Received August 16, 2019, accepted August 29, 2019, date of publication September 2, 2019, date of current version September 18, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2939120*

# Matching-Based Resource Allocation for Critical MTC in Massive MIMO LTE Networks

**MOHAMMED Y. ABDELSADEK[1], (Student Member, IEEE),**
**YASSER GADALLAH[2], (Senior Member, IEEE),**
**AND MOHAMED H. AHMED[1], (Senior Member, IEEE)**
[1]Department of Electrical and Computer Engineering, Memorial University of Newfoundland, St. John's, NL A1B 3X5, Canada
[2]Department of Electronics and Communications Engineering, The American University in Cairo, Cairo 11835, Egypt

Corresponding author: Mohammed Y. Abdelsadek (mabdelsadek@mun.ca)

This work was supported by the Natural Sciences and Engineering Research Council (NSERC), Canada, under Discovery Grant 2014-03638.

**ABSTRACT** Supporting critical Machine-Type Communications (MTC) in addition to Human-Type Communications (HTC) is a major target for LTE networks to fulfill the 5G requirements. However, guaranteeing a stringent Quality-of-Service (QoS) for MTC, in terms of latency and reliability, while not sacrificing that of HTC is a challenging task from the radio resource management perspective. In this paper, we optimize the resource allocation process through exploiting the additional degrees of freedom introduced by massive Multiple-Input Multiple-Output (MIMO) techniques. We utilize the effective bandwidth and effective capacity concepts to provide statistical guarantees for the QoS, in terms of probability of delay-bound violation, of critical MTC in a cross-layer design manner. In addition, we employ the matching theory to solve the formulated combinatorial problem with much lower computational complexity compared to that of the global optimal solution so that the proposed scheme can be used in practice. In this regard, we analyze the computational complexity of the proposed algorithms and prove their convergence, stability and optimality. The results of extensive simulations that we performed show the ability of the proposed matching-based scheme to satisfy the strict QoS requirements of critical MTC with no impact on those of HTC. In addition, the results show a close-to-global optimal performance while outperforming other algorithms that belong to different scheduling strategies in terms of the adopted performance indicators.

**INDEX TERMS** Critical machine-type communications, ultra-reliable low-latency communications, massive MIMO LTE.

## I. INTRODUCTION

In order to accommodate all communicating elements to be connected to the network and form the Internet of Things (IoT), the evolution of the communication networks to incorporate Machine-Type Communications (MTC) in addition to Human-Type Communications (HTC) has become inevitable. MTC can be categorized into two major classes, massive MTC and critical MTC. The former is about connecting a massive number of low-complexity and low-cost devices such as sensors and wearables. It supports the IoT applications that require low data rate and latency-tolerant transmissions. On the other hand, critical MTC represent those types of communications that require very low latency, ultra-high reliability, and high network availability. Therefore, they are also known as Ultra-Reliable Low-Latency Communications (URLLC). Supporting such type of MTC opens the door to many applications such as traffic safety, industry automation, emergency and disaster response, e-health services, and many other yet-to-appear applications.

Among the different wireless technologies, cellular networks are considered one of the most convenient technologies to provide the connectivity of critical MTC devices (MTCDs). This is by virtue of their advanced Radio Resource Management (RRM) techniques and the availability of licensed spectrum that can guarantee the required stringent Quality of Service (QoS). Accordingly, the International Telecommunication Union (ITU) targets URLLC as a major use case, in addition to enhanced Mobile Broadband (eMBB) and massive MTC, in the requirements for the International Mobile Telecommunications 2020 and beyond (IMT-2020) [1]. The Third Generation Partnership Project (3GPP) is

The associate editor coordinating the review of this article and approving it for publication was Liang Yang.

working on evolving the current Long-Term Evolution (LTE) standard, in addition to the New Radio (NR), to fulfill the Fifth-Generation (5G) requirements with backward compatibility [2]. Therefore, several enhancements in the PHYsical (PHY) and Medium Access Control (MAC) layers have been introduced in 3GPP Releases 14 and 15 to support critical MTC in LTE [3]. For instance, the concept of short transmission time intervals and supporting reduced processing time are considered in [4], in addition to fast uplink access on MAC in [5], as techniques to reduce the latency in LTE to serve critical MTC efficiently.

Massive Multiple-Input Multiple-Output (MIMO) is considered as a major technology to improve the spectral efficiency, processing complexity, and energy efficiency of LTE systems to fulfill the 5G requirements. Therefore, 3GPP targets employing tens of antennas at the eNodeB (eNB) to utilize the massive MIMO techniques [6]. These MIMO enhancements in LTE are standardized under the official name of Full-Dimension MIMO (FD-MIMO) [7]. In this case, the additional degrees of freedom introduced by massive MIMO can be exploited to serve critical MTC efficiently [8]. As analyzed in [9], the spatial degrees of freedom created by massive MIMO enable several beneficial properties for critical MTC such as high signal-to-noise ratio (SNR) links, spatial division multiplexing, and quasi-deterministic links that are immune to fast fading. In this regard, the study in [10] investigates the feasibility of the massive antenna systems to fulfill the stringent requirements of critical MTC in the uplink direction, testing different multi-antenna schemes such as coherent and non-coherent receivers. On the other hand, the satisfaction of the requirements of critical MTC should be without sacrificing the QoS of the HTC traffic. This is due to the fact that the characteristics of critical MTC traffic is different than those of HTC in several aspects such as the data rate, the packet size, the latency-tolerance, and the reliability requirements. Therefore, and to achieve the goal of fulfilling the stringent QoS requirements of critical MTC without negative effects on HTC, RRM techniques should be optimized to serve both types of communications efficiently without degrading the system utility as well. Hence, in this paper, we optimize the resource allocation and scheduling process for critical MTC, considering the coexistence of the HTC traffic, through exploiting massive MIMO techniques.

### A. RELATED WORK

Several recent studies consider the resource allocation problem of critical MTC without considering the coexistence of HTC traffic. In [11], the authors propose a downlink scheduler for reliable low latency users. First, they subdivide the users into two groups, high and low priority, according to the possibility of satisfying their QoS requirements in terms of maximum delay and packet error rate. Therefore, they serve the users who have QoS requirements that can be satisfied in the scheduling period first. However, they consider a special case of channel status feedback, in which a wideband report is used for the whole bandwidth. The study in [12] maximizes the energy efficiency in the downlink of Frequency Division Multiple Access (FDMA) systems that serve URLLC while considering their end-to-end delay and packet loss requirements. This is achieved by optimizing the transmit power, bandwidth and the number of active antennas. They adopt a finite blocklength analysis to approximate the achievable data rates of the users. Nevertheless, they do not consider Orthogonal FDMA (OFDMA)-based systems such as LTE. In [13], the study maximizes the energy efficiency of URLLC in OFDMA-based radio access systems considering their QoS requirements of packet loss and latency. For this purpose, they optimize the packet dropping, power allocation, and bandwidth allocation policies. The authors in [14] extend the work in [12] and [13] by exploiting the multi-user diversity. However, they consider the downlink of FDMA-based cellular systems similar to [12]. In [15], the MTCDs are clustered based on their QoS characteristics, requirements and transmission protocols. Then, the aggregate data rate is maximized while considering the minimum data rate requirements of the devices. Nevertheless, separating the resource allocation processes for HTC and critical MTC, as discussed in the aforementioned works, does not optimize the overall resource allocation and reduces the gain resulting from multiuser diversity. Furthermore, this approach does not consider the impact of satisfying the stringent requirements of critical MTC on HTC traffic.

Therefore, studies consider the coexistence of MTC and HTC traffic types in the resource allocation problem. In [16], the authors consider splitting the radio resources between both types of users based on their buffer sizes. Then, every type of communication is scheduled separately. The resources are allocated fairly on the MTCDs considering their transmission deadlines. However, such splitting process of the radio resources before allocation does not optimize the allocation process at the system level. In [17], [18], the authors optimally maximize the aggregate data rate of the HTC traffic while considering the QoS requirements of all users. Nevertheless, they do not consider multiple antenna configurations that complicate the resource allocation process. This is due to the interference that can occur between users co-scheduled on the the same radio resources. Hence, the selection of the co-scheduled users should be taken into consideration while optimizing the resource allocation process. Moreover, they do not consider effective bandwidth and effective capacity concepts that can be used to provide statistical guarantees for the QoS of critical MTC as will be discussed in Section II-B.

### B. PAPER CONTRIBUTIONS AND ORGANIZATION

The major contributions of this paper can be summarized in the following:

- We formulate the resource allocation problem of critical MTC coexistent with HTC in massive MIMO LTE networks such that the system utility is maximized while satisfying the different QoS requirements of both types of communications. In this regard, we use the effective bandwidth [19] and effective capacity [20] concepts to

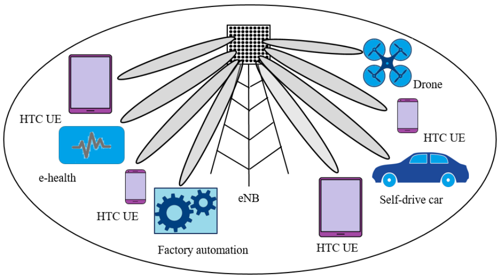


**FIGURE 1.** An eNB with massive antennas serving critical MTCDs coexistent with HTC UEs in an LTE cell.

design the resource allocation constraints from a cross-layer perspective to provide statistical guarantees for the QoS requirements of critical MTC in terms of probability of delay-bound violation. This considers both the PHY layer parameters and the buffer dynamics of the devices. Then, we formulate an equivalent instantaneous resource allocation problem exploiting the ergodicity of the service processes. However, an exponential computational complexity is required to calculate the global optimal solution of the formulated optimization problem that is NP-hard, as will be discussed in Section II.

