scheduling and RB allocation with a relatively similar structure in MLWDF. An OFDMA version of EXP can be represented as

$$(\phi^*, j^*) = \arg \max_{\substack{\phi\in\Phi_{\rm RT},\\ 1\le j\le N}} \frac{b_\phi^{(j)}[k]}{\bar{r}_\phi[k]} \, {\rm e}^{\left(\frac{\mu_\phi d_\phi^{\rm HOL}[k]}{1+\left(\overline{d^{\rm HOL}[k]}\right)^\eta}\right)}, \quad (27)$$

where, in each frame k, a flow ϕ^* is selected to be transmitted on an RB on subchannel j^*, $0<\eta<1$, μ_ϕ is defined the same as in (25) (inversely proportional to delay deadline), and

$$\overline{d^{\rm HOL}}[k] = \frac{1}{|\Phi_{\rm RT}|} \sum_{\phi=1}^{|\Phi_{\rm RT}|} d_\phi^{\rm HOL}[k] \quad (28)$$

is the average of HOL delays over RT flows.

The HOL delay term $d_\phi^{\rm HOL}[k]$, in (22) and (27) can be replaced with a queue-length term as $q_\phi[k]$ to obtain the queue-length driven versions of the aforementioned algorithms. Similar to EDF, both MLWDF and EXP are heuristics designed for delay-sensitive flows. EXP and MLWDF in conjunction with virtual token queues (with constant deterministic arrival rate) can be used to guarantee a minimum bit-rate [27], [41]. When queue-lengths of flows are equal or close (see [22] for its formal definition), EXP and MLWDF reduce to PF [22]. EXP is suitable for the cases where the delay equalization is preferable. However, there has been analysis [63] showing that EXP sacrifices the asymptotic system throughput when the queues grow asymptotically as the cost of emphasis on delay equalization. As discussed earlier, the structure of EXP and MLWDF is based on the division of HOL delay by its deadline, whereas the structure of EDF is based on the difference of HOL delay and its deadline.

The sequential EXP and GPF is constructed by serving the NRT flows by GPF, after RT flows are served by EXP [14].

D. Mean Delay-Based MDU

An algorithm, which is called MDU, based on two different functions on the Little's delay for RT and NRT flows has been used in [42] and [64] for resource allocation of RT and NRT flows. As an example of RT, [42] and [64] used

$$\frac{\partial D_\phi^{\bar{\rm d}}\left(\bar{d}_\phi[k]\right)}{\partial x_\phi^{(j)}[k]} = b_\phi^{(j)}[k]$$

$$* \begin{cases} \bar{d}_\phi[k], & \text{if } \bar{d}_\phi[k] \le \frac{\bar{d}_\phi^{\max}}{4},\ \phi\in\Phi_{\rm VoIP}, \\ \left(\bar{d}_\phi[k]\right)^{1.5} - \left(\frac{\bar{d}_\phi^{\max}}{4}\right)^{1.5} + \frac{\bar{d}_\phi^{\max}}{4}, & \text{if } \frac{\bar{d}_\phi^{\max}}{4} \le \bar{d}_\phi[k],\ \phi\in\Phi_{\rm VoIP}, \end{cases} \quad (29)$$

for the VoIP disutility gradient, where $\bar{d}_\phi[k]$ is the Little's delay for flow ϕ in frame k, and $\bar{d}_\phi^{\max}$ is the maximum tolerable Little's delay for flow ϕ. Note that we generalized their utility for general $\bar{d}_\phi^{\max}$ as the mean delay threshold.

In parallel, as an example of NRT, [44] and [62] used

$$\frac{\partial D_\phi^{\bar{\mathrm{d}}}\left(\bar{d}_\phi[k]\right)}{\partial x_\phi^{(j)}[k]} = b_\phi^{(j)}[k].\begin{cases}\left(\bar{d}_\phi[k]\right)^{0.5}, & \text{if } \bar{d}_\phi[k] \le 100,\ \phi \in \Phi_{\mathrm{BE}},\\ 100^{0.5}, & \text{if } 100 \le \bar{d}_\phi[k],\ \phi \in \Phi_{\mathrm{BE}},\end{cases} \tag{30}$$

for the gradient of the disutility of the best effort (BE) traffic. Notations Φ_{VoIP} and Φ_{BE} denote the VoIP and BE flows sets, respectively. Accordingly, the flow $\phi^*[k]$ is selected to be transmitted on an RB in subchannel $j^*[k]$ based on

$$(\phi^*[k], j^*[k]) = \arg \max_{\substack{1 \le \phi \le |\Phi|,\\ 1 \le j \le N}} \frac{\partial D_\phi^{\bar{\mathrm{d}}}\left(\bar{d}_\phi[k]\right)}{\partial x_\phi^{(j)}[k]}. \tag{31}$$

Note that the Little's delay ($\bar{d}_\phi[k]$) can, in fact, be approximated [42], [53] by

$$\bar{d}_\phi[k] \approx \frac{\bar{q}_\phi[k]}{\bar{r}_\phi[k]}. \tag{32}$$

Therefore, MDU can be considered as a special case of the proposed approach. Nevertheless, since the averaging is a low-pass filter and the Little's delay is essentially an averaging mechanism, RT flows suffer when the Little's delay is only used, as an argument, for their disutility functions. In reality, RT flows sense the HOL delay, rather than the Little's delay or mean delay. For handling heterogeneous HOL delay deadlines, the design needs to incorporate the HOL delays, which are the relevant delay measures for RT flows. We also note that, since MDU design is merely based on mean delay, it suffers from bit-rate fairness point of view, in high input loads.

E. Delay Fairness Through WGPF

Similar to bit-rate fairness, delay fairness is a concept referring to mechanisms that equalize delay measures, among flows. A framework for achieving mean delay fairness has been studied in [53]. We note that, with the selection of

$$F_\phi^{\bar{\mathrm{r}}}(\bar{r}_\phi[k]) = (\bar{r}_\phi[k])^{\nu^{\bar{\mathrm{d}}}}, F_\phi^{\bar{\mathrm{q}}}(\bar{q}_\phi[k]) = (\bar{q}_\phi[k])^{\nu^{\bar{\mathrm{d}}}}, \quad \text{if } \phi \in \Phi_{\mathrm{NRT}} \tag{33}$$

and $\xi = \infty$, the proposed gradient in (16) reduces to the Little's delay driven disutility function in [53]. Indeed, parameter $1 \le \nu^{\bar{\mathrm{d}}}$ is controlling the tradeoff between mean delay fairness and throughput (or resource efficiency). It has been proved in [53] that sufficiently large $\nu^{\bar{\mathrm{d}}}$ corresponds to the *minmax* mean delay fairness.

On the other hand, with the selection of disutility over HOL delay as

$$\frac{F_\phi^{\mathrm{HOL}}\left(d_\phi^{\mathrm{HOL}}[k]\right)}{F_\phi^{\bar{\mathrm{r}}}\left(\bar{r}_\phi[k]\right)} = \frac{\left(d_\phi^{\mathrm{HOL}}[k]\right)^{\nu^{\mathrm{d^{HOL}}}}}{\bar{r}_\phi[k]}, \quad \text{if } \phi \in \Phi_{\mathrm{RT}}, \tag{34}$$

our framework will be reduced to the HOL delay fairness proposed in [65], similar to the mean delay fairness in [53]. The same observation in [53] for tradeoff between delay fairness and throughput (or equivalently resource efficiency) can be seen for HOL delay fairness, by controlling the parameter $\nu^{\mathrm{d^{HOL}}}$ ($1 \le \nu^{\mathrm{d^{HOL}}}$) [59]. Reference [65] used the HOL delay fairness and the GPF mechanism sequentially (or with static separation) for RT and NRT flows packet scheduling and RB allocation.

The proposed disutility, in (16), with using GPF functions for the NRT flows can be adjusted between the sum bit-rate maximization and the maxmin bit-rate fairness for NRT flows. As the counterpart to NRT flows, the proposed disutility, in (12), with using weighted GPF (WGPF) functions for the RT flows can be adjusted between the sum delay minimization and the minmax delay fairness for RT flows (see [19] for the proofs).

F. Time-Utility Function

An idea based on TUF for joint RT and NRT flows packet scheduling and RB allocation has been proposed in [48]–[50]. In general, RT scheduler should transmit RT packets any time within their deadline for satisfying the delay requirement, not necessarily with EDF. References [48] and [50] used the z-shape TUF adopted from [66] as the urgency criterion for RT resource allocation and a channel efficiency criterion for NRT resource allocation. Based on TUF, RT flows are transmitted, just before their deadline, within a short interval ι_{RT} defined as

$$\iota_{\mathrm{RT}} = \left[d_\phi^{\mathrm{HOL}^{\max}} - l_\phi, d_\phi^{\mathrm{HOL}^{\max}}\right], \tag{35}$$

where $d_\phi^{\mathrm{HOL}^{\max}}$ is the flow ϕ deadline, and l_ϕ determines the length of interval for RT flows transmissions. NRT packets are transmitted during the remaining time interval ι_{NRT} defined as

$$\iota_{\mathrm{NRT}} = \left[0, d_\phi^{\mathrm{HOL}^{\max}} - l_\phi\right). \tag{36}$$

The TUF approach transmits RT packets near the deadline without channel information. Therefore, it cannot fully exploit the multiuser diversity, and it increases the number of dropped packets due to channel-blindness attribute in TUF for RT flows. In [67], a similar approach in [48] is also used where [67] introduces a transmission guard interval, which gives high priority (overriding the NRT packets) to RT packets approaching the delay deadline. Note that, in our approach, generally, any combination of RT and NRT packets transmissions can take place in time, and RT and NRT packets are not necessarily separated in time, in contrast to the TUF approach.

G. Others

References [28] and [29] considered the unit-step utility functions for non-BE flows and concave nondecreasing utility functions for BE flows and proved three theorems for bounds on the optimality of their proposed algorithms, based on the inverse of the utility functions. However, [28] and [29] did not consider the queue information, or delay information, in their framework. Accordingly, they used bit-rate driven utilities with the sequential approach for mixed traffic.

H. Notes

Generalized Flow Concept: The joint RT and NRT flows packet scheduling and RB allocation through disutility functions not only increases the efficiency of the system but also has a futuristic application. It is common to have the HOL delay deadline requirement for RT flows. However, most NRT flows are not completely insensitive to delay. Based on our approach, one can introduce different levels of delay sensitivity for NRT flows as well. We use the term generalized nonrealtime (GNRT) for those NRT flows that have a mean delay deadline $\bar{d}_{\phi}^{\max}$. The concept GNRT is a good model for NRT flows QoS measures, such as the file transfer time. We further elaborated on GNRT concept in [55]. The joint RT and NRT flows packet scheduling and RB allocation enables to define future QoS classes and to accommodate differentiated services between pure RT flows and pure NRT flows.

Interference: It is worth highlighting that the proposed joint approach, in this paper, is designed based on the static interference assumption. In fact, the intercell interference coordination (ICIC) [68] works in a much longer timescale, in comparison with the packet scheduling and RB allocation algorithm, to specify which RB should be muted or demuted for each cell or sector. Note that the ICIC schemes often do not aim for the QoS requirements or sophisticated fairness notions. Therefore, integrating ICIC (see [68] and the references therein) right into the packet scheduling and RB allocation core, through a systematic design, can be suggested as a promising future direction.