- Therefore, we propose a computationally-efficient algorithm for the formulated resource allocation problem that can be implemented in practice. For this purpose, we utilize the matching theory [21] to formulate the resource allocation problem as a matching process that can be solved efficiently with much lower computational complexity compared to that of the global optimal solution. In this regard, we analyze the computational complexity of the proposed algorithms in big-O notation and discuss and prove the convergence and stability of the proposed matching processes. In addition, the optimality of the proposed resource allocation scheme is investigated. Moreover, we run extensive simulations to evaluate the performance of the proposed matching-based resource allocation technique and compare it with other algorithms from different scheduling techniques. The statistics of the major parameters impacting the computational complexity of the proposed algorithms are calculated.

The rest of the paper is organized as follows. In Section II, we discuss the adopted system model and formulate the resource allocation problem. The proposed matching-based resource allocation technique is presented in Section III. Then, in Section IV, the proposed scheme is analyzed form the practical and computational perspective. The simulation results are presented and discussed in Section V. Finally, the study is concluded in Section VI.

## II. SYSTEM MODEL AND PROBLEM FORMULATION

### A. SYSTEM MODEL AND GENERAL FORMULATION

We consider the resource allocation and scheduling of the uplink transmissions of single-antenna users in a single LTE cell that is served by a single eNB, as shown in Fig. 1.

**TABLE 1.** Frequently used symbols and notations.

| Symbol | Description |
|---|---|
| $\mathcal{K},\mathcal{U},\mathcal{H},\mathcal{M}$ | Sets of PRBs, users, HTC UEs, MTCDs, respectively |
| $K,U,H,M$ | Cardinalities of $\mathcal{K},\mathcal{U},\mathcal{H},\mathcal{M}$, respectively |
| $\mathcal{K}_u$ | Subset of PRBs assigned to user $u$ |
| $\mathcal{C}_k$ | Set of co-scheduled users on PRB $k$ |
| $A$ | Number of antennas at eNB |
| $y_k$ | Received signal on $k$th PRB |
| $h_{u,k}$ | Channel gain of $u$th user on $k$th PRB |
| $\gamma_{u,k}$ | SNR of $u$th user on $k$th PRB |
| $P_{u,k}$ | Transmit power of $u$th user on $k$th PRB |
| $N_k$ | Power spectral density of AWGN on $k$th PRB |
| $v_{u,k}$ | Beamforming vector of $u$th user on $k$th PRB |
| $T$ | Period of one TTI |
| $i$ | TTI index |
| $\lambda_u$ | Average arrival rate of $u$th user |
| $\mathcal{A}$ | Arrival process |
| $\mathcal{S}$ | Service process |
| $\bar{R}_u^{min}$ | Minimum average rate of the $u$th user, $u \in \mathcal{H}$ |
| $K_u^{max}$ | Maximum no. of PRBs can be assigned to $u$th user |
| $C_k^{max}$ | Maximum no. of users co-scheduled on $k$th PRB |
| $D_u$ | Delay of $u$th user |
| $D_u^{max}$ | Delay bound of $u$th user |
| $\varepsilon_u$ | Maximum allowed PDBV of $u$th user |
| $R_u$ | Achievable data rate of the $u$th user |
| $R_{u,k}$ | Achievable data rate of the $u$th user on $k$th PRB |
| $x_{u,k}$ | Indicator weather PRB $k$ is assigned to user $u$ or not |
| $\theta_u$ | QoS exponent of the $u$th user |
| $\Lambda_u$ | Effective bandwidth of $u$th user |
| $\kappa_u$ | Effective capacity of $u$th user |
| $\mu$ | Assignment operation of the matching process |
| $\varrho_{u,k}$ | Desirability between user $u$ and PRB $k$ |
| $\Psi$ | System utility function |

Assume that the set of users is indexed by $\mathcal{U} = \mathcal{H} \cup \mathcal{M} = \{1, \cdots, u, \cdots, U\}$, where $\mathcal{H}$ is a set of HTC UEs and $\mathcal{M}$ is a set of critical MTCDs. Suppose that the number of HTC UEs and critical MTCDs in the cell are $H$ and $M$, respectively. The system bandwidth is divided into Physical Resource Blocks (PRBs) of 180 KHz bandwidth that are indexed by $\mathcal{K} = \{1, \cdots, k, \cdots, K\}$. A user can use a PRB for uplink transmission for a time period known as the Transmission Time Interval (TTI). The frequently used symbols and notations are summarized in Table 1.

Assume that the eNB uses $A$ antennas, where $A \gg U$. Such a massive number of antennas is deployed to utilize beamforming at the eNB for the uplink reception. Therefore, a set, $\mathcal{C}_k$ of users can be co-scheduled on the same PRB $k$. That is, $\mathbf{y}_k \in \mathbb{C}^{A\times 1}$, which is the received signal vector at the eNB on the $k$th PRB, is calculated by

$$\mathbf{y}_k = \sum_{u\in\mathcal{C}_k} \mathbf{h}_{u,k}\sqrt{P_{u,k}}s_{u,k} + \mathbf{n}_k, \tag{1}$$

where $s_{u,k} \in \mathbb{C}$ is the data signal transmitted by the $u$th user on the $k$th PRB, which is normalized to unit power, $\mathbf{n}_k \in \mathbb{C}^{A\times 1}$ is the receiver AWGN noise vector on the $k$th PRB, which is a complex Gaussian vector with zero mean and covariance matrix of $N_k\mathbf{I}_A$, where $\mathbf{I}_A$ is the identity matrix of size $A$, and $P_{u,k}$ is the transmit power on the $k$th PRB by the $u$th user. The channel between the eNB and the $u$th user on the $k$th PRB is represented by $\mathbf{h}_{u,k} \in \mathbb{C}^{A\times 1}$

which is calculated by

$$\mathbf{h}_{u,k} = \sqrt{Z_u/L_u}\mathbf{f}_{u,k}, \tag{2}$$

where $L_u$ is the power path loss, $Z_u$ is the shadowing power gain, and $\mathbf{f}_{u,k}$ is the small-scale fading between the device and the eNB on the $k$th PRB, which is assumed to be independent and identically distributed complex Gaussian.

The received signal, $\mathbf{y}_k$, is multiplied by a unit-norm receive beamforming vector, $\mathbf{v}_{u,k} \in \mathbb{C}^{A\times 1}$, to spatially discriminate the signal sent by the $u$th user on the $k$th PRB from the interfering signals of other co-scheduled users on the same PRB, $\{u' \neq u : u' \in \mathcal{C}_k\}$. Therefore, the uplink SINR of the signal from the $u$th user on the $k$th PRB can be calculated by [22]

$$\gamma_{u,k} = \frac{\frac{P_{u,k}}{N_k}|\mathbf{h}_{u,k}^H \mathbf{v}_{u,k}|^2}{\sum_{\forall u' \neq u, u' \in \mathcal{C}_k} \frac{P_{u',k}}{N_k}|\mathbf{h}_{u',k}^H \mathbf{v}_{u,k}|^2 + \mathbf{v}_{u,k}^H \mathbf{I}_A \mathbf{v}_{u,k}}. \tag{3}$$

Consequently, the maximum achievable data rate of user $u$ over PRB $k$ is

$$R_{u,k} = B\log_2(1+\gamma_{u,k}), \tag{4}$$

where $B = 180$ KHz, is the bandwidth of one PRB.

Every TTI, the scheduler in the eNB assigns the PRBs to the users such that the system utility is maximized while satisfying the QoS requirements of the users in the cell. According to [1], achieving high data rates for critical MTC is of low importance since their transmissions are characterized by their low data rate [23] and small packet size [24]. However, satisfying their latency and reliability requirements is crucial. On the other hand, the QoS of HTC improves by increasing their data rates. Therefore, maximizing the data rate of all users in the cell impacts the resource utilization negatively. This is because maximizing the data rate of critical MTC does not improve their QoS, given that their latency requirements are satisfied. Nevertheless, this data rate is at the expense of that of the HTC UEs.

As a consequence, we formulate the resource allocation problem such that the aggregate data rate of the HTC traffic is maximized while considering the QoS requirements of all users as constraints. That is, the optimization problem of the resource allocation process is formulated as follows:

$$\max_{\{\mathcal{K}_1,\cdots,\mathcal{K}_U\}\in\mathcal{K}} \sum_{u\in\mathcal{H}} R_u \tag{5}$$

$$\text{s.t. } \mathbb{E}\{R_u\} \geq \bar{R}_u^{min}, \quad \forall u \in \mathcal{H} \tag{5a}$$

$$\Pr[D_u \geq D_u^{max}] \leq \varepsilon_u, \quad \forall u \in \mathcal{M} \tag{5b}$$

$$|\mathcal{K}_u| \leq K_u^{max}, \quad \forall u \in \mathcal{M} \tag{5c}$$

$$|\mathcal{C}_k| \leq C_k^{max}, \quad \forall k \in \mathcal{K}, \tag{5d}$$

where $\mathcal{K}_u \in \mathcal{K}$ is the subset of PRBs assigned to the $u$th user, $R_u$ is the maximum achievable data rate of user $u$ over the subset of PRBs assigned to it, and $\mathbb{E}\{R_u\}$ is its average rate. To guarantee a minimum average rate for each HTC user, constraint (5a) is used, where $\bar{R}_u^{min}$ is the required minimum average rate of user $u$. On the other hand, we use accurate statistical guarantees for the latency requirements of critical MTC. For this purpose, we ensure that the probability of delay bound violation (PDBV) of each critical MTCD is under a certain threshold $\varepsilon_u$ as in constraint (5b), where $D_u^{max}$ is the delay bound for the $u$th MTCD. Therefore, given that the packets that miss their deadlines are dropped, the parameter $\varepsilon_u$ represents one component of the reliability guarantees of MTCD $u$. Constraint (5c) is used to ensure a maximum number of allowed PRBs to be assigned to MTCDs. For example, in LTE Release 13, the number of PRBs that are assigned to MTCDs is limited to 6. Constraint (5d) is expressed to limit the number of co-scheduled users on PRBs as used in the framework of users pairing as in [25], for instance. In (5c) and (5d), $K_u^{max}$ is the maximum number of PRBs that can be assigned to MTCD $u$ and $C_k^{max}$ is the maximum number of co-scheduled users allowed on PRB $k$. As discussed in [26], non-contiguous resource allocations are allowed in the uplink of LTE-Advanced. This enhances the spectral efficiency as discussed in [27] thanks to using frequency-selective scheduling.

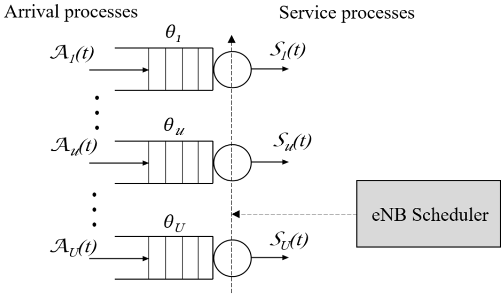


**FIGURE 2.** A cross-layer perspective of the eNB scheduler.

### B. CROSS-LAYER DESIGN AND FORMULATION

To provide statistical guarantees for the satisfaction of the latency requirements of critical MTC, a cross-layer design is required to consider their buffer dynamics as well as PHY layer parameters. For this purpose, we use the effective bandwidth and effective capacity concepts.

The resource allocation and scheduling process determines the data rate of every user in every TTI and controls the dynamics of the queues of the devices, as shown in Fig. 2. Let us define the arrival and service processes, in bits, of user $u$ as $\mathcal{A}_u(t)$ and $\mathcal{S}_u(t)$, respectively. According to the large deviations theory, the PDBV of the queue can accurately be approximated by [19]:

$$\Pr[D_u(t) \geq D_u^{max}] \approx e^{-\theta_u \delta_u D_u^{max}}, \tag{6}$$

where $\delta_u$ depends on both the arrival and service processes as will be discussed below and $\theta_u$ is known as the QoS exponent that characterizes the queue length decaying rate where a smaller $\theta_u$ represents a looser QoS constraint and vice versa.

The effective bandwidth [19] of the arrival process of user $u$ is defined as the minimum constant service rate that can serve that process with a guaranteed QoS exponent $\theta_u$ such that

$$\Pr[D_u(t) \geq D_u^{max}] \approx e^{-\theta_u \delta_u D_u^{max}} \leq \varepsilon_u, \tag{7}$$

and is calculated by

$$\Lambda_u(\theta_u) = \lim_{t\to\infty} \frac{1}{t\theta_u} \ln \mathbb{E}\{e^{\theta_u \mathcal{A}_u(t)}\}. \tag{8}$$

In a similar manner, the effective capacity [20] of the service process of the $u$th user is defined as the maximum constant arrival rate that can be served by the process with a guaranteed QoS exponent $\theta_u$, and is calculated by

$$\kappa_u(\theta_u) = -\lim_{t\to\infty} \frac{1}{t\theta_u} \ln \mathbb{E}\{e^{-\theta_u \mathcal{S}_u(t)}\}. \tag{9}$$

Therefore, the effective capacity of a wireless channel converges to the ergodic capacity when the QoS constraints are relaxed as discussed in [28].

The parameter $\delta_u$ can be calculated by deriving the rate at which the effective capacity and effective bandwidth curves intersect [29]. That is, $\delta_u = \kappa_u(\theta_u^*) = \Lambda(\theta_u^*)$. For instance, for a Poisson process, the parameter $\delta_u$ can be calculated as follows [30]:

$$\delta_u = \lambda_u \left( \frac{e^{\theta_u^*} - 1}{\theta_u^*} \right), \tag{10}$$

where $\lambda_u$ is the arrival rate of the Poisson process.

Accordingly, to guarantee a certain QoS exponent for an MTCD $u$, the effective capacity of the service process should satisfy the following inequality

$$\kappa_u(\theta_u) \geq \Lambda_u(\theta_u), \tag{11}$$

where the guaranteed QoS exponent, $\theta_u$, represents the required QoS level $(D_u^{max}, \varepsilon_u)$ and can be derived from (7) as

$$\theta_u = \frac{-\ln \varepsilon_u}{\delta_u D_u^{max}}. \tag{12}$$

To derive the effective capacity of the service process that represents the serviced bits at time $t$, we assume that the data rate of user $u$ at the $i$th TTI is $R_u[i]$. Therefore, the sequence $\{R_u[i]T : i = 1, 2, 3, \cdots\}$, where $T$ is the TTI period, is a discrete-time stationary and ergodic random process. Hence, the service process for the $u$th user is

$$\mathcal{S}_u[t] = \sum_{i=1}^{t} R_u[i]T. \tag{13}$$

Due to the fact that the sequence $\{R_u[i]T : i = 1, 2, 3, \cdots\}$ is uncorrelated, the effective capacity of the $u$th user in (9) reduces to [31]:

$$\kappa_u(\theta_u) = \frac{-1}{\theta_u} \ln \mathbb{E}\{e^{-\theta_u R_u[i]T}\}. \tag{14}$$

From the previous discussion, the PDBV constraint of critical MTCDs in (5b) can be expressed in a cross-layer perspective using (11) and (14). That is, the equivalent optimization problem to that in (5) is

$$\max_{\mathbf{X}} \sum_{u\in\mathcal{H}} \sum_{k=1}^{K} R_{u,k} x_{u,k} \tag{15}$$

$$\text{s.t. } \mathbb{E}\{R_u\} \geq \bar{R}_u^{min}, \quad \forall u \in \mathcal{H} \tag{15a}$$

$$\frac{-1}{\theta_u} \ln \mathbb{E}\{e^{-\theta_u R_u[i]T}\} \geq \Lambda_u, \quad \forall u \in \mathcal{M} \tag{15b}$$

$$\sum_{k=1}^{K} x_{u,k} \leq K_u^{max}, \quad \forall u \in \mathcal{M} \tag{15c}$$

$$\sum_{u=1}^{U} x_{u,k} \leq C_k^{max}, \quad \forall k \in \mathcal{K} \tag{15d}$$

$$x_{u,k} \in \{0, 1\}, \quad \forall u \in \mathcal{U},\ k \in \mathcal{K}, \tag{15e}$$

where $\mathbf{X}$ is a $U \times K$ binary indicator matrix such that $x_{u,k}$ indicates whether PRB $k$ is assigned to user $u$. Constraints (15a)–(15d) are equivalent to (5a)–(5d), respectively. Constraint (15e) is used to restrict $x_{u,k}$ to binary values.

The optimization problem in (15) falls in the Binary Nonlinear Programming (BNLP) category. This type of problems can be optimally solved using exhaustive search or algorithms such as the Branch and Bound (BB). However, the computational complexity of such algorithms is exponential which makes the problem NP-hard [32]. Therefore, these algorithms cannot be used in real-time processing such as in resource allocation and scheduling. Therefore, we propose computationally-efficient algorithms as a trade-off between the complexity and the performance so that they can be used in practice as resource allocation and scheduling schemes.

## III. MATCHING-BASED RESOURCE ALLOCATION

In this section, we formulate an instantaneous resource allocation problem that can be solved every TTI such that the long-term constraints (15a) and (15b) are satisfied. Then, utilizing the matching theory, we formulate the instantaneous problem as a two-sided matching process. Finally, we propose a complete matching-based resource allocation algorithm.

### A. FORMULATION OF THE INSTANTANEOUS RESOURCE ALLOCATION PROBLEM

Theorem 3.1 can be used to restrict the instantaneous data rates of the users such that their average data rate or PDBV constraints be satisfied in the long-term. That is, we derive data rate constraints equivalent to the constraints in (15a) and (15b) as follows.