V. Proposed Joint Realtime and Nonrealtime Flows Packet Scheduling and Resource Block Allocation Algorithm

Given the optimization of RT and NRT flows in Section III and the design of the disutilities in (12) and (16), we develop the algorithm here. To make the best change in the objective by increasing only one optimization variable, the variable with the steepest gradient should be chosen as

$$(\phi^*, j^*) = \arg \max_{\substack{1\le\phi\le|\Phi|,\\ 1\le j\le N}} \frac{\partial D_{\text{net}}^{\text{joint}}(\cdot)}{\partial x_{\phi}^{(j)}[k]}. \tag{37}$$

This equation repeatedly determines the flow ϕ^* to be transmitted on an RB on subchannel j^*. The corresponding values of (37), or equivalently (12) and (16), for RT and NRT flows will be denoted by $\Gamma_{\phi\in\Phi_{\text{RT}}}^{(j)}$ and $\Gamma_{\phi\in\Phi_{\text{NRT}}}^{(j)}$, respectively, in the proposed algorithm.

In this part, we propose the algorithm named **Algorithm** Joint RT-and-NRT Flows Packet Scheduling and RB Allocation (PSRA). Step 1 makes a copy of AMC values, in frame k, and the number of available slots in each subchannel. Step 2 and Step 12 implement a loop until all RBs are assigned in frame k. Step 3 and Step 4 fill the matrix $\Gamma_{\phi}^{(j)}$ for NRT and RT flows, respectively. Step 5 decides which flow, denoted by ϕ^*, should be emptied on which subchannel, denoted by j^*, based on the largest element in $\Gamma_{\phi}^{(j)}$. Variable $\Gamma_{\phi}^{(j)}$ is equal to the gradient of RT flows and the gradient of NRT flows for $\phi \in \Phi_{\text{RT}}$ and $\phi \in \Phi_{\text{NRT}}$, respectively. The ties are broken with a uniform random variable. Step 6 executes the decided schedule. The number of unassigned RBs is updated in Step 7. Step 8 updates the intermediate quantities based on the last decision. In other words, the packet scheduling and RB allocation algorithm makes decisions one RB at a time and updates queues and other quantities, such as HOL delay, after each assignment and before finding the next flow for the next RB. Steps 9–11 make the AMC copy of the fully occupied subchannels to zero so that RBs on the corresponding subchannels are not selected again.

Algorithm Joint RT-and-NRT Flows PSRA

1: $\forall\, j, \phi : \tilde{b}_{\phi}^{(j)} \leftarrow b_{\phi}^{(j)}[k],\ \forall\, j : T^{(j)} = T.$
2: **while** $\exists\, T^{(j)} > 0$ **do**
3: $\quad \forall\, \phi \in \Phi_{\text{NRT}} :\ \Gamma_{\phi}^{(j)} \leftarrow \min\left[\xi, \tilde{b}_{\phi}^{(j)} \frac{F_{\phi}^{\bar{q}}(\bar{q}_{\phi}[k])}{F_{\phi}^{\bar{r}}(\bar{r}_{\phi}[k])}\right].$
4: $\quad \forall\, \phi \in \Phi_{\text{RT}} : \Gamma_{\phi}^{(j)} \leftarrow \left(\tilde{b}_{\phi}^{(j)}\right)^{\kappa} \frac{F_{\phi}^{d^{\text{HOL}}}(d_{\phi}^{\text{HOL}}[k])}{F_{\phi}^{\bar{r}}(\bar{r}_{\phi}[k])}.$
5: $\quad (\phi^*, j^*) \leftarrow \arg\max_{\phi, j} \Gamma_{\phi}^{(j)}.$
6: $\quad$ Empty flow, ϕ^*, to an empty RB on subchannel, j^*. $x_{\phi^*}^{(j^*)}[k] \leftarrow x_{\phi^*}^{(j^*)}[k] + 1.$
7: $\quad T^{(j^*)} \leftarrow T^{(j^*)} - 1.$
8: $\quad$ Update $d_{\phi}^{\text{HOL}}[k]$, $\bar{q}_{\phi}[k]$, and $\bar{r}_{\phi}[k]$.
9: $\quad$ **if** $T^{(j^*)} = 0$ **then**
10: $\quad\quad \forall\, \phi \in \Phi : \tilde{b}_{\phi}^{(j^*)} \leftarrow 0.$
11 $\quad$ **end if**
12: **end while**

Computational Complexity: We note that the computational complexity of the **Algorithm** Joint RT-and-NRT Flows PSRA depends on the exact implementation, particularly for the execution of the arg max in Step 5. We estimate the computational complexity of the algorithm based on the worst case complexity of finding the maximum element in an array. This gives a bound on the average computational complexity. In each execution of the "while loop," Steps 3 and 4 have five operations for $N|\Phi_{\text{NRT}}|$ and five operations for $N|\Phi_{\text{RT}}|$ elements, respectively, which makes $5N|\Phi_{\text{NRT}}| + 5N|\Phi_{\text{RT}}| = 5N|\Phi|$ number of operations. It also takes $\mathcal{O}(N|\Phi|)$ operations to find the maximum in Step 5 (worst case). Taking into account that there are NT RBs, the computational complexity of the algorithm is estimated as $\mathcal{O}(N^2T(6|\Phi| + 1))$.

VI. Behavioral Study of Input–Output Bit-Rates

Here, the behavior of the RT and NRT flows output bit-rates versus total input bit-rate will be analyzed. We start by a discussion on how the capacity of the system depends on input and the structure of the algorithm. Subsequently, the input model, the RT output, and the NRT output will be analyzed. Finally, the underload region, the sat. I region, and the sat. II region will be identified and explained.

This study sheds light in understanding the system input–output bit-rates dynamics, in identifying different load regions, and in explaining the simulation results in Section VII.

A. Capacity Definitions and Their Dependence on the Load and on the Algorithm

The capacity of the system depends on the structure of the algorithm, as well as on the input traffic mixture. In other words, the capacity depends on the way the algorithm allocates resources to the RT and NRT mixture in the input and shapes the mixture in the output. We denote the capacity when the system only allocate resources to RT flows by Ω_{RT}. This case happens either when there is no resource remaining for NRT flows (RT flows occupy the system capacity and the system is in overload) or when the input only consists of RT-only flows. Similarly, the capacity of the system in NRT-only traffic is denoted by Ω_{NRT}. This capacity is realized when the input only consists of NRT-only flows. In addition to Ω_{RT} and Ω_{NRT}, when the server is *saturated with the lowest load (SLL)*, its capacity is denoted by Ω_{SLL}. The server is at SLL when the input load to the system reaches the point that the system is at the edge of the overload and the underload. This is when the server is full with the lowest input bit-rate. From this point on, the system cannot serve the total arrivals. Generally

$$\Omega_{\mathrm{RT}} \leq \Omega_{\mathrm{SLL}} \leq \Omega_{\mathrm{NRT}}. \tag{38}$$

The underlying reason for the first inequality in (38) is twofold: First, when the system capacity allows having both RT and NRT flows in the output, the multiuser diversity level is higher than that when RT-only flows are in the output. Second, serving RT-only flows when the system capacity is reached reduces the opportunity to wait for RBs with better links SINR due to the RT flows' delay deadlines stress. The second inequality in (38) is due to the similar higher level of multiuser diversity when we have NRT flows (no deadline) in comparison with the case when we have both RT and NRT flows. In other words, HOL delay deadlines in RT flows override the opportunistic transmissions and reduce multiuser diversity. Inequality (38) will be further explained in Section VI-C, after a preliminary discussion in Section VI-B.

B. Input Model and RT Fraction

Assume that the total input bit-rate $\Lambda_{\Sigma}^{\mathrm{in}}$ is composed of RT bit-rate $\Lambda_{\mathrm{RT}}^{\mathrm{in}}$ and NRT bit-rate $\Lambda_{\mathrm{NRT}}^{\mathrm{in}}$, i.e.,

$$\Lambda_{\Sigma}^{\mathrm{in}} = \Lambda_{\mathrm{RT}}^{\mathrm{in}} + \Lambda_{\mathrm{NRT}}^{\mathrm{in}}, \tag{39}$$

with

$$\Lambda_{\mathrm{RT}}^{\mathrm{in}} = f_{\mathrm{RT}} \Lambda_{\Sigma}^{\mathrm{in}}, \tag{40}$$

and

$$\Lambda_{\mathrm{NRT}}^{\mathrm{in}} = (1 - f_{\mathrm{RT}}) \Lambda_{\Sigma}^{\mathrm{in}}, \tag{41}$$

all in bits per second, where $0 \leq f_{\mathrm{RT}} \leq 1$, which is similar to the input model in [13].

C. RT Output Bit-Rate

Given the aforementioned definitions in the input model, RT output bit-rate $\Lambda_{\mathrm{RT}}^{\mathrm{out}}$ is limited to its input $\Lambda_{\mathrm{RT}}^{\mathrm{in}}$ and, naturally, to the system capacity for RT-only traffic Ω_{RT}, i.e.,

$$\Lambda_{\mathrm{RT}}^{\mathrm{out}} = \min\left(\Lambda_{\mathrm{RT}}^{\mathrm{in}}, \Omega_{\mathrm{RT}}\right). \tag{42}$$

Normalized Input Load Coefficient: Since we are after the general behavior of the output bit-rates, we normalize the total input bit-rate to the SLL capacity, when the saturation happens with the lowest possible input load. This defines the normalized input load coefficient as

$$\rho = \frac{\Lambda_{\Sigma}^{\mathrm{in}}}{\Omega_{\mathrm{SLL}}}. \tag{43}$$

Two specific ρ values, namely

$$\rho_{\mathrm{SLL}} = 1, \tag{44}$$

and

$$\rho_{\mathrm{RTO}} = \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}} \tag{45}$$

are important in determining the different load regions and capacity of the system. Input load coefficient $\rho_{\mathrm{SLL}} = 1$ corresponds to the server saturation with the lowest possible input load. Nevertheless, from this point on, two different saturation regions can be identified. For the load coefficient larger than 1 but smaller than ρ_{RTO}, the server can still serve a portion of NRT flows. For the load coefficient larger than ρ_{RTO}, however, the server output consist of RT-only flows, despite the fact that there exist NRT flows in the input. Input load coefficient ρ_{RTO} is equal to $\rho_{\mathrm{RTO}} = \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}}$, which corresponds to the point where the RT output is equal to the RT-only capacity. This happens when the total input bit-rate is high enough that its RT flows portion ($f_{\mathrm{RT}} \Lambda_{\Sigma}^{\mathrm{in}}$) is equal to the RT-only capacity (Ω_{RT}) and results in (45) [see (40) and (42)]. We will elaborate more on the two saturation regions in Section VI-G.