*Theorem 3.1: The long-term constraints in (15a) and (15b) for the HTC and critical MTC, respectively, can be fulfilled if the following necessary and sufficient set of constraints is satisfied:*

$$R_u[i] \geq R_u^{min}[i], \quad \forall u \in \mathcal{U}, \tag{16}$$

*where, $R_u^{min}[i]$ is the instantaneous minimum data rate at the ith TTI for the uth user to fulfill its long-term constraint and is*

$$R_u^{min}[i] = \begin{cases} i\bar{R}_u^{min} - (i-1)R_u^{avg}[i-1], & \forall u \in \mathcal{H} \\ \dfrac{1}{\theta_u T} \ln\left(ie^{-\theta_u \Lambda_u} - (i-1)\Phi_u^{avg}[i-1]\right), & \forall u \in \mathcal{M} \end{cases}, \tag{17}$$

*calculated by (17), as shown at the top of the this page, where*

$$R_u^{avg}[i] = \begin{cases} \dfrac{R_u[i] + (i-1)R_u^{avg}[i-1]}{i}, & i \geq 2 \\ R_u[i], & i = 1, \end{cases} \tag{18}$$

$$\Phi_u^{avg}[i] = \begin{cases} \dfrac{\Phi_u[i] + (i-1)\Phi_u^{avg}[i-1]}{i}, & i \geq 2 \\ \Phi_u[i], & i = 1, \end{cases} \tag{19}$$

$$\Phi_u[i] = e^{-\theta_u R_u[i] T}. \tag{20}$$

*Proof:* To derive the minimum instantaneous rate of the set of HTC UEs, define $R_u^{avg}[i]$ as the cumulative moving average (CMA) of $R_u[i]$ at the $i$th TTI. This can be calculated using (18). This CMA represents the estimation of $\mathbb{E}\{R\}$, at the $i$th TTI, due to the ergodicity of the random process composed by the sequence $\{R_u[i] \; : \; i = 1, 2, 3, \cdots\}$. Therefore, the constraint in (15a) can be satisfied by fulfilling the following instantaneous constraint

$$R_u^{avg}[i] \geq \bar{R}_u^{min}, \; \forall u \in \mathcal{H}. \tag{21}$$

Using (18), we can write (21) as

$$R_u[i] \geq i\bar{R}_u^{min} - (i-1)R_u^{avg}[i-1], \; \forall u \in \mathcal{H}. \tag{22}$$

Therefore, the equivalent minimum instantaneous rate for the HTC UEs can be given by (17).

Similarly, to derive the minimum instantaneous data rate of the MTCDs, we define $\Phi_u[i]$ as in (20) and $\Phi_u^{avg}[i]$ as the CMA of $\Phi_u[i]$ at the $i$th TTI as given in (19). Similarly, this represents the estimation of $\mathbb{E}\{e^{-\theta_u R_u[i] T}\}$ since the random process $\{R_u[i]T \; : \; i = 1, 2, 3, \cdots\}$ is ergodic. Thus, the constraint in (15b) can be expressed as

$$\Phi_u^{avg}[i] \leq e^{-\theta_u \Lambda_u}. \tag{23}$$

Using (19), (23) can be rewritten in the following form

$$e^{-\theta_u R_u[i] T} \leq ie^{-\theta_u \Lambda_u} - (i-1)\Phi_u^{avg}[i-1]. \tag{24}$$

The last inequality can be written as in the form used in (16). Therefore, the minimum instantaneous data rate of the critical MTCDs can be derived as in (17). ∎

Using the equivalent set of constraints as in (16) in place of (15a) and (15b), we can derive an instantaneous resource allocation problem that is equivalent to (15) at the $i$th TTI as follows

$$\max_{\mathbf{X}} \sum_{u \in \mathcal{H}} \sum_{k=1}^{K} R_{u,k}[i] x_{u,k} \tag{25}$$

$$\text{s.t. } R_u[i] \geq R_u^{min}[i], \quad \forall u \in \mathcal{U} \tag{25a}$$

$$\sum_{k=1}^{K} x_{u,k} \leq K_u^{max}, \quad \forall u \in \mathcal{M} \tag{25b}$$

$$\sum_{u=1}^{U} x_{u,k} \leq C_k^{max}, \quad \forall k \in \mathcal{K} \tag{25c}$$

$$x_{u,k} \in \{0, 1\}, \quad \forall u \in \mathcal{U}, \; k \in \mathcal{K}. \tag{25d}$$

To solve the equivalent instantaneous problem in (25), we utilize the matching theory to devise a computationally-efficient algorithm.

### B. MATCHING MODEL AND FORMULATION

To formulate the resource allocation problem in (25) as a centralized matching process, we assume that $\mathcal{U}$ and $\mathcal{K}$ are two disjoint sets of agents that are willing to maximize their utilities and satisfy their minimum requirements. After the PRB assignment process is complete, we say that $(u, k)$ is a matched pair if PRB $k$ is assigned to user $u$. Therefore, a two-sided matching $\mu$ for the considered resource allocation problem in (25) can be defined as follows.

*Definition 3.2: A matching $\mu$ that is equivalent to the resource allocation problem in (25) is defined as a mapping from the set $\mathcal{U} \cup \mathcal{K}$ into the set $\mathcal{U} \cup \mathcal{K}$ such that for any $u \in \mathcal{U}$ and $k \in \mathcal{K}$:*

(i) $\mu(u) \subseteq \mathcal{K}$,

(ii) $\mu(k) \subseteq \mathcal{U}$,

(iii) $|\mu(k)| \leq C_k^{max}, \; \forall k \in \mathcal{K}$,

(iv) $|\mu(u)| \geq q_u^{min}, \; \forall u \in \mathcal{H}$,

(v) $q_u^{min} \leq |\mu(u)| \geq q_u^{max}, \; \forall u \in \mathcal{M}$,

(vi) $k \in \mu(u) \iff u \in \mu(k)$.

Condition (i) indicates that every user $u \in \mathcal{U}$ can be matched to a set of PRBs. Also, every PRB $k \in \mathcal{K}$ can be matched to a set of users as indicated in condition (ii). Therefore, this matching process falls in the many-to-many matching category. Condition (iii) represents the maximum number of co-scheduled users per PRB $k$. Condition (iv) represents the minimum rate requirement of the HTC UEs, where the minimum quota, $q_u^{min}$, is the cardinality of the set of PRBs that satisfy this constraint. Similarly, condition (v) is formulated for the minimum rate and the maximum number of PRBs constraints of the MTCDs. Condition (vi) indicates that if a PRB $k$ is matched to a user $u$, then it should be in its matched set of PRBs as well.

Due to the interference between users, the matching of every user $u$ to every PRB $k$ does not depend only on its channel conditions on this PRB. That is, every user $u$ cares about other users that are matched to the same PRBs. Therefore, similar to [33], we use a weighted, directed social network graph to model the relationship of every user to other users on every PRB to represent the interference between them as follows.

*Definition 3.3: The friendship network among users on every PRB $k$ is modeled as a weighted graph*

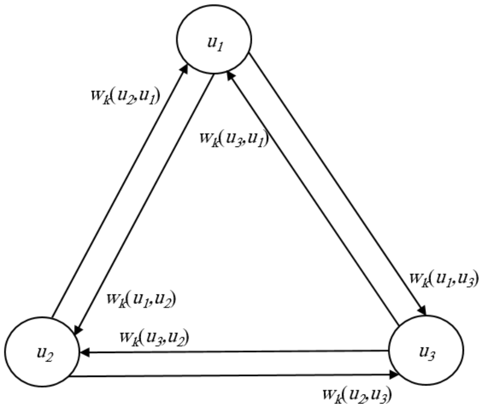


**FIGURE 3.** Weighted directed friendship network between users.

*$G = (\mathcal{N}, \Xi_k, w_k)$, where $\mathcal{N} = \mathcal{U}$ is the set of nodes, $\Xi_k$ is the set of arcs between them on PRB k, and $w_k$ are the weights that represent the relationship between users on the kth PRB, as shown in Fig. 3. The relationship between user u and $u'$ on PRB k is weighted by*

$$w_k(u, u') = \frac{P_{u',k}}{N_k} |\mathbf{h}^H_{u',k} \mathbf{v}_{u,k}|^2. \tag{26}$$

To define the utility of agents, we first define the desirability between user $u$ and PRB $k$, $\varrho_{u,k}$, as follows

$$\varrho_{u,k} = \frac{P_{u,k}}{N_k} |\mathbf{h}^H_{u,k} \mathbf{v}_{u,k}|^2. \tag{27}$$

Therefore, the utility of user $u$ on PRB $k$ depends on the desirability of user $u$ and PRB $k$, and the weight of the relationship between $u$ and other users co-scheduled on the same PRB, $\{u' \neq u : u' \in \mathcal{C}_k\}$. That is, the utility of user $u$ on PRB $k$, $\Psi_{u,k}$, can be calculated by

$$\Psi_{u,k} = B \log_2 \left( 1 + \frac{\varrho_{u,k}}{\sum_{\forall u' \neq u, u' \in \mathcal{C}_k} w_k(u, u') + \mathbf{v}^H_{u,k} \mathbf{I}_A \mathbf{v}_{u,k}} \right). \tag{28}$$

On the other hand, the utility of every PRB depends on the utilities of the HTC UEs scheduled on this PRB. Therefore, the utility of PRB $k$ can be calculated as

$$\Psi_k = \sum_{u \in \mu(k) \cap \mathcal{H}} \Psi_{u,k}. \tag{29}$$

Accordingly, to maximize the aggregate data rate of the HTC users, the matching assignment $\mu$ should maximize the system utility $\Psi$ that is defined as follows

$$\Psi = \sum_{k=1}^{K} \Psi_k, \tag{30}$$

subject to the conditions in Definition 3.2.