Dependence of the Capacity on Input Load: For finding the NRT output bit-rate, we need knowledge on the capacity of the system for the input loads larger than $\rho_{\mathrm{SLL}} = 1$. We show the capacity when the input load coefficient is larger than 1 ($\rho_{\mathrm{SLL}} \leq \rho$) but smaller than the load coefficient point where the RT-only flows fill the server ($\rho \leq \rho_{\mathrm{RTO}}$) by Ω_{ρ}. This capacity is showing the dependence of capacity on input load. The capacity Ω_{ρ} is a nonincreasing function over ρ due to the decrease in multiuser diversity and the increase in the pressure of the RT flows' delay deadlines. The tighter the RT flows' requirements, the more the degradation in capacity when the system serves RT-only flows, in comparison with when it serves NRT flows. In fact, the RT flows' requirements are casted as $d_{\phi}^{\mathrm{HOL}^{\max}}$. The tighter the values of $d_{\phi}^{\mathrm{HOL}^{\max}}$, the lower the $\Omega_{\mathrm{RT}}/\Omega_{\mathrm{SLL}}$. A similar observation of the degradation in capacity due to RT requirements has been reported in [69]. The discussion on the dependence of the capacity on the input bit-rate of RT and NRT flows mixture is not within the scope of this study. In fact, for many combined packet scheduling and RB allocation of RT and NRT flows, the capacity has a complex behavior versus total input bit-rate (we will observe this in simulation experiments in Section VII). Moreover, this section does not consider the effects of packet deletion due to passed deadline and/or finite-buffer assumptions. However, we show a linear model, with respect to ρ, for the general behavior of the capacity, after SLL until the RTO. We will see in the simulation section that this model can show the general behavior of the system.

As discussed earlier, the capacity denoted by Ω_{ρ} is equal to Ω_{SLL} at the saturation with the lowest possible input load ($\rho = \rho_{\mathrm{SLL}}$) and is equal to Ω_{RT} when the output consists of RT-only

($\rho = \rho_{\mathrm{RTO}}$). Therefore, the general behavior of the capacity can be represented as

$$\Omega_\rho = \begin{cases} \Omega_{\mathrm{SLL}}, & \rho \le 1 \\ \frac{f_{\mathrm{RT}}\Omega_{\mathrm{SLL}}(\Omega_{\mathrm{RT}}-\Omega_{\mathrm{SLL}})}{\Omega_{\mathrm{RT}}-f_{\mathrm{RT}}\Omega_{\mathrm{SLL}}}(\rho-1) & \\ \quad +\Omega_{\mathrm{SLL}}, & 1 \le \rho \le \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}} \\ \Omega_{\mathrm{RT}}, & \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}} \le \rho. \end{cases} \tag{46}$$

This model shows that the server capacity decreases versus input bit-rate when $1 \le \rho \le \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}}$. Note that we use this model only to show the general system behavior. The simulation experiments, in the following, is independent of the capacity model here. The capacity model from this section is neither used nor necessary in simulation experiments.

D. NRT Output Bit-Rate

Having discussed the RT output bit-rate and the dependence of the capacity of system on input load, we will analyze the NRT output bit-rate here. The NRT output bit-rate $\Lambda^{\mathrm{out}}_{\mathrm{NRT}}$ is limited to its input $\Lambda^{\mathrm{in}}_{\mathrm{NRT}}$, the capacity of the system in NRT-only traffic Ω_{NRT}, and the remaining capacity after serving RT flows (equal to $\max(0, \Omega_\rho - \Lambda^{\mathrm{in}}_{\mathrm{RT}})$), i.e.,

$$\Lambda^{\mathrm{out}}_{\mathrm{NRT}} = \min\left(\Lambda^{\mathrm{in}}_{\mathrm{NRT}}, \Omega_{\mathrm{NRT}}, \max\left(0, \Omega_\rho - \Lambda^{\mathrm{in}}_{\mathrm{RT}}\right)\right) \tag{47}$$

or by substituting (40) and (41) into (47), i.e.,

$$\Lambda^{\mathrm{out}}_{\mathrm{NRT}} = \min\left((1-f_{\mathrm{RT}})\Lambda^{\mathrm{in}}_{\Sigma}, \Omega_{\mathrm{NRT}}, \max\left(0, \Omega_\rho - f_{\mathrm{RT}}\Lambda^{\mathrm{in}}_{\Sigma}\right)\right). \tag{48}$$

We note that, since the input bit-rate $\Lambda^{\mathrm{in}}_{\Sigma}$ is unbounded, the term max in $\max(0, \Omega_\rho - f_{\mathrm{RT}}\Lambda^{\mathrm{in}}_{\Sigma})$ is necessary to keep the remaining capacity nonnegative.

E. Discussion on the RT and NRT Output Bit-Rates

Understanding the behavior of RT and NRT output bit-rates versus input bit-rate is important. Equations (42) and (48) describe this behavior for RT and NRT, respectively. Generally, it is convenient to depict the behavior of normalized RT and NRT output based on normalized input bit-rate, which is defined in (43), as

$$\frac{\Lambda^{\mathrm{out}}_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}}} = \min\left(f_{\mathrm{RT}}\rho, \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}}}\right), \tag{49}$$

and

$$\frac{\Lambda^{\mathrm{out}}_{\mathrm{NRT}}}{\Omega_{\mathrm{SLL}}} = \min\left((1-f_{\mathrm{RT}})\rho, \frac{\Omega_{\mathrm{NRT}}}{\Omega_{\mathrm{SLL}}}, \max(0, 1-f_{\mathrm{RT}}\rho)\right). \tag{50}$$

Normalization helps to focus on the general behavior and needs fewer assumptions for absolute values of the capacities.

Figs. 1 and 2 show RT, NRT, and total output bit-rates versus total input bit-rate, all normalized to Ω_{SLL}, when RT input bit-rate is dominant ($f_{\mathrm{RT}} > 0.5$) and vice versa ($f_{\mathrm{RT}} < 0.5$), respectively. The value of $\Omega_{\mathrm{RT}}/\Omega_{\mathrm{SLL}}$ is assumed to be equal to 0.85 in Figs. 1 and 2 for showing a typical behavior. The saturation region, for input loads larger than ρ_{SLL} but smaller than ρ_{RTO}, is highlighted in Figs. 1 and 2. This region is labeled as sat. I.

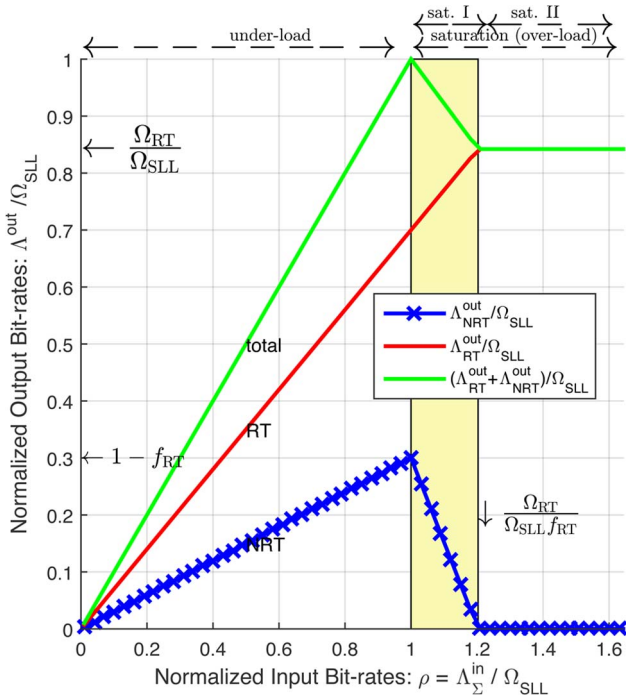

Fig. 1. Normalized output bit-rates of RT flows, NRT flows, and total versus normalized input bit-rate, with the assumption of $f_{\mathrm{RT}} = 0.7$ and $\Omega_{\mathrm{RT}}/\Omega_{\mathrm{SLL}} = 0.85$, for the behavioral study.

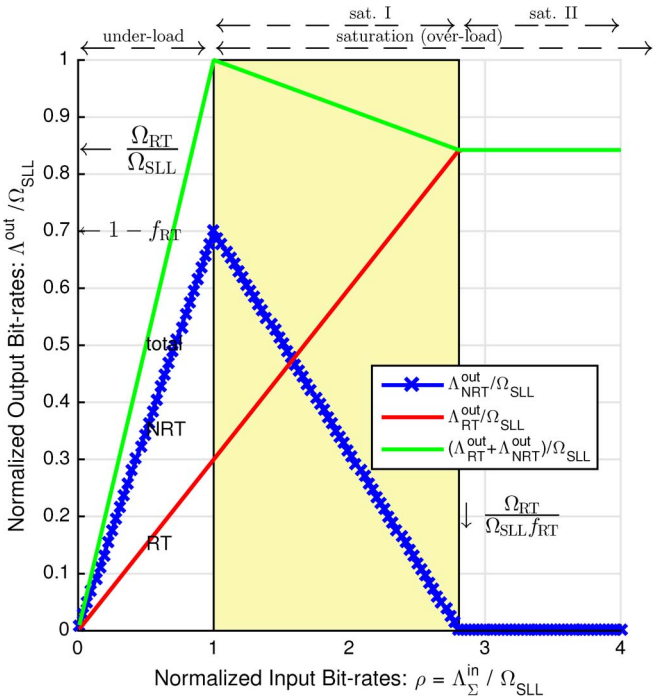

Fig. 2. Normalized output bit-rates of RT flows, NRT flows, and total versus normalized input bit-rate, with the assumption of $f_{\mathrm{RT}} = 0.3$ and $\Omega_{\mathrm{RT}}/\Omega_{\mathrm{SLL}} = 0.85$, for the behavioral study.

Note that the absolute values of the capacities depend not only on server attributes (such as channels) but also on the RT flows's requirements and fairness parameters. As an example, for tighter values of $d_\phi^{\mathrm{HOL}^{\max}}$, the RT-only capacity degrades more when passing RTO load coefficient (ρ_{RTO}). In other words, for tighter values of $d_\phi^{\mathrm{HOL}^{\max}}$, the value of $\Omega_{\mathrm{RT}}/\Omega_{\mathrm{SLL}}$ decreases. As we discussed earlier in this paper, decreasing

the parameter ξ makes the algorithm go toward the sequential approach. Therefore, decreasing ξ also decrease Ω_{SLL}.

In Sections VI-F and G, we elaborate more on RT and NRT output bit-rates versus ρ for both Figs. 1 and 2, in different load regions, namely, underload and overload regions (sat. I region and sat. II region).

F. Underload Region

In the underload region, as shown in Figs. 1 and 2, RT output bit-rate $\Lambda_{\mathrm{RT}}^{\mathrm{out}}$ is equal to its input (goes up with slope equal to f_{RT}) until the input reaches RT-only system capacity Ω_{RT} and becomes constant at $\Lambda_{\Sigma}^{\mathrm{in}} = \frac{\Omega_{\mathrm{RT}}}{f_{\mathrm{RT}}}$, or equivalently, at $\rho = \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}}$.

The NRT output is equal to the NRT input until the input load reaches the server capacity, i.e., $\Lambda_{\Sigma}^{\mathrm{in}} = \Omega_{\mathrm{SLL}}$, or equivalently, $\rho = 1$. At this point, $\rho = 1$, the NRT output reaches $\Lambda_{\mathrm{NRT}}^{\mathrm{out}} = (1 - f_{\mathrm{RT}})\Omega_{\mathrm{SLL}}$, and the RT output reaches $\Lambda_{\mathrm{RT}}^{\mathrm{out}} = f_{\mathrm{RT}}\Omega_{\mathrm{SLL}}$. The system becomes saturated, which means that all the resources in OFDMA server are occupied. Note that, although the system is saturated, the output of the system still consists of both RT and NRT flows.