**Algorithm 1** Proposed Matching-Based Scheduling Algorithm

```
for all TTIs do
    Construct the instantaneous equivalent resource allocation problem as in (25) by calculating the minimum instantaneous data rate required for all $u \in \mathcal{U}$ at current TTI using (17).
    Formulate the instantaneous problem as a two-sided matching process by constructing the preference lists of all $u \in \mathcal{U}$ over $k \in \mathcal{K}$ and all $k \in \mathcal{K}$ over $u \in \mathcal{U}$ according to $\varrho_{u,k}$.
    Matching Phase 1
    Use Algorithm 2 to match the agents so that their minimum rate constraints are satisfied.
    Matching Phase 2
    Use Algorithm 3 to match the agents to maximize the data rate of the HTC users.
end for
```

### C. MATCHING-BASED RESOURCE ALLOCATION ALGORITHM

We now propose the resource allocation algorithm that is based on the matching process formulated in Section III-B. The matching process is a many-to-many assignment. However, two major challenges arise in this matching process. The first one is the lower quota bounds that are used for the minimum rate constraints. The second challenge is the externalities in the problem since the allocation of the PRBs to a certain user affects the other users that are co-scheduled on the same PRBs. To address these challenges, we perform the matching process in two phases, where each phase addresses one of the challenges. Algorithm 1 summarizes the proposed approach for solving the resource allocation problem in (15) and how the two phases of matching can be used to overcome the difficulty of the matching process.

In Algorithm 1, in every TTI, we construct the instantaneous resource allocation problem as in (25) and then derive a matching that solves it, as discussed in Section III-B. To establish the matching process, every agent $u \in \mathcal{U}$, or $k \in \mathcal{K}$, composes its preference list $\mathcal{P}(u)$, or $\mathcal{P}(k)$, respectively, in which the agents in the opposite set are ordered. Therefore, we say that PRB $k$ is preferred to $k'$ by user $u$ which is expressed as $k \succ_u k'$, if $k$ precedes $k'$ in $u$'s preference list, $\mathcal{P}(u)$. Similarly, if user $u$ precedes $u'$ in $k$'s preference list $\mathcal{P}(k)$, we say that $u \succ_k u'$. The ordering of the agents of the opposite set depends on the desirability between the two agents, $\varrho_{u,k}$, as in (27). That is,

$$k \succ_u k' \iff \varrho_{u,k} > \varrho_{u,k'}, \tag{31}$$

$$u \succ_k u' \iff \varrho_{u,k} > \varrho_{u',k}. \tag{32}$$

The preferences of the agents are transitive. That is, if $u \succ_k u'$ and $u' \succ_k u''$, then $u \succ_k u''$.

In Phase 1 of the matching, the users are matched such that their minimum instantaneous rate requirements are

**Algorithm 2** Satisfy Minimum Rate Constraints

```
Set the status of all $u \in \mathcal{U}$ that have minimum rate constraints to 1 and others to 0, where a status of 1 indicates that the user is willing to propose and otherwise is status 0.
while any user's status is 1 do
    Listen to the first user willing to propose. Assume it is $u^*$ and its preferred PRB is $k^*$ which is not in current $\mu(u^*)$.
    if $|\mu(k^*)| < C_k^{max}$ then
        Match $u^*$ to $k^*$ by updating their match lists $\mu(u^*)$ and $\mu(k^*)$ respectively.
        for all $u \in \mu(k^*)$ do
            Update its rate on the current PRB $k^*$ considering the interference of other co-scheduled users.
            if $R_u < R_u^{min}[i]$ and $|\mu(u)| < K_u^{max}$ then
                Set the status of $u$ to 1.
            else
                Set the status of $u$ to 1.
            end if
        end for
    else
        Let $k^*$ select the preferred set of users from the matched set and the candidate one $u^*$ according to its preference list.
        if the rejected user is the proposing one $u^*$ then
            Remove $k^*$ from the preference list of $u^*$ and clear its status if its preference list became empty.
        else
            Update the rate and status of the accepted set.
            Update the preference list, rate, and status of the rejected user.
        end if
    end if
end while
Feasibility test
if any $u \in \mathcal{U}$ still not satisfied then
    Problem is infeasible.
end if
```

**Algorithm 3** Maximize the HTC Data Rate

```
while There is still approved addition/deletion/swap do
    Step 1: Add users to improve the utility of PRBs
    for all $k \in \mathcal{K}$ do
        Determine the unmatched users and sort them according to the preference list of the PRB $k$, $\mathcal{P}(k)$.
        Consider the users in this candidate list in order, to be added to the current match $\mu(k)$. The approved user must yield a better utility of the PRB, $\Psi_k$, without violating the minimum rate requirements of the currently matched users, $\mu(k)$, in addition to the other conditions in Definition 3.2.
        Add the approved users to the current match.
        Update the rate of the new matched set of users.
        Update the utility of PRB $k$, $\Psi_k$.
    end for
    Step 2: Delete users to improve the utility of PRBs
    for all $k \in \mathcal{K}$ do
        Search in the matched users, $\mu(k)$, for the ones that can be unmatched to PRB $k$ such that the utility function of the PRB, $\Psi_k$, would improve without violating their minimum rate requirements.
        Unmatch the approved users from PRB $k$.
        Update the rate of the matched and rejected users on PRB $k$.
        Update the utility of PRB $k$, $\Psi_k$.
    end for
    Step 3: Swap users to improve the system utility
    for all $u \in \mathcal{U}$ do
        Search $\mathcal{U} \backslash \{u\}$ for an approved swap that can improve the system utility function, $\Psi$, without violating the minimum rate constraints of the users.
        Implement the approved swaps.
        Update the rate of the affected users.
        Update the utilities of affected PRBs.
    end for
end while
```

satisfied without considering the maximization of the aggregate data rate of the HTC UEs. For this purpose, we use Algorithm 2 that is based in principle on the one-to-many Gale-Shapley algorithm [34] after adapting it to the many-to-many problem and considering the externalities and lower quota bounds. Then, in Phase 2, the aggregate data rate of the HTC users is maximized by utilizing Algorithm 3. In Algorithm 3, the users are added, deleted, and swapped such that the system utility $\Psi$ is maximized without violating the minimum rate constraints that were satisfied in Phase 1. These three operations are similar to the swap-matching techniques studied for one-to-many problems in [33] to overcome the externalities in the problem. However, we use addition and deletion operations in addition to the swap operation in our many-to-many problem. Also, we consider the lower quota bounds in all operations.

## IV. ANALYSIS OF THE PROPOSED METHODS

In this section, we analyze the performance of the proposed resource allocation scheme from a practical perspective. For this purpose, we analyze the stability and convergence of the proposed matching algorithms. In addition, we discuss the optimality and computational complexity of the proposed scheme.

### A. STABILITY

The stability of the proposed resource allocation scheme in Algorithm 1 depends on that of the matching phase in Algorithm 3. This is because the other matching phase in Algorithm 2 is used mainly to satisfy the minimum instantaneous rate requirements of the users. To define the stability of

Algorithm 3, we first define the swap, addition, and deletion matchings as follows.

*Definition 4.1: Swap, $\mu^s_{u_1,u_2}$, addition, $\mu^a_{u,k}$, and deletion, $\mu^d_{u,k}$, matchings are defined respectively as follows:*

- $\mu^s_{u_1,u_2} = \{\mu \backslash \{(u_1, k_1), (u_2, k_2)\} \cup \{(u_1, k_2), (u_2, k_1)\}\}$, $k_1 \in \mu(u_1)$, $k_2 \in \mu(u_2)$,
- $\mu^a_{u,k} = \mu \cup (u, k)$, *and*
- $\mu^d_{u,k} = \mu \backslash (u, k)$.

Given the definition of swap, addition, and deletion matchings, the stability of the matching scheme in Algorithm 3 can be defined as follows.

*Definition 4.2: A matching $\mu$ is stable if and only if there are no $u$, $u'$, $k$ such that*

1) $\Psi(\mu^s_{u,u'}) > \Psi(\mu)$,
2) $\Psi(\mu^a_{u,k}) > \Psi(\mu)$, *or*
3) $\Psi(\mu^d_{u,k}) > \Psi(\mu)$.

*This is given that the matchings $\mu^s_{u,u'}$, $\mu^a_{u,k}$, and $\mu^d_{u,k}$ satisfy the minimum instantaneous rate requirements of all users in addition to the other conditions in Definition 3.2.*

The stability of Algorithm 3 is analyzed as follows.

*Lemma 4.3: If the matching scheme in Algorithm 3 converges to a matching $\mu^*$. Then, this matching $\mu^*$ is stable as defined in Definition 4.2.*

*Proof:* Assume that there are $u'$, $u''$, $k'$ that can yield $\Psi(\mu^a_{u',k'}) > \Psi(\mu)$, $\Psi(\mu^d_{u',k'}) > \Psi(\mu)$, or $\Psi(\mu^s_{u',u''}) > \Psi(\mu)$, and the new matchings satisfy the conditions in Definition 3.2. Then, this new matching would be approved in Step 1, 2, or 3, respectively, in Algorithm 3. This is because Steps 1, 2, and 3 in Algorithm 3 search for all approved addition, deletion, and swap operations, respectively, which improves the system utility $\Psi$ without violating the conditions in Definition 3.2. Accordingly, these $u$, $u'$, $k$ cannot exist given that the algorithm converged to a matching $\mu^*$. Consequently, the matching $\mu^*$ is stable. ■

## B. CONVERGENCE

The convergence of the proposed resource allocation scheme depends on that of the matching algorithms in Algorithm 2 and Algorithm 3. Therefore, in Theorem 4.4 we discuss the convergence of Algorithms 2 and 3 as follows.

*Theorem 4.4: The proposed matching schemes in Algorithm 2 and Algorithm 3 converge after a finite number of iterations.*

*Proof:* In Algorithm 2, every user $u$ proposes to its preferred PRB in its preference list, $\mathcal{P}(u)$, in order. If it is rejected by a PRB, it deletes it from its preference list and proposes to the next one until it satisfies its requirements, or its preference list becomes empty. Since the number of PRBs is limited, the preference list of every user $u$ is limited as well. Therefore, the number of proposals, and hence iterations, is limited. Consequently, Algorithm 2 converges after a finite number of iterations.

In Algorithm 3, after every approved addition, deletion, or swap operation, the new matching improves the system utility. That is, if the matching after every approved operation is as follows

$$\mu^{(1)},\ \mu^{(2)},\ \cdots,\mu^{(j-1)},\ \mu^{(j)},\ \cdots,\mu^{(final)}, \tag{33}$$

then $\Psi(\mu^{(j)}) > \Psi(\mu^{(j-1)})$. In other words, the system utility improves from every matching to the next. Due to the limited number of users and PRBs, the number of matchings is finite. In addition, the sum rate of the HTC UEs, which is the system utility, $\Psi(\mu)$, has an upper bound. Therefore, there is a round in which there is no further operation can be approved by the algorithm. Consequently, Algorithm 3 converges after a finite number of approved operations. ■

## C. OPTIMALITY

To analyze the optimality of the proposed resource allocation technique in Algorithm 1, we investigate how Algorithms 2 and 3 are used to get to a final solution for the problem in (25). As previously discussed, Algorithm 2 is mainly used to find a feasible solution that satisfies the minimum instantaneous rate requirements in addition to the remaining constraints in (25). However, Algorithm 3 is used to maximize the aggregate data rate of the HTC users, which is the objective function of (25), without violating the feasibility of the solution. Hence, the optimality of Algorithm 1 depends on that of Algorithm 3.

To analyze the optimality of Algorithm 3, we first discuss the relationship between the local maxima of the problem in (25) and the stability of the solution as a matching scheme as follows.

*Theorem 4.5: All local maxima of the objective function of the problem in (25) represent a stable matching as defined in Definition 4.2.*

*Proof:* Assume that a resource allocation pattern, that is represented by the matching $\mu^*$, is a local maximum to the optimization problem in (25). If $\mu^*$ is not a stable matching, then, according to Definition 4.2, there is at least one addition, deletion, or swap operation that can yield a better matching that has a better system utility function $\Psi(\mu)$. Since the system utility function $\Psi$ is the same as the objective function of the problem in (25), this contradicts the assumption that $\mu^*$ is a local maximum. Therefore, $\mu^*$ must be a stable matching. ■

Consequently, the optimality of Algorithm 1 can be proved as in the following lemma.

*Lemma 4.6: The matching-based resource allocation scheme in Algorithm 1 yields a local optimal solution for the optimization problem in (25).*

*Proof:* This is a direct the result of Theorem 4.5 and the stability proof of Algorithm 1 that is based on Lemma 4.3. ■

## D. COMPUTATIONAL COMPLEXITY

To analyze the computational complexity of the proposed resource allocation scheme in Algorithm 1, we calculate the worst case computational complexity of every step in Algorithm 1 in terms of big-O notation. For this purpose, we first analyze the computational complexity of the steps of Algorithm 2 and 3.

The worst case computational complexity of the steps of Algorithm 2 can be summarized as follows:

- step 1 requires $\mathcal{O}(U)$,
- steps 2-23 require $\mathcal{O}(UKC_k^{max})$, and
- steps 24-26 require $\mathcal{O}(U)$.

The steps of Algorithm 3 have the following computational complexity:

- steps 2-8 require $\mathcal{O}(K(min(H, C_k^{max}))^2)$,
- steps 9-14 require $\mathcal{O}(KC_k^{max})$, and
- steps 15-20 require $\mathcal{O}(U(U-1)C_k^{max}K_u^{max})$.

Accordingly, we can analyze the worst case computational complexity of every step of Algorithm 1 as follows:

- step 2 requires $\mathcal{O}(U)$,
- step 3 requires $\mathcal{O}(UK^2) + \mathcal{O}(KU^2)$,
- step 4 (Algorithm 2) requires $\mathcal{O}(KUC_k^{max})$, and
- step 5 (Algorithm 3) requires $\mathcal{O}(r(U^2 K_u^{max} C_k^{max}))$,

where $r$ is the number of rounds implemented in Algorithm 3. Numerical evaluations for this parameter are presented in Section V.

Therefore, step 5 dominates the total complexity of the proposed resource allocation scheme. In fact, this is the computational complexity of the swap operations that are used to maximize the aggregate data rate of the HTC users. However, the complexity of the algorithm is still much lower than that of the global optimal solution. This is because, as mentioned above, the computational complexity of the global optimal solution of BNLP problem is exponential, which makes the problem NP-hard [32].

## V. EXPERIMENTAL RESULTS

In this section, we present and discuss the results of the simulation experiments performed to evaluate the performance of the proposed matching-based resource allocation scheme. We compare the performance of the proposed algorithms with that of the global optimal allocation, the solution calculated by the Genetic Algorithm (GA), and the Proportional Fairness (PF) scheduler for multi-user MIMO systems as in [36]. In addition, we evaluate the computational complexity of the proposed algorithm by discussing the statistics of the major parameters that affect the complexity.

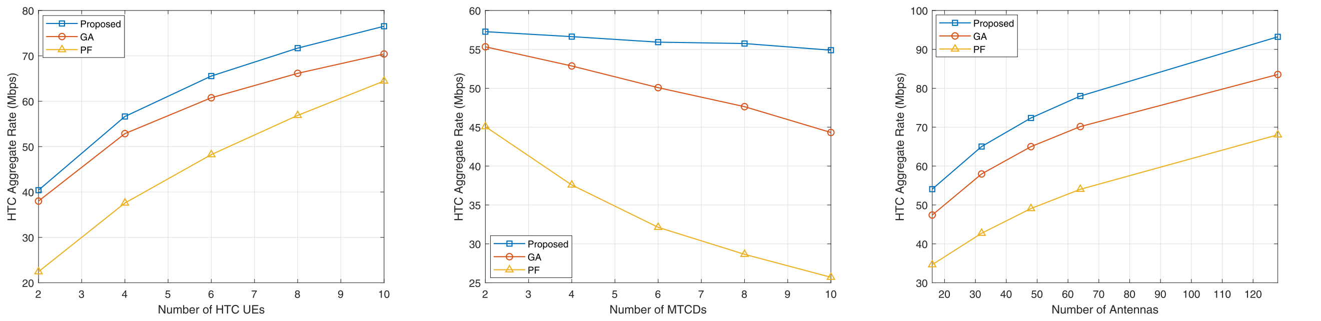
In the simulations, we uniformly distribute a set of single-antenna HTC and critical MTC users in a single LTE cell with a radius of 500 m. The users are served by a single eNB that contains a massive number of antennas which are used to simultaneously schedule more than one user on the same PRB. Without loss of generality, we use maximal ratio combining (MRC) receive beamforming vectors in the simulations. The users generate uplink transmissions with Poisson arrivals with average arrival rate uniformly picked from the sets as in Table 2, which summarizes the simulation parameters. As discussed in Section II, we assume that the HTC UEs have minimum average rate requirements and the critical MTCDs have minimum PDBV requirements. Therefore, the aggregate achievable data rate of the HTC UEs and the average PDBV of the MTCDs in the cell are the metrics used to evaluate the performance of the proposed resource allocation scheme. The confidence interval of the estimation of the HTC aggregate rate ranges from 0.2586 Mbps to 0.7083 Mbps with an average of 0.4544 Mbps. For the estimation of the average PDBV of the MTCDs, the confidence interval varies from $1.56 \times 10^{-4}$ to $2.70 \times 10^{-3}$ with an average of $1.12 \times 10^{-3}$.