This explanation summarizes the first part of the outputs, where RT output $\frac{\Lambda_{\mathrm{RT}}^{\mathrm{out}}}{\Omega_{\mathrm{SLL}}} \in (0, f_{\mathrm{RT}})$, and NRT output $\frac{\Lambda_{\mathrm{NRT}}^{\mathrm{out}}}{\Omega_{\mathrm{SLL}}} \in (0, 1 - f_{\mathrm{RT}})$. See Figs. 1 and 2 in the region where $\rho < 1$. This region is labeled as underload.

G. Overload Regions: Saturation I and Saturation II Regions

For $\Lambda_{\Sigma}^{\mathrm{in}} > \Omega_{\mathrm{SLL}}$, or equivalently, $\rho > 1$, the priority of RT flows makes the NRT output bit-rate break and go down until $\Lambda_{\Sigma}^{\mathrm{in}} = \frac{\Omega_{\mathrm{RT}}}{f_{\mathrm{RT}}}$, or equivalently, $\rho = \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}}$, despite the increase in its input. In saturation regions (or equivalently, $\rho > 1$, which is labeled as saturation), two subregions can be identified, which are further labeled as sat. I and sat. II. For the region where $\frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}} > \rho > 1$, the system is saturated but is still able to pass a portion of NRT flows. This region is depicted in shaded color. It is important to note that networks are expected to operate in moderate saturation region (which is sat. II), where the operator benefits from its investment efficiently, while clients can have both RT and NRT flows. For the region where $\frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}} \leq \rho$, RT flows merely fill the system capacity, and the system will not even be able to pass all RT flows. This region is labelled by sat. II.

This explanation outlines the second part of the outputs, where $\frac{\Lambda_{\mathrm{RT}}^{\mathrm{out}}}{\Omega_{\mathrm{SLL}}} \in (f_{\mathrm{RT}}, \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}}})$, and $\Lambda_{\mathrm{NRT}}^{\mathrm{out}}/\Omega_{\mathrm{SLL}}$ goes down from $1 - f_{\mathrm{RT}}$ to 0. See Figs. 1 and 2 in the regions where $1 \leq \rho \leq \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}}$ (shaded color), and $\rho \leq \frac{\Omega_{\mathrm{RT}}}{\Omega_{\mathrm{SLL}} f_{\mathrm{RT}}}$.

In the case when RT and NRT flows share the equal input bit-rate ($f_{\mathrm{RT}} = 0.5$), RT output $\Lambda_{\mathrm{RT}}^{\mathrm{out}}$ and NRT output $\Lambda_{\mathrm{NRT}}^{\mathrm{out}}$ go up with same slope equal to $f_{\mathrm{RT}} = 0.5$ until $\frac{\Lambda_{\Sigma}^{\mathrm{in}}}{\Omega_{\mathrm{SLL}}} = 1$. Then, $\Lambda_{\mathrm{RT}}^{\mathrm{out}}/\Omega_{\mathrm{SLL}}$ continues to increase, but $\Lambda_{\mathrm{NRT}}^{\mathrm{out}}/\Omega_{\mathrm{SLL}}$ decreases until $\frac{\Lambda_{\Sigma}^{\mathrm{in}}}{\Omega_{\mathrm{SLL}}} = 2$, when RT flows saturate the system, and there is no remaining capacity for NRT flows. We did not include the graphical representation of this case, due to its less importance in comparison with Figs. 1 & 2 and due to space limitations.

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Fading	Rayleigh [72]
Shadowing	Log-normal, s.d. 5 dB
Doppler shift	37 Hz
Path loss	$38.4 + 2.35 \log_{10}(d)$ dB
Total bandwidth	10 MHz
Number of sub-channels	$N = 20$
Sub-carrier bandwidth	25 kHz
Sub-carriers per sub-channel	20
Slots per frame	$T = 1$
Frame duration	1 ms
Cell radius	1000 m
Close-in minimum distance	35 m
Transmit power	30 dBm BS
Antenna gain	5 dBi BS, 0 dBi WTs
Noise figure	2 dB BS, 2 dB WTs

In this paper, we have not considered finding analytic expressions for capacities, analytic expressions for delay performances, or an analytic approach for prediction of the multiuser diversity gain (see [61], [70], and [71] for some analytic evaluation of legacy designs in much simpler settings). Associated analytic expressions could be research topics for future work.

VII. SIMULATION

We have developed a comprehensive simulation platform for packet delay simulation for a single cell in MATLAB. The platform incorporates correlated fading in time and frequency with Rayleigh fading, shadowing, and path loss based on [72]. The total OFDMA bandwidth is 10 MHz divided into 20 subchannels, each subchannel consists of 20 subcarriers, each with 25-kHz span in frequency. We used 14 AMC levels (including zero) for AMC table (represented by ϝ in (1) in the system model), which is the result of QPSK, 16-QAM, and 64-QAM in conjunction with 14 code rates from 0.105 up to 0.801 [73].

A. SINR Distribution

We test the algorithm based on equal average SINR for all flows. For saving time on wireless channel simulation, we used a three-step method: First, we find a high-resolution SINR distribution, resulting from large-scale fading. Second, one cell is simulated by finding the SINRs from the aforementioned SINR distribution. Third, we generate small-scale time and frequency Rayleigh fading, independently for each flow, and add the corresponding average SINRs (result of the large-scale fading, drawn in the second step) to get the instantaneous SINR [73]. Simulation parameters are summarized in Table II.

B. Simulation Assumptions

We tested the proposed joint RT and NRT flows packet scheduling and RB allocation with three arrival scenarios, four RT flows, four NRT flows, and a mixture of two RT flows and two NRT flows. The assumption for delay requirements is 20 ms (or 20 frames) on RT flows. The arrival bit-rates are proportional to [1, 2, 3, 4] according to the total load in the system as $\Lambda_1^{\mathrm{in}} = \frac{1}{10}\Lambda_{\Sigma}^{\mathrm{in}}, \Lambda_2^{\mathrm{in}} = \frac{2}{10}\Lambda_{\Sigma}^{\mathrm{in}}, \Lambda_3^{\mathrm{in}} = \frac{3}{10}\Lambda_{\Sigma}^{\mathrm{in}}$, and $\Lambda_4^{\mathrm{in}} = \frac{4}{10}\Lambda_{\Sigma}^{\mathrm{in}}$, where $\Lambda_{\Sigma}^{\mathrm{in}}$ is the total input bit-rate. Higher

(lower) loads have the same input pattern, but with higher (lower) total input bit-rate $\Lambda_{\Sigma}^{\text{in}}$. We investigated the output bit-rate and the 99th percentile of the HOL delay cumulative distribution function (CDF) versus total input bit-rate. The results of complete separation of RT and NRT flows with EDF-PF, MLWDF-PF, and EXP-PF are also produced for comparison.

To find the load range that covers interesting load regions, we estimate the capacity of our OFDMA system based on the 0.3 portion of the highest AMC level on RBs. Based on this estimation, we found a load coefficient range by multiplying this capacity from 0.05 to 5. The interval [0.05, 5] was large enough to cover all the interesting load situations.

We assume that the RT packets will be discarded, if their deadlines are passed. We also use a finite-buffer assumption for both RT and NRT flows, which is equal to 40 Mb. However, the probability of overflow for RT flows is very low due to their deadline timescale.

C. RT-Only and NRT-Only Traffic

For the proof of concept, we first test the algorithm with either RT-only or NRT-only flows. We expect the proposed algorithm to reduce to the channel aware version of EDF [see (10)] for RT-only flows and PF for NRT-only flows. Note that, in RT-only flows, finding the highest value from (12) with $\kappa = 1$ is equivalent to finding the highest value from (10). We observed that the proposed algorithm outperforms the baseline (which is EDF for RT flows) in this case and releases the potential increase in admissible bit-rates without any compromise in the 99th percentile of the HOL delay. Moreover, we also observed that, irrespective of the load situation and link qualities, the output bit-rates per flow are in the same order of the input bit-rates. We also tested the proposed approach versus PF for four NRT-only flows. In NRT-only experiment, we observed that, in underload situation, the output bit-rates are proportional to input bit-rates. However, in an overload situation, the flows with better wireless links get more RBs, in compliance with the expected bit-rate fairness. Figures for the RT-only experiment and the NRT-only experiment are not included, due to their less importance in comparison with mixed traffic.

D. Mixed-Traffic Experiment 1

We tested the algorithm in the mixed scenario of two RT flows and two NRT flows. Flows 1 & 2 are RT, and flows 3 & 4 are NRT in the experiments. We tested the proposed approach versus sequential EDF-PF in mixed traffic in experiment 1. The output bit-rate per flow, the total output bit-rate of RT and NRT flows, and the 99th percentile of the HOL delay are shown in Figs. 3–5, respectively.

Discussion on Output Bit-Rate: Based on the analysis in Section VI and since RT flows occupy

$$f_{\text{RT}} = \frac{\sum_{\phi \in \Phi_{\text{RT}}} \Lambda_{\phi}^{\text{in}}}{\sum_{\phi \in \Phi} \Lambda_{\phi}^{\text{in}}} \tag{51}$$

fraction (which is equal to $\frac{1+2}{1+2+3+4}$) of the total input bit-rate, the saturation of the system with RT flows (when there

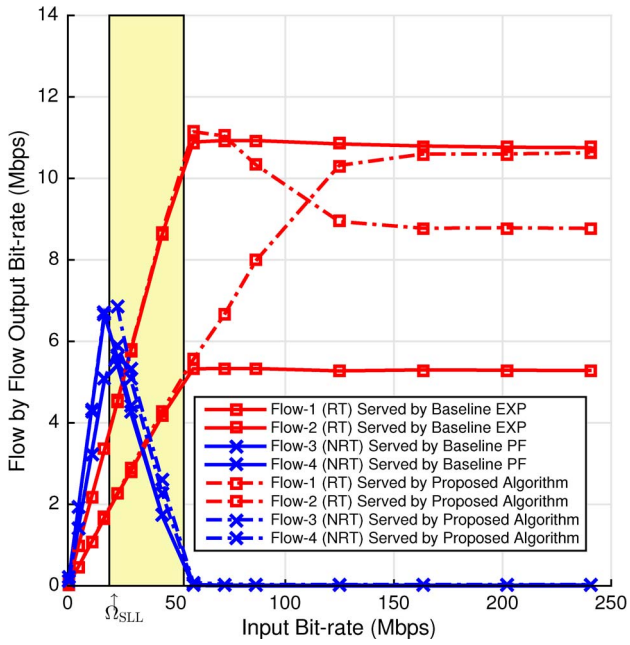

Fig. 3. Flow-by-flow output bit-rates versus input bit-rate in mixed scenario, in comparing EDF-PF (as the baseline) with the proposed algorithm.

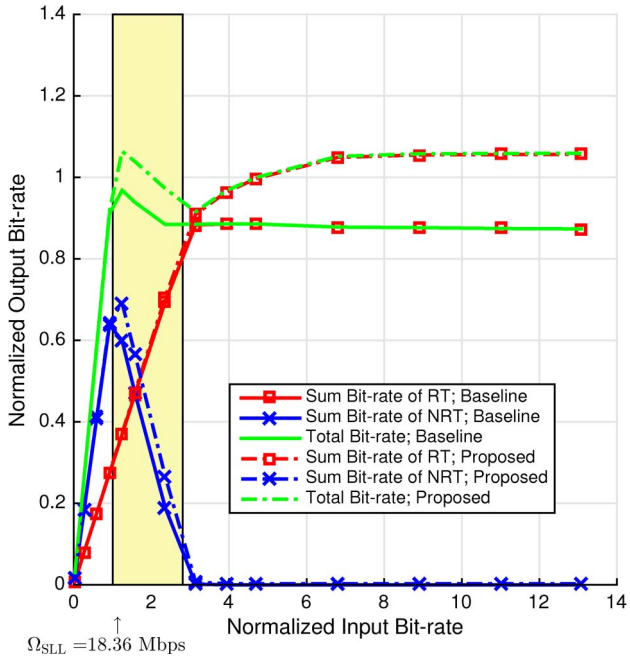

Fig. 4. Sum RT and NRT output bit-rates versus input bit-rate in mixed scenario, in comparing EDF-PF (as the baseline) with the proposed algorithm.

is only RT flows in the output) happens around normalized load coefficient that is equal to $\frac{1}{f_{\text{RT}}} = \frac{10}{3}$ for the baseline (EDF-PF). Note that the better estimate is $\frac{\Omega_{\text{RT}}}{\Omega_{\text{SLL}} f_{\text{RT}}}$. In addition, the NRT output bit-rate at its highest point is around the normalized load coefficient that is equal to $1 - f_{\text{RT}} = 0.7$. Fig. 4 verifies the predicted results from Section VI. Due to the fact that figures are crowded with plots, we inevitably used colored plots (see the soft copy). The red lines indicate the total output bit-rate of the RT flows. The blue lines indicate the total output

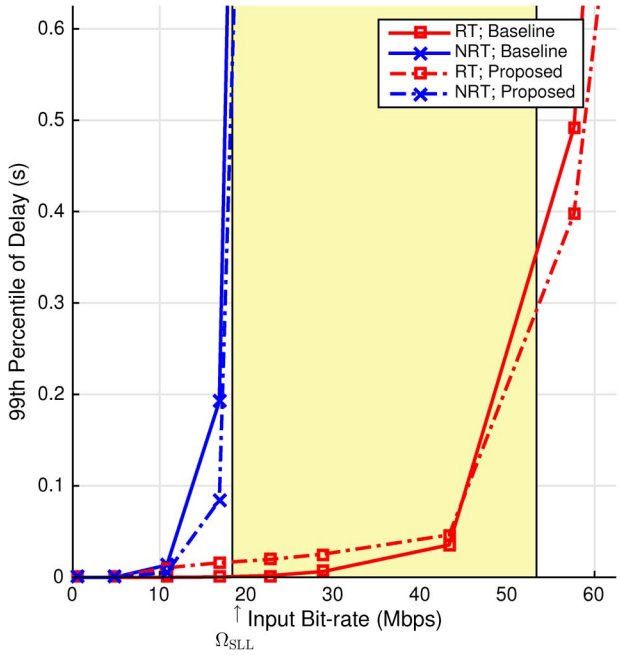

Fig. 5. Ninety-ninth percentile of delay CDF versus input bit-rate in mixed scenario in comparing EDF-PF (as the baseline) with the proposed algorithm.

bit-rate of the NRT flows. For both red and blue lines, the solid lines belong to the baseline algorithm, and the dotted lines belong to the proposed algorithm. The green lines are the sum of the red and blue lines, which is equal to the server capacity, in saturation regions. In the sat. I region, the NRT output bit-rate decreases until RT flows saturate the OFDMA server. When the system goes to the sat. II region (NRT flows output is zero), the RT output (red line) is the same as the green lines (sum of RT and NRT). As we anticipated, the framework allows for potential increase in admissible bit-rates, which show themselves as a gap between green lines. In the sat. I region, the gain (gap) is due to the better exploitation of multiuser diversity by the joint RT and NRT flows packet scheduling and RB allocation in a common pool of RBs, in comparison with the sequential approach. In the sat. II region, the sustained gain is due to the channel awareness of the design, in comparison with EDF. As discussed in Section VI, networks are designed to operate in moderate saturation region (sat. I), where the operator benefits from its investment efficiently, while clients can have both RT and NRT flows. The joint approach enables increasing the capacity of moderate saturation region (sat. I).

We note that the achieved gain (through exploiting higher level of multiuser diversity) can be consumed (through modifying the design parameters) over improving either output bit-rates, delay performance, or fairness measures. This makes a 3-D tradeoff among efficiency, fairness, and performance of fulfilling the requirements.

Discussion on Delay: Fig. 5 shows that the average of the delay requirements of RT is not compromised in a significant portion of the load situation. It is worth mentioning that based on the range of the input load to the system, one algorithm potentially can have better delay (or output bit-rate) performance with respect to others, even in a single class of RT flows, or a single class of NRT flows. We note that the stationary and reliable delays are valid up to a fraction of the simulation time. Since we run the simulation for 10 s, we took the first 2500 frames in delay as reliable delays (see Fig. 5). Beyond that, the system is in overload situation, and delays are not stationary and not reliable.

Capacity Dependence on the Load and on the Algorithm: We observe that, for the sequential EDF-PF, the capacity of the system at the SLL is equal to $\Omega_{\mathrm{SLL}} = 17.78$ Mbps and is larger than RTO saturation $\Omega_{\mathrm{RT}} = 16.14$ Mbps. When the system capacity allows having both RT and NRT in the output, the multiuser diversity level is higher. Serving RT-only flows with input bit-rate around the system capacity reduces the opportunity to wait for better RBs, due to the RT flows's delay deadlines. This is when the queues start to build up and the delays start to pass their deadlines constantly. Beyond this point, to bear with the congested RT queues, the system pays more attention to the delay than being more opportunistic. In this overload situation, the multiuser diversity cannot be used as much as before because of delay stress.

We also observed that, when delay deadlines of RT flows become tighter, the increase in system capacity due to the joint RT and NRT flows packet scheduling and RB allocation vanishes. In other words, the tighter the delay deadlines, the lesser the chances to wait for a good RB, thus reducing the multiuser diversity. Note that, for the proposed algorithm, some of the sophisticated effects in output bit-rate cannot be predicted by the model in Section VI, namely, the nonlinearity and the recurring increase in the total output bit-rate after the RTO saturation in Fig. 4 (for the total output bit-rate versus total input bit-rate). This is due to the unavoidable simplification in the Section VI, including not considering the packet deletion due to passed deadline, static separation assumption, and nonlinearity of AMC table versus SINR. Particularly, we think that the recurring increase in the total output bit-rate after the RTO saturation is due to the inevitable constant increase in the packet loss ratio in sat. II, which lessens the delay deadlines stress. As stated previously, sat. I is the most likely operational region.

E. Mixed-Traffic Experiment 2

To further test our proposed approach, we reproduced the algorithms MLWDF-PF and EXP-PF (described in Section IV) and compared the proposed algorithm against them. Figs. 6–8 show the flow-by-flow output bit-rates, the total output bit-rates, and the delay performance, respectively, for the proposed approach against the baseline algorithm MLWDF-PF. In parallel, Figs. 9–11 show the flow-by-flow output bit-rates, the total output bit-rates, and the delay performance, respectively, for the proposed approach against the baseline algorithm EXP-PF. The same observations in Section VII-D are valid in comparing the proposed joint approach against the baseline algorithms MLWDF-PF and EXP-PF. The claimed gain is also observable for the proposed approach in the sat. I region, in comparison with both MLWDF-PF and EXP-PF, with comparable delay performances. Nevertheless, we anticipate that, in a real network with several flows with many heterogeneous delay deadlines and heterogeneous links, higher gain is possible, due to

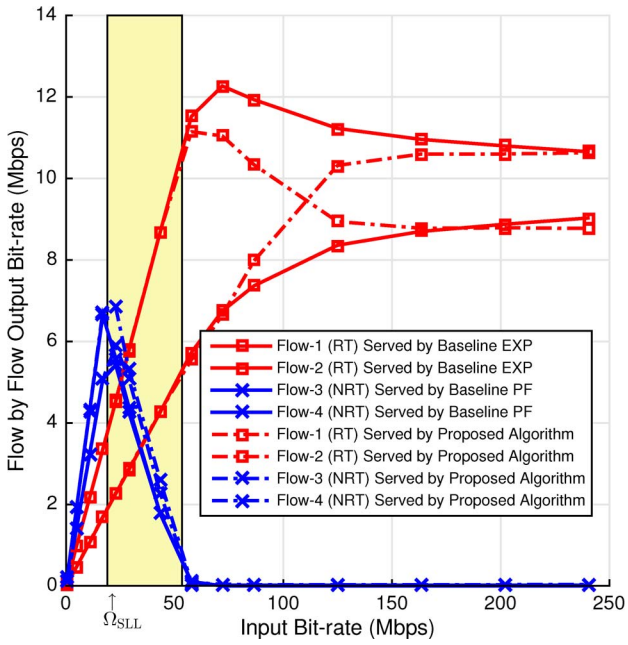

Fig. 6. Flow-by-flow output bit-rates versus input bit-rate in mixed scenario in comparing MLWDF-PF (as the baseline) with the proposed algorithm.

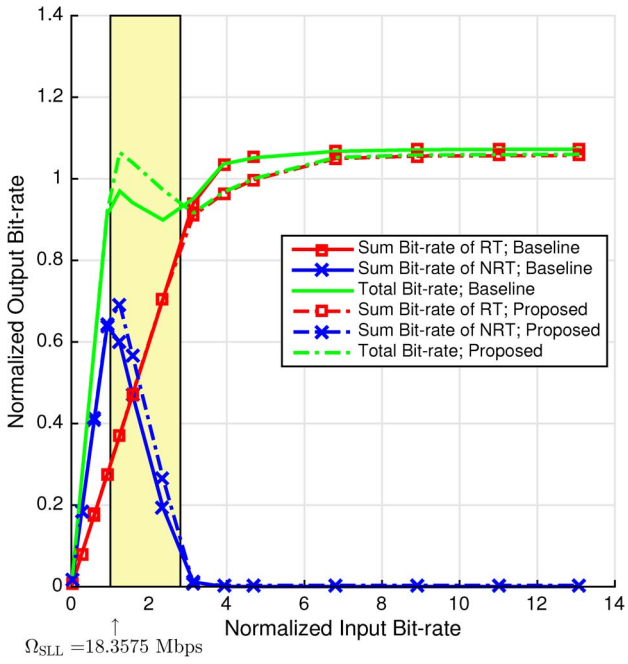

Fig. 7. Sum RT and NRT output bit-rates versus input bit-rate in mixed scenario in comparing MLWDF-PF (as the baseline) with the proposed algorithm.