**TABLE 2. Simulation parameters.**

| Parameter | Value |
|---|---|
| Cell radius ($C$) | 500 m |
| Number of eNBs | 1 |
| Simulation time | 200 TTI |
| Number of runs | 10 |
| Path loss ($PL0 + 10n \log_{10} d$) | $128.1 + 37.6 \log_{10}(d)$,<br>$d$ in km [35] |
| Standard deviation of shadowing ($\sigma$) | 8 dB |
| Transmitter power ($P$) | 15 dBm |
| Power spectral density of noise | $-174$ dBm/Hz |
| Noise figure | 18 dB |
| Number of PRBs ($K$) | 25 |
| Distribution of MTCDs/UEs | Fixed and uniform |
| HTC arrival rate ($\lambda_u,\ u \in \mathcal{H}$) | $64, 128, 192, 256$ kbps |
| MTC arrival rate ($\lambda_u,\ u \in \mathcal{M}$) | $10, 20, 30, 40$ kbps |
| Arrival rates distribution | Uniform |
| Delay bound ($D_u^{max},\ u \in \mathcal{M}$) | 0.2 ms |
| Maximum PDBV ($\varepsilon_u,\ u \in \mathcal{M}$) | $10^{-2}$ |
| $\bar{R}_u^{min},\ u \in \mathcal{H}$ | $\lambda_u$ |

Fig. 4 shows the aggregate achievable data rate of the HTC UEs in the cell using the proposed matching-based, the GA-based, and the PF schedulers. Increasing the number of HTC UEs or antennas allows the scheduler to co-schedule more HTC UEs on the same PRB. This results in an improvement in the HTC sum-rate as shown in Figs. 4a and 4c. On the other hand, scheduling more MTCDs in the cell degrades the HTC sum-rate since fulfilling their QoS requirements come at the expense of the HTC data rate, as Fig. 4b reveals. In all cases, the matching-based resource allocation achieves better aggregate HTC data rate compared to the other schedulers. This is because the PF scheduler allocates the PRBs in a fair manner to all users by maximizing their data rate based on their average throughput. Nevertheless, maximizing the data rate of the MTCDs after satisfying their QoS requirements is inefficient and impacts that of the HTC as discussed in Section II. On the other side, both the matching-based and the GA-based schemes maximize the data rate of the HTC UEs while satisfying the QoS requirements of all users. The GA yields a local maximum to the optimization problem but with lower objective value than the matching-based algorithm.

To show how close the solution of the matching-based algorithm is to the global optimal solution, we compare the HTC sum-rate, which is the objective function of the optimization problem, with that of the global maximum. For this purpose, we use the BARON solver [37] to solve the optimization problem in every TTI, i.e., the problem in (25). BARON adopts a polyhedral branch-and-cut approach to calculate the global optimal solution of the handled optimization problem [37]. Due to the exponential computational complexity of calculating the global optimal solution of such a

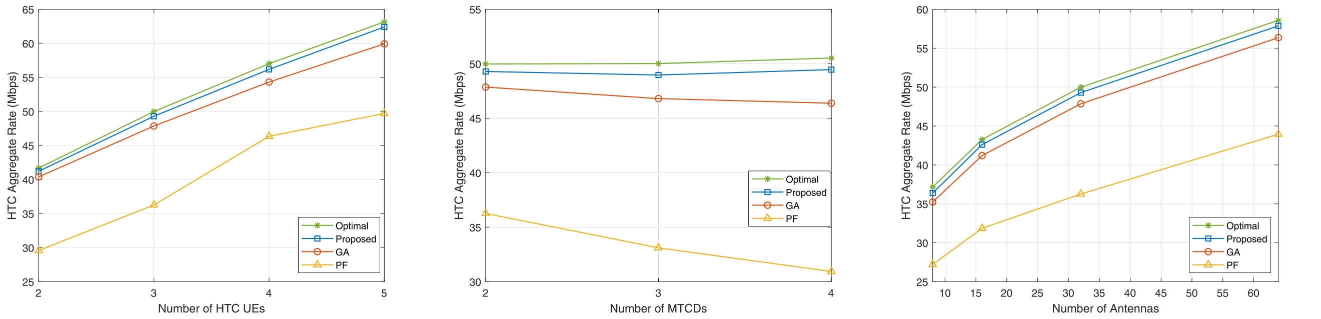


**(a)**H2H sum-rate versus number of HTC UEs ($A = 32,\ M = 4$).

**(b)**HTC sum-rate versus number of MTCDs ($A = 32,\ H = 4$).

**(c)**HTC sum-rate versus number of antennas ($H = 6,\ M = 6$).

**FIGURE 4.** Aggregate HTC achievable data rate.

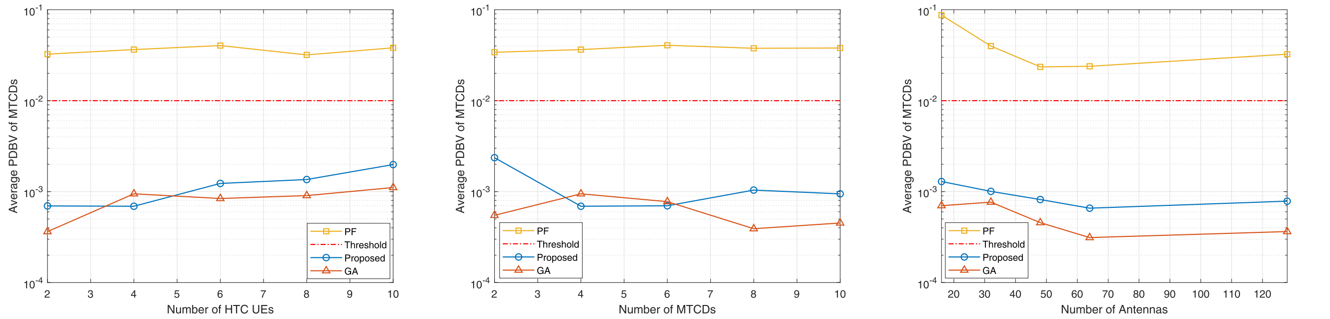


**(a)**Aggregate HTC data rate versus number of HTC UEs ($A = 32,\ M = 2$).

**(b)**Aggregate HTC data rate versus number of MTCDs ($A = 32,\ H = 3$).

**(c)**Aggregate HTC data rate versus number of antennas ($H = 3,\ M = 2$).

**FIGURE 5.** Comparison with the global optimal solution (3 runs and 100 TTIs).



**(a)**Average PDBV versus number of HTC UEs ($A = 32,\ M = 4$).

**(b)**Average PDBV versus number of MTCDs ($A = 32,\ H = 4$).

**(c)**Average PDBV versus number of antennas ($H = 6,\ M = 6$).

**FIGURE 6.** Average PDBV of MTCDs in the cell.

problem, we run the simulation on a small-size problem as demonstrated in Fig. 5. The figure shows the aggregate data rate of the HTC UEs in the cell versus the number of HTC UEs, the MTCDs, and the antennas. As the figure reveals, the sum-rate achieved by utilizing the matching-based algorithm is close to the global optimal rate and always better than that of the GA-based algorithm, as discussed before.

The satisfaction of the QoS requirements of the critical MTC is demonstrated in Fig. 6 which shows the average PDBV of the MTCDs in the cell versus the number of the HTC UEs, MTCDs, and antennas for the scheduling algorithms. As expected, both the matching-based and the GA-based algorithms satisfy the required level of QoS in all cases. This is due to the fact that any feasible solution to an optimization problem must satisfy its constraints and the constraints of the problem in (15) are formulated to fulfill the QoS requirements of the MTCDs. This fulfillment of the constraints could be with equality or as an inequality based on what maximizes the objective function. However, the PF scheduler targets a fair allocation on all users without

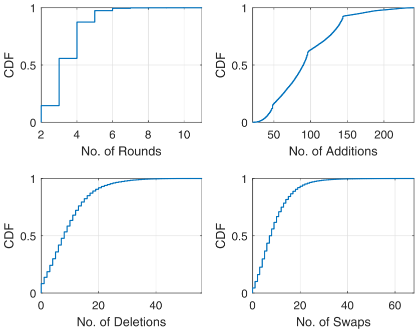


**FIGURE 7. CDF of the major parameters of the matching algorithm based on 28, 315 samples.**

considering latency requirements. Consequently, the stringent latency requirements are violated.

In addition to analyzing the computational complexity of the proposed matching-based resource allocation scheme in big-O notation, as discussed in Section IV-D, we calculate statistics of the major parameters affecting the complexity using simulations. For this purpose, we calculate the cumulative distribution function (CDF) of the number of rounds, additions, deletions, and swaps performed in Algorithm 3. This is because these parameters mainly determine the complexity of Algorithm 3 that represents the major component of the complexity of the proposed scheme. Fig. 7 shows the CDF of the parameters after executing the matching-based scheme 28, 315 times during the simulation using different combinations of numbers of users and antennas. As the figure reveals, the maximum number of the rounds, additions, deletions, and swaps was 11, 241, 56, 68, respectively. This shows the order of those parameters and the reduced computational complexity of the proposed scheme compared to the global optimal solution that has an exponential complexity.

## VI. CONCLUSION

In this paper, we utilized the effective bandwidth and effective capacity theories to formulate a cross-layer resource allocation problem for critical MTC coexistent with HTC in LTE networks with massive MIMO deployments. Then, we employed the matching theory to solve the formulated problem with much lower complexity compared to that of the global optimal solution. Therefore, the proposed matching-based resource allocation scheme can be used in practice in LTE networks. To this end, we analyzed the computational complexity, the convergence, the stability, and the optimality of the proposed algorithms. The analysis showed that the proposed scheme converges to a local optimal allocation in a polynomial time. Extensive simulations proved the efficiency of the proposed scheme in satisfying the different types of QoS of both types of communications (HTC and critical MTC) while maximizing the system utility. The results revealed the superiority of the matching-based resource allocation compared to other algorithms of different scheduling strategies while achieving a close-to-global optimal performance. Moreover, the statistics of the major parameters that impact the computational complexity of the proposed algorithms showed the feasibility of applying the proposed scheme in practice.

## REFERENCES

[1] *IMT Vision—Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, document ITU-R M.2083, Sep. 2015.

[2] C. Hoymann, D. Astely, M. Stattin, G. Wikstrom, J.-F. Cheng, A. Hoglund, M. Frenne, R. Blasco, J. Huschke, and F. Gunnarsson, "LTE release 14 outlook," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 44–49, Jun. 2016.

[3] J. C. S. Arenas, T. Dudda, and L. Falconetti, "Ultra-low latency in next generation LTE radio access," in *Proc. 11th Int. ITG Conf. Syst., Commun. Coding (SCC)*, Feb. 2017, pp. 1–6.

[4] *Work Item on Shortened TTI and Processing Time for LTE*, document RP-161299, 3GPP, Jun. 2016.

[5] *Work Item on L2 Latency Reduction Techniques for LTE*, document RP-160667, 3GPP, Mar. 2016.

[6] Y. Kim, H. Ji, J. Lee, Y.-H. Nam, B. L. Ng, I. Tzanidis, Y. Li, and J. Zhang, "Full dimension MIMO (FD-MIMO): The next evolution of MIMO in LTE systems," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 26–33, Apr. 2014.

[7] H. Ji, Y. Kim, J. Lee, E. Onggosanusi, Y. Nam, J. Zhang, B. Lee, and B. Shim, "Overview of full-dimension MIMO in LTE-advanced pro," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 176–184, Feb. 2017.

[8] P. Popovski, J. J. Nielsen, C. Stefanovic, E. De Carvalho, E. Strom, K. F. Trillingsgaard, A.-S. Bana, D. M. Kim, R. Kotaba, and J. Park, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Netw.*, vol. 32, no. 2, pp. 16–23, Mar./Apr. 2018.

[9] P. Popovski, C. Stefanovic, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.

[10] S. R. Panigrahi, N. Bjorsell, and M. Bengtsson, "Feasibility of large antenna arrays towards low latency ultra reliable communication," in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2017, pp. 1289–1294.

[11] E. Khorov, A. Krasilov, and A. Malyshev, "Reliable low latency communications in LTE networks," in *Proc. IEEE Int. Black Sea Conf. Commun. Netw. (BlackSeaCom)*, Jun. 2017, pp. 1–5.

[12] C. Sun, C. She, and C. Yang, "Energy-efficient resource allocation for ultra-reliable and low-latency communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[13] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.

[14] C. Sun, C. She, and C. Yang, "Exploiting multi-user diversity for ultra-reliable and low-latency communications," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–6.

[15] F. Ghavimi, Y.-W. Lu, and H.-H. Chen, "Uplink scheduling and power allocation for M2M communications in SC-FDMA-based LTE-A networks with QoS guarantees," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6160–6170, Jul. 2017.

[16] A. M. Maia, D. Vieira, M. F. de Castro, and Y. Ghamri-Doudane, "A fair QoS-aware dynamic LTE scheduler for machine-to-machine communication," *Comput. Commun.*, vol. 89, pp. 75–86, Sep. 2016.

[17] M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "An LTE-based optimal resource allocation scheme for delay-sensitive M2M deployments coexistent with H2H users," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2017, pp. 139–144.

[18] M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "Optimal cross-layer resource allocation for critical MTC traffic in mixed LTE networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5944–5956, Jun. 2019.

[19] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[20] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[21] M. David, *Algorithmics of Matching Under Preferences*, vol. 2. Singapore: World Scientific, 2013.

[22] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]," *IEEE Signal Process. Mag.*, vol. 31, no. 4, pp. 142–148, Jul. 2014.

[23] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. 1st Int. Conf. 5G Ubiquitous Connectivity*, Nov. 2014, pp. 146–151.

[24] *Study on Scenarios and Requirements for Next Generation Access Technologies, Technical Specification Group Radio Access Network*, document TR 38.913, 3GPP, Oct. 2016.

[25] A. Mehbodniya, W. Peng, and F. Adachi, "An adaptive multiuser scheduling and chunk allocation algorithm for uplink SIMO SC-FDMA," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 2861–2866.

[26] N. Abu-Ali, A. M. Taha, M. Salah, and H. Hassanein, "Uplink scheduling in LTE and LTE-advanced: Tutorial, survey and evaluation framework," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1239–1265, Dec. 2014.

[27] *PUSCH Resource Allocation for Clustered DFT-Spread OFDM*, document R1-101211, 3GPP, Feb. 2010.

[28] L. Liu and J.-F. Chamberland, "On the effective capacities of multiple-antenna Gaussian channels," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 2583–2587.

[29] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2318–2328, Jun. 2008.

[30] F. Kelly, S. Zachary, and I. Ziedins, *Stochastic Networks: Theory and Applications*. London, U.K.: Oxford Univ. Press, 1996.

[31] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.

[32] C. H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*. Chelmsford, MA, USA: Courier, 1998.

[33] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *Proc. Int. Symp. Algorithmic Game Theory*. Berlin, Germany: Springer, 2011, pp. 117–129.

[34] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *Amer. Math. Monthly*, vol. 69, no. 1, pp. 9–15, Jan. 1962.

[35] *Further Advancements for E-UTRA Physical Layer Aspects*, document TR 36.814, 3GPP, 2010.

[36] L. Liu, Y.-H. Nam, and J. Zhang, "Proportional fair scheduling for multi-cell multi-user MIMO systems," in *Proc. 44th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2010, pp. 1–6.

[37] M. Tawarmalani and N. V. Sahinidis, "A polyhedral branch-and-cut approach to global optimization," *Math. Program.*, vol. 103, no. 2, pp. 225–249, 2005.

**MOHAMMED Y. ABDELSADEK** (S'17) received the B.Sc. and M.Sc. degrees in electrical engineering from Assiut University, Asyut, Egypt, in 2011 and 2014, respectively. He is currently pursuing the Ph.D. degree with the Memorial University of Newfoundland, St. John's, NL, Canada. His research interests include machine-type communications, radio resource management in cellular networks, and cognitive radio networks.

**YASSER GADALLAH** (S'91–M'05–SM'09) is currently the Chair of the Department of Electronics and Communications Engineering, The American University in Cairo (AUC). Before he joined the AUC, he was an Associate Professor with the Electronics and Communications Department, Misr International University, Cairo, Egypt. Earlier, he was an Assistant Professor of networking systems with the College of Information Technology, UAE University. Prior to joining the UAE University, He was a Research Scientist in the area of wireless sensor networks with the Communications Research Centre, Ottawa, ON, Canada. He has also worked for many years in the networking and telecommunications industry. He was as a Software Development Manager with Cisco Systems for several years, where he was involved in numerous research and development projects, which resulted in the introduction of several high-profile Cisco products. He was also with the Nortel Networks as a Software Project Leader and prior to that as a Software Designer. He has authored/coauthored several technical and scientific publications in the networking, communications, and IT fields. His current research interests include the Internet of Things issues, machine-to-machine communications, wireless sensor networks, mobile ad hoc networks, and broadband wireless access networks. He has served as an Organizer, a Technical Program Committee Member, and a Reviewer for several international conferences and journals.

**MOHAMED H. AHMED** received the Ph.D. degree in electrical engineering from Carleton University, Ottawa, ON, Canada, in 2001. From 2001 to 2003, he was a Senior Research Associate with Carleton University. In 2003, he joined the Faculty of Engineering and Applied Science, Memorial University, where he is currently a Full Professor. He authored/coauthored more than 140 articles in international journals and conferences. His research interests include radio resource management in wireless networks, multi-hop relaying, cooperative communication, vehicular ad hoc networks, cognitive radio networks, and wireless sensor networks. His research is sponsored by NSERC, CFI, QNRF, Bell/Aliant, and other governmental and industrial agencies. He was a recipient of the Ontario Graduate Scholarship for Science and Technology, in 1997, the Ontario Graduate Scholarship, from 1998 to 2000, and the Communication and Information Technology Ontario Graduate Award, in 2000. He served as the Co-Chair for the Signal Processing Track in ISSPIT14, the Transmission Technologies Track in VTC10-Fall, and the Multimedia and Signal Processing Symposium in CCECE09. He has served as a Guest Editor for the Special Issue on Fairness of Radio Resource Allocation, JWCN (EURASIP), in 2009, and the Special Issue on Radio Resource Management in Wireless Internet, *Wireless Communications and Mobile Computing* (Wiley), in 2003. He serves as an Editor for the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS and an Associate Editor for the *International Journal of Communication Systems* (Wiley) and *Wireless Communications & Mobile Computing* (Wiley). He is also a Registered Professional Engineer (P.Eng.) in the province of Newfoundland, Canada.