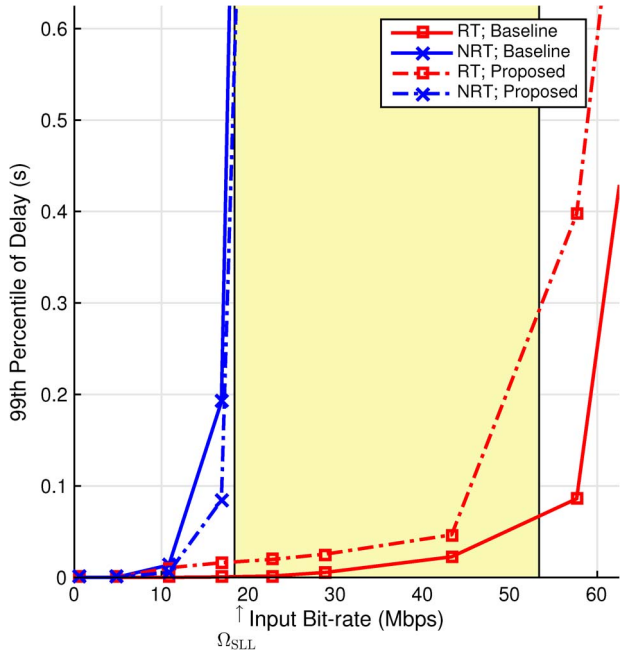

Fig. 8. Ninety-ninth percentile of HOL delay CDF versus input bit-rate in mixed scenario in comparing MLWDF-PF (as the baseline) with the proposed algorithm.

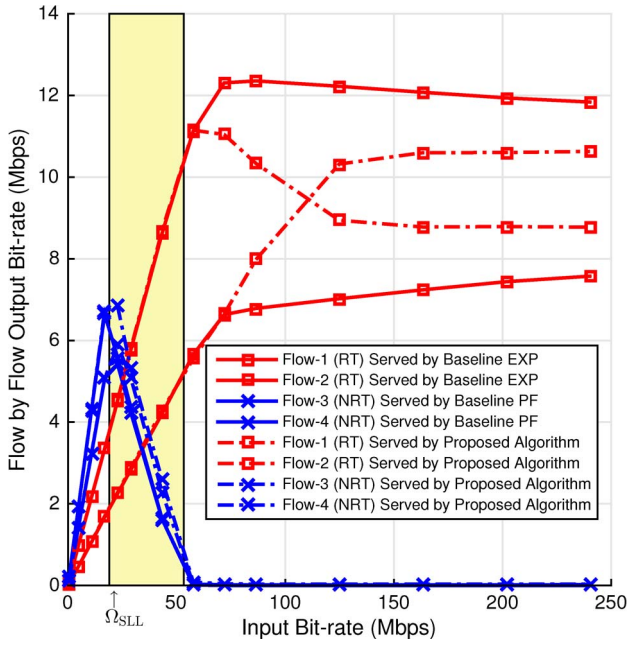

Fig. 9. Flow-by-flow output bit-rates versus input bit-rate in mixed scenario in comparing EXP-PF (as the baseline) with the proposed algorithm.

higher level of potent multiuser diversity exploitable by the proposed joint approach. Having evaluated various algorithms, novel joint RT and NRT flow packet scheduling and RB allocation algorithms that can change their core structure based on the input load situation is also recommended for future work.

F. RB Utilization

Generally, channel RB utilization is the ratio of the consumed RBs over the total available RBs in the OFDMA plane. In packet scheduling and RB allocation, when the number of bits that are assigned to a certain RB is less than its capacity, it is possible that an RB is partially filled. Therefore, we used two measures for RB utilization: one that only considers the number of RBs and one that accounts for the portion for which the RBs are filled.

The first one, which is denoted by $\mathfrak{U}^{\text{non-frag}}$ (called nonfragmented utilization), is based on the number of consumed RBs

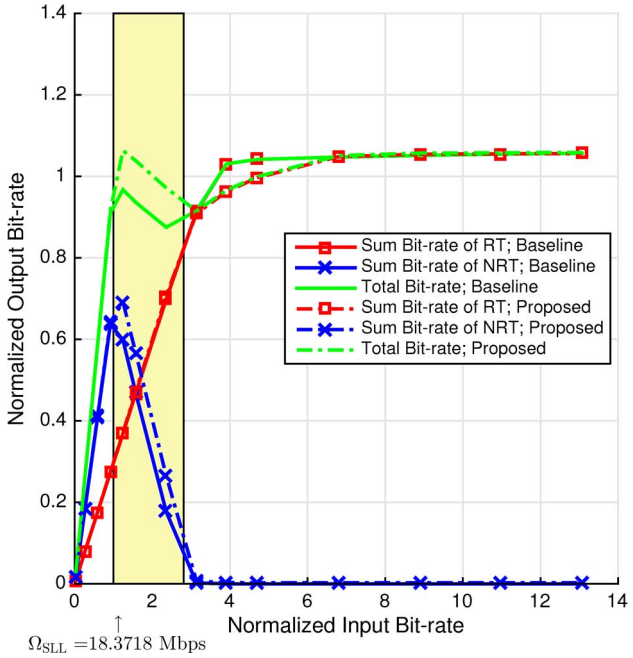

Fig. 10. Sum RT and NRT output bit-rates versus input bit-rate in mixed scenario in comparing EXP-PF (as the baseline) with the proposed algorithm.

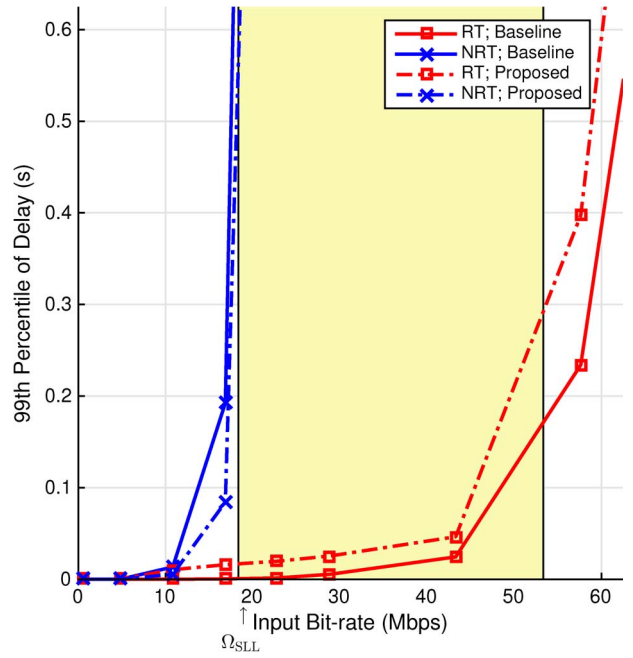

Fig. 11. Ninety-ninth percentile of HOL delay CDF versus input bit-rate in mixed scenario in comparing EXP-PF (as the baseline) with the proposed algorithm.

over the total available RBs in the OFDMA plane, i.e.,

$$\mathfrak{U}^{\text{non-frag}} = \frac{N_{\text{used}}}{N_{\text{total}}}, \tag{52}$$

where N_{used} is the number of used RBs, and N_{total} is the total number of available RBs in a frame.

For the second one, since RBs can be filled partially, we adjusted a novel RB utilization measure, which is denoted by $\mathfrak{U}^{\text{frag}}$ (called fragmented utilization), as the average value of

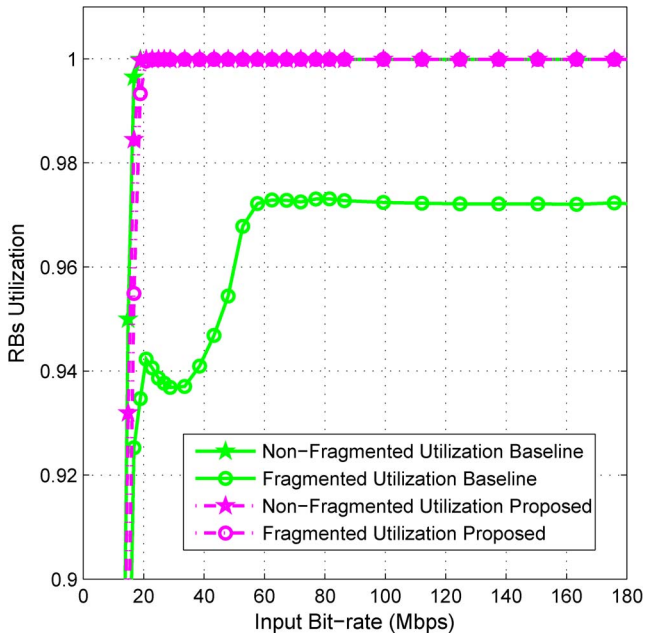

Fig. 12. Channel RB utilization versus input bit-rate in comparing EDF-PF (as the baseline) with the proposed algorithm.

the number of the transmitted bits on each RB over the capacity of each RB defined by

$$\mathfrak{U}^{\text{frag}} = \frac{1}{N_{\text{total}}} \sum_{n=1}^{N_{\text{total}}} \frac{B_n^{\text{tx}}}{B_n^{\text{cap}}}, \tag{53}$$

where B_n^{tx} is the number of transmitted bits on the nth RB, B_n^{cap} is the capacity of the nth RB, and N_{total} is the total number of RBs.

Fig. 12 shows the RB utilization versus the input bit-rate. The nonfragmented utilization in the proposed approach is lower than the baseline, although by a small margin. This observation has been also reported in [22], where they used EXP. Interestingly, it can be observed that fragmented utilization in the proposed algorithm can use the whole capacity of the available RBs, whereas the baseline algorithm is not able to achieve that. This is due to the fact that the proposed algorithm uses the RBs with a higher level of multiuser diversity in comparison with the baseline algorithms.

G. Fairness Evaluation Through Jain's Index

We also evaluated the Jain's index of bit-rates and the Jain's index of the 99th percentile of the HOL delay. We observe that the Jain's index of the *weighted* NRT flows' output bit-rates for the load coefficient smaller than 1 (underload), the Jain's index of the NRT flows' output bit-rates for the load coefficient larger than 1 (overload), and the Jain's index of the 99th percentile of HOL delay for RT flows are in the range of 0.93–1 in the proposed approach, depending on the input load situation. This proves that NRT flows are treated bit-rate fair and RT flows are treated delay fair. Note that, for the normalized input load coefficient smaller than 1 (underload), since there is idle capacity in the server and, more importantly, since the arrival bit-rates are intentionally heterogeneous, the system serves the NRT flows in accordance to their input bit-rates. Therefore, we

used the weighted Jain's index [74] for measuring the NRT output bit-rate's fairness, in underload, as

$$J\left(w_1\Lambda_1^{\text{out}},\ldots,w_{|\Phi|}\Lambda_{|\Phi|}^{\text{out}}\right)=\frac{\left(\sum_{\phi=1}^{|\Phi|}w_\phi\Lambda_\phi^{\text{out}}\right)^2}{|\Phi|\sum_{\phi=1}^{|\Phi|}\left(w_\phi\Lambda_\phi^{\text{out}}\right)^2} \tag{54}$$

which compensates the input bit-rate heterogeneity. The weights are inversely proportional to the input arrival proportions, i.e., $w_\phi=\frac{\Lambda_\Sigma^{\text{in}}}{\Lambda_\phi^{\text{in}}}$. However, in the overload region, the system should ignore the input bit-rate heterogeneity and should provide the service based on bit-rate fairness. The bit-rate fairness of the NRT flows, in the overload region, is due to the structure in (16), where the term $F_\phi^{\bar{r}}(\bar{r}_\phi[k])$ equalizes the output NRT flow's bit-rates. We used pure Jain's index, or equivalently, $w_\phi=1$, in this load situation to assess the bit-rate fairness.

Interestingly, the bit-rate fairness is also observable from flow-by-flow output bit-rates, for example, in Fig. 3. In fact, in the underload region, any arrival to the system (either RT or NRT) will be served. In other words, the slope of each red or blue lines is equal to their proportion in the arrival (that is, $\frac{1}{10}$, $\frac{2}{10}$, $\frac{3}{10}$, and $\frac{4}{10}$). The fairness is not interesting or challenging in this underload situation. When the system goes to saturation regions, the fairness enforcement kicks in. Now, since the system is full, fairness governs the packet scheduling and RB allocation. This is the underlying reason that the output bit-rate of RT flows shows convergent property (see Fig. 3). In fact, for $\phi=2$ (the one that has two tenths of the arrival proportion), the output decreases until it reaches the output of $\phi=1$ (the one that has one tenth of the arrival proportion). The convergence of RT flows is because of fairness feature, embedded in the structure.

We conclude this section by noting that the overall tradeoff in the system design is a three-dimensional tradeoff, for each QoS element (such as bit-rate or delay measure). The tradeoff is among the efficiency of the QoS element (total sum over flows), the fairness of the QoS element (over flows), and guaranteeing the QoS element requirement.

VIII. Conclusion

We have developed a channel aware and delay aware framework and designed appropriate disutilities for the joint RT and NRT flows packet scheduling and RB allocation. The choice of disutility functions results in sorting and making the prioritization within a global set of RT and NRT flows. Exploiting the inherent diversity, including time, frequency, spatial, and multiuser diversity in a wireless system is a key to improving its performance. Joint RT and NRT flows packet scheduling and RB allocation exploits higher levels of multiuser diversity, in comparison with the conventional approach. The joint RT and NRT flows approach is particularly relevant in the context of VoLTE. Wireless systems have been evolved from circuit switching to all-IP packet switching (such as VoLTE). Nevertheless, VoLTE is still implemented based on a sequential approach. We envision the next enhancement step based on the proposed joint approach.

The effectiveness of the proposed approach is validated by extensive simulations. The simulations showed that the proposed algorithm is able to exploit the potent intra- and interclass multiuser diversity, as well as the heterogeneity of traffic in time and among flows. This increases the effective system capacity without significant compromise in delay performance. It is worth highlighting that whenever there is heterogeneity in the resources, including (but not limited to) multiuser diversity or any large-scale or small-scale signal variation, the proposed joint RT and NRT flows approach can offer higher performance, in comparison with the sequential approach. We anticipate that the increase in the heterogeneity of the requirements (when there is heterogeneity in the resources) and/or the increase in the number of wireless terminals increases the gain of the proposed joint approach. We also observed that one fixed algorithm rule is not likely to outperform in terms of output bit-rate and delay performance over all the input loads. Accordingly, the design of packet scheduling and the RB allocation rule that can change its structure, depending on the input load situation, is suggested as future work. The framework can be also extended to incorporate other QoS vector elements, such as other statistics of the delay measures, in the structure.

We have also developed a novel model and analyzed input–output bit-rate behavior of the mixture of RT and NRT flows. This model sheds light to the understanding of the system in a simple and intuitive manner and elaborates on different capacity definitions, capacity dependence on the input load, identifying load regions (namely, underload region, two saturation regions), and the general trends of the output bit-rate of RT and NRT flows.

Acknowledgment

The authors would like to thank Dr. G. Senarath, Dr. P. Djukic, and Dr. H. T. Cheng for their helpful comments. We would also like to thank the associate editor and anonymous reviewers for their valuable suggestions, which helped improve the presentation and the content of this paper.

References

[1] H. Yanikomeroglu and J. Zhang, "Beyond-4G cellular networks: Advanced radio access network (RAN) architectures, advanced radio resource management (RRM) techniques, and other enabling technologies," in *Proc. WWRF21 Meeting*, Oct. 2008, pp. 13–15.

[2] "Internet Traffic Management Practices Guidelines for Responding to Complaints and Enforcing Framework Compliance by Internet Service Providers," Can. Radio Telecommun. Commiss., Ottawa, ON, Canada, Sep. 2011. [Online]. Available: http://www.crtc.gc.ca/eng/archive/2011/2011-609.htm

[3] Chief Judge Tatel, U.S. Court of Appeal, "Comcast vs. FCC," Apr. 2010. [Online]. Available: https://www.eff.org/files/Comcast

[4] P. P. Bhattacharya and A. Ephremides, "Optimal scheduling with strict deadlines," *IEEE Trans. Autom. Control*, vol. 34, no. 7, pp. 721–728, Jul. 1989.

[5] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *J. Oper. Res. Soc.*, vol. 49, no. 3, pp. 237–252, Mar. 1998.

[6] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1277–1294, Jun. 2002.

[7] W. Rhee and J. M. Cioffi, "Increase in capacity of multiuser OFDM system using dynamic sub-channel allocation," in *Proc. IEEE 51st VTC-Spring*, May 2000, pp. 1085–1089.

[8] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[9] A. Sharifian, P. Djukic, H. Yanikomeroglu, and J. Zhang, "Resource Allocation Methods and Devices for Amplify-and-Forward Relay Network," US Patent US 8,477,679 B2, Jul. 2013. [Online]. Available: http://www.google.com/patents/US8477679

[10] A. Sharifian, P. Djukic, H. Yanikomeroglu, and J. Zhang, "Generalized proportionally fair scheduling for multiuser amplify-and-forward relay networks," in *Proc. IEEE 71st VTC-Spring*, May 2010, pp. 1–5.

[11] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of realtime and nonrealtime data in HDR," in *Proc. Int. Teletraffic Congr.*, Dec. 2001, pp. 793–804.

[12] N. Ruangchaijatupon and Y. Ji, "Integrated approach to proportional-fair resource allocation for multiclass services in an OFDMA system," in *Proc. IEEE GLOBECOM*, Nov. 2009, pp. 1–6.

[13] R. Schoenen and A. Otyakmaz, "QoS and flow management for future multi-hop mobile radio networks," in *Proc. IEEE 72nd VTC-Fall*, Sep. 2010, pp. 1–5.

[14] O. S. Shin and K. B. Lee, "Packet scheduling over a shared wireless link for heterogeneous classes of traffic," in *Proc. IEEE ICC*, Jun. 2004, pp. 58–62.

[15] H. Holma and A. Toskala, *WCDMA for UMTS*. London, U.K.: Wiley, 2002.

[16] D. Bültmann, T. Andre, and R. Schoenen, "Analysis of 3GPP LTE-Advanced cell spectral efficiency," in *Proc. IEEE Int. Symp. PIMRC*, Sep. 2010, pp. 1–5.

[17] G. Fettweis and S. Alamouti, "5G: Personal mobile Internet beyond what cellular did to telephony," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 140–145, Feb. 2014.

[18] Rohde and Schwarz, "Voice and SMS in LTE technology," Rohde and Schwarz, White Paper, 2011. [Online]. Available: rohde-schwarz.com

[19] A. Sharifian, "Utility-based packet scheduling and resource allocation algorithms with heterogeneous traffic for wireless OFDMA networks," Ph.D. dissertation, Carleton Univ., 2014.

[20] S. Shakkottai, T. Rappaport, and P. Karlsson, "Cross-layer design for wireless networks," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 74–80, Oct. 2003.

[21] S. Shakkottai and T. Rappaport, "Research challenges in wireless networks: A technical overview," in *Proc. Int. Symp. Wireless Pers. Multimedia Commun.*, Oct. 2002, pp. 12–18.

[22] S. Bilal, M. Ritesh, and S. Ashwin, "Downlink scheduling for multiclass traffic in LTE," *EURASIP J. Wireless Commun. Netw.*, vol. 2009, pp. 1:9–18:9, Nov. 2009.

[23] M. Andrews, "A Survey of Scheduling Theory in Wireless Data Networks," ser. The IMA Volumes in Mathematics and its Applications in *Wireless Communications*. New York, NY, USA: Springer-Verlag, 2007.

[24] N. Enderle and X. Lagrange, "User satisfaction models and scheduling algorithms for packet-switched services in UMTS," in *Proc. IEEE 57th VTC-Spring*, May 2003, pp. 1704–1709.

[25] L. Chen, B. Wang, X. Chen, X. Zhang, and D. Yang, "Utility-based resource allocation for mixed traffic in wireless networks," in *Proc. IEEE INFOCOM Workshop*, Apr. 2011, pp. 91–96.

[26] P. Svedman, S. K. Wilson, L. J. Cimini, and B. Ottersten, "Opportunistic beamforming and scheduling for OFDMA systems," *IEEE Trans. Commun.*, vol. 55, no. 5, pp. 941–952, May 2007.

[27] R. Agarwal, V. Majjigi, R. Vannithamby, and J. Cioffi, "Efficient scheduling for heterogeneous services in OFDMA downlink," in *Proc. IEEE GLOBECOM*, Nov. 2007, pp. 3235–3239.

[28] W.-H. Kuo and W. Liao, "Utility-based optimal resource allocation in wireless networks," in *Proc. IEEE GLOBECOM*, Dec. 2005, pp. 3508–3512.

[29] W.-H. Kuo and W. Liao, "Utility-based resource allocation in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3600–3606, Oct. 2007.

[30] G. Song and Y. Li, "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks," *IEEE Commun. Mag.*, vol. 43, no. 12, pp. 127–134, Dec. 2005.

[31] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks—Part I: Theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 614–624, Mar. 2005.

[32] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks—Part II: Algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625–634, Mar. 2005.

[33] H. Kushner and P. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE J. Wireless Commun.*, vol. 3, no. 4, pp. 1250–1259, Jul. 2004.

[34] R. Madan, S. Boyd, and S. Lall, "Fast algorithms for resource allocation in wireless cellular networks," *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 973–984, Jun. 2010.

[35] A. Bin Sediq, R. Gohary, and H. Yanikomeroglu, "Optimal tradeoff between efficiency and Jain's fairness index in resource allocation," in *Proc. IEEE Int. Symp. PIMRC*, Sep. 2012, pp. 577–583.

[36] A. Bin Sediq, R. Gohary, R. Schoenen, and H. Yanikomeroglu, "Optimal tradeoff between sum-rate efficiency and Jain's fairness index in resource allocation," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3496–3509, Jun. 2013.

[37] E. Rodrigues and F. Casadevall, "Adaptive radio resource allocation framework for multiuser OFDM," in *Proc. IEEE 69th VTC-Spring*, Apr. 2009, pp. 1–6.

[38] E. B. Rodrigues and F. Casadevall, "Rate adaptive resource allocation with fairness control for OFDMA networks," in *Proc. Eur. Wireless Conf.*, Apr. 2012, pp. 18–20.

[39] A. Pantelidou and A. Ephremides, "What is optimal scheduling in wireless networks," in *Proc. Annu. Int. Conf. Wireless Internet*, Nov. 2008, pp. 1–8.

[40] M. Andrews, "Instability of the proportional fair scheduling algorithm for HDR," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1422–1426, Sep. 2004.

[41] M. Andrews, L. Qian, and A. Stolyar, "Optimal utility based multiuser throughput allocation subject to throughput constraints," in *Proc. IEEE INFOCOM*, Mar. 2005, pp. 2415–2424.

[42] G. Song, Y. Li, J. Cimini, L. J., and H. Zheng, "Joint channel aware and queue-aware data scheduling in multiple shared wireless channels," in *Proc. IEEE WCNC*, Mar. 2004, pp. 1939–1944.

[43] G. Song, Y. Li, and L. J. Cimini, "Joint channel- and queue-aware scheduling for multiuser diversity in wireless OFDMA networks," *IEEE Trans. Commun.*, vol. 57, no. 7, pp. 2109–2121, July 2009.

[44] H. Hou, W. Zhou, S. Zhou, and J. Zhu, "Cross-layer resource allocation for heterogeneous traffic in multiuser OFDM based on a new QoS fairness criterion," in *Proc. IEEE 66th VTC-Fall*, Sep. 2007, pp. 1593–1597.

[45] S. Shakkottai and A. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *Amer. Math. Soc. Translations*, vol. 207, pp. 185–202, Dec. 2002.

[46] S. Shakkottai and A. Stolyar, "Optimal utility based multiuser throughput allocation subject to throughput constraints," in *Proc. Int. Teletraffic Congr.*, Sep. 2001, pp. 793–804.

[47] M. Andrews *et al.*, "Dynamic bandwidth allocation algorithms for high-speed data wireless networks," *Bell Labs Tech. J.*, vol. 3, no. 3, pp. 30–49, Jul. 1998.

[48] S. Ryu, B. Ryu, H. Seo, and M. Shin, "Urgency and efficiency based packet scheduling algorithm for OFDMA wireless system," in *Proc. IEEE ICC*, May 2005, pp. 2779–2785.

[49] S. Ryu, B. Ryu, H. Seo, and M. Shi, "Urgency and efficiency based wireless downlink packet scheduling algorithm in OFDMA system," in *Proc. IEEE 61st VTC-Spring*, Jun. 2005, pp. 1456–1462.

[50] S. Ryu, B. Ryu, and H. Seo, "Adaptive and QoS downlink multimedia packet scheduling for broadband wireless systems," in *Advances in Multimedia Information Processing*. New York, NY, USA: Springer-Verlag, 2005, pp. 417–428.

[51] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Utility-based adaptive radio resource allocation in OFDM wireless networks with traffic prioritization," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 66–71, Jan. 2009.

[52] M. Katoozian, K. Navaie, and H. Yanikomeroglu, "Optimal utility-based resource allocation for OFDM networks with multiple types of traffic," in *Proc. IEEE 67th VTC-Spring*, May 2008, pp. 2223–2227.

[53] A. Sharifian and H. Yanikomeroglu, "On the delay fairness through scheduling for wireless OFDMA networks," in *Proc. IEEE 73rd VTC-Spring*, May 2011, pp. 1–5.

[54] F. Kelly, "Fairness and stability of end-to-end congestion control," *Eur. J. Control*, vol. 9, no. 2, pp. 159–176, Apr. 2003.

[55] A. Sharifian, R. Schoenen, H. Yanikomeroglu, G. Senarath, H. T. Cheng, and P. Djukic, System and Method for Network Resource Allocation Considering Users Experience, Satisfaction, and Operators Interest," Huawei, ON, Canada, U.S. Patent US2014/0120974 A1, Feb. 2014. [Online]. Available: http://www.google.com/patents/US20140229210

[56] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE ICC*, Jun. 1995, pp. 331–335.

[57] S. Low and D. Lapsley, "Optimization flow control. I. Basic algorithm and convergence," *IEEE/ACM Trans. Netw.*, vol. 7, no. 6, pp. 861–874, Dec. 1999.

[58] A. Sharifian, P. Djukic, H. Yanikomeroglu, and J. Zhang, "Max-min fair resource allocation for multiuser amplify-and-forward relay networks," in *Proc. IEEE 72nd VTC-Fall*, Sep. 2010, pp. 1–5.

[59] E. B. Rodrigues, F. R. Lima, F. Casadevall, and F. R. P. Cavalcanti, "Capacity, fairness, and QoS tradeoffs in wireless networks with applications to LTE," in *Resource Allocation and MIMO for 4G and Beyond*. New York, NY, USA: Springer-Verlag, 2014, pp. 157–211.

[60] M. Andrews *et al.*, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.
[61] A. L. Stolyar and K. Ramanan, "Largest weighted delay first scheduling: Large deviations and optimality," *JSTOR Ann. Appl. Probability*, vol. 11, no. 1, pp. 1–48, Feb. 2001.
[62] M. Andrews *et al.*, "CDMA data QoS scheduling on the forward link with variable channel conditions," in *Bell Labs Techn. J.*, Boulogne-Billancourt, France: Bell Labs., Lucent Technol., 2000, pp. 1–45.
[63] B. Sadiq, S. J. Baek, and G. De Veciana, "Delay-optimal opportunistic scheduling and approximations: The log rule," *IEEE/ACM Trans. Netw.*, vol. 19, no. 2, pp. 405–418, Aug. 2011.
[64] G. Song, "Cross-layer resource allocation and scheduling in wireless multicarrier networks," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, USA, 2005.
[65] E. Rodrigues and F. Casadevall, "Control of the tradeoff between resource efficiency and user fairness in wireless networks using utility-based adaptive resource allocation," *IEEE Commun. Mag.*, vol. 49, no. 9, pp. 90–98, Sep. 2011.
[66] T. Schwarzfischer, "Quality and utility—Towards a generalization of deadline and anytime scheduling," in *Proc. Int. Conf. Autom. Planning Sched.*, Jun. 2003, pp. 277–286.
[67] H. Lei, M. Yu, A. Zhao, Y. Chang, and D. Yang, "Adaptive connection admission control algorithm for LTE systems," in *Proc. IEEE 67th VTC-Spring*, May 2008, pp. 2336–2340.
[68] M. Rahman and H. Yanikomeroglu, "Enhancing cell edge performance: A downlink dynamic interference avoidance scheme with inter-cell coordination," *IEEE Trans. Wireless Commun.*, vol. 9, no. 4, pp. 1414–1425, Apr. 2010.
[69] S. Patil and G. de Veciana, "Managing resources and quality of service in heterogeneous wireless systems exploiting opportunism," *IEEE/ACM Trans. Netw.*, vol. 15, no. 5, pp. 1046–1058, Oct. 2007.
[70] P. Liu, R. Berry, and M. Honig, "Delay-sensitive packet scheduling in wireless networks," in *Proc. IEEE WCNC*, Mar. 2003, pp. 1627–1632.
[71] L. Yang and M.-S. Alouini, "Performance analysis of multiuser selection diversity," *IEEE Trans. Veh. Technol.*, vol. 55, no. 6, pp. 1848–1861, Nov. 2006.
[72] L. Hentila, P. Kyasti, M. Koske, M. Narandzic, and M. Alatossava, "Matlab Implementation of the WINNER Phase II Channel Model ver1.1," Dec. 2007. [Online]. Available: http://projects.celtic-initiative.org/winner+/phase_2_model.html
[73] R. Schoenen, A. Bin Sediq, H. Yanikomeroglu, G. Senarath, C. Zhijun, and H. T. Cheng, "Spectral efficiency and fairness tradeoffs in cellular networks with realtime+nonrealtime traffic mix using stochastic Petri nets," in *Proc. IEEE 76th VTC-Fall*, Sep. 2012, pp. 1–5.
[74] H. T. Cheng and W. Zhuang, "An optimization framework for balancing throughput and fairness in wireless networks with QoS support," *IEEE Trans. Wireless Commun.*, vol. 7, no. 2, pp. 584–593, Feb. 2008.

Alireza Sharifian (M'14) was born in Isfahan, Iran. He received the B.Sc. degree in electronics and the M.Sc. degree in communications from Isfahan University of Technology, in 2005 and 2008, respectively, and the Ph.D. in electrical and computer engineering, from Carleton University, Ottawa, ON, Canada, in 2014, with a Senate medal for outstanding academic achievement.

Throughout his Ph.D. studies, he has been a member of a strategic collaborative research project with the Huawei Shenzhen and Huawei Ottawa R&D offices, supported by industry and the Ontario Research Fund—Research Excellence Grant from the Ontario Government. He is currently a Postdoctoral Fellow with the University of Toronto, Toronto, ON, Canada. His research interests include radio resource allocation, packet scheduling, quality of service, interference coordination, heterogeneous networks, relay networks, optimization, signal processing, orthogonal frequency division multiple access (OFDMA), Long-Term Evolution (LTE), LTE Advanced, and conversations on possible technologies for fifth generation wireless networks.

Rainer Schoenen (SM'13) received the Dipl.-Ing. and Ph.D. degrees in electrical engineering from RWTH Aachen University, Aachen, Germany, in 1995 and 2000, respectively.

His Ph.D. dissertation was entitled "System Components for Broadband Universal Networks with QoS Guarantee," which was written within the ISS Group of Prof. H. Meyr. In 2000, he became self-employed. From 2005 to 2009, he was a Senior Researcher with the Communication Networks Research Group, RWTH Aachen University, with Prof. Walke, working on computer networks, queuing theory, Petri nets, LTE-Advanced, frequency division duplex relaying, scheduling, OSI layer 2 (MAC), and IMT-Advanced Evaluation within WINNER+. From 2010 to 2014, he was a Project Manager with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada, supporting Prof. H. Yanikomeroglu. Since 2015, he has been a Professor with Hamburg University of Applied Sciences (HAW), Hamburg, Germany. His research interests include stochastic Petri nets and queuing systems, asynchronous transfer mode, Transmission Control Protocol/Internet Protocol, switching, flow control, quality of service, tariffs, user in the loop, wireless resource and packet scheduling, and the medium access control layer of 4+5G systems.

Halim Yanikomeroglu (S'96–M'98–SM'12) was born in Giresun, Turkey, in 1968. He received the B.Sc. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 1990 and the M.A.Sc. degree in electrical engineering (now ECE) and the Ph.D. degree in electrical and computer engineering from the University of Toronto, Toronto, ON, Canada, in 1992 and 1998, respectively.

During 1993–1994, he was with the R&D Group, Marconi Kominikasyon A.S., Ankara. Since 1998, he has been with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada, where he is currently a Full Professor. During 2011–2012, he was a Visiting Professor with TOBB University of Economics and Technology, Ankara. He has given a many tutorial lectures and invited talks on wireless technologies in leading international conferences. In recent years, his research has been funded by Huawei, Telus, Blackberry, Samsung, the Communications Research Centre Canada, and Nortel. This collaborative research has resulted in about 20 patents (granted and applied). His research interests include many aspects of wireless technologies, with a special emphasis on cellular networks.

Dr. Yanikomeroglu has been involved in the organization of the IEEE Wireless Communications and Networking Conference (WCNC) from its inception, including serving as a Steering Committee Member and the Technical Program Chair or the Cochair for WCNC 2004 (Atlanta, GA, USA), WCNC 2008 (Las Vegas, NV, USA), and WCNC 2014 (Istanbul, Turkey). He was the General Cochair for the IEEE Vehicular Technology Conference in the Fall of 2010, held in Ottawa. He has served on the editorial boards of the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and *IEEE Communications Surveys and Tutorials*. He was the Chair of the IEEE Technical Committee on Personal Communications (now called the Wireless Technical Committee). He is a Distinguished Lecturer of the IEEE Communications Society (2015–2016) and of the IEEE Vehicular Technology Society (2011–2015). He received the Carleton University Research Achievement Award in 2009, the Carleton University Graduate Students Association Excellence Award in Graduate Teaching in 2010, the Carleton University Faculty Graduate Mentoring Award in 2010, and the IEEE Ottawa Section Outstanding Educator Award in 2014. He is a Registered Professional Engineer in the Province of Ontario, Canada.