IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 21, 2020, accepted September 24, 2020, date of publication September 29, 2020, date of current version October 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027626

Performance Comparison of QoS Deployment Strategies for Cellular Network Services

ENGIN ZEYDAN[1], (Senior Member, IEEE), JOSEP MANGUES-BAFALLUY[1], OMER DEDEOGLU[2], AND YEKTA TURK[3]

[1]Centre Tecnològic de Telecomunicacions de Catalunya, 08860 Barcelona, Spain
[2]Radio Network Planning Department, Türk Telekomünikasyon A. S., 34889 İstanbul, Turkey
[3]Mobile Network Architect-Based in Istanbul, 34906 İstanbul, Turkey

Corresponding author: Engin Zeydan (engin.zeydan@cttc.cat)

This work was supported in part by the Spanish MINECO under Grant TEC2017-88373-R (5G-REFINE), and in part by the Generalitat de Catalunya under Grant 2017 SGR 1195.

ABSTRACT Differentiated quality-of-Service (QoS) techniques are widely used to distinguish between different service classes and prioritize service needs in mobile networks. Mobile Network Operators (MNOs) utilize QoS techniques to develop strategies that are supported by the mobile network infrastructure. However, QoS deployment strategy can ensure that the radio resources provided by the base station are easily consumed if it is not used correctly or when different techniques are used all together. In this paper, we propose a scheduling algorithm and compare two different QoS deployment strategies for prioritized User Equipment (UEs) (with higher scheduling rates and dedicated bandwidth) that MNOs can use in the current infrastructure, using a commercial real-time Long Term Evolution (LTE) network in different test scenarios. Moreover, we expose the real-time user experience in terms of uplink throughput and analyze results of the UE's real-time key performance indicators (KPIs) in detail. Experiment results are evaluated considering the implications of different QoS support types on network coverage and capacity planning optimizations. Our results demonstrate that even though pre-configured resource allocations can be given to prioritize UEs, the experience of all the UEs can be affected unexpectedly in the presence of many UEs who have received different QoS deployment support. Our experimental observations have revealed that location of the UEs with respect to Base Station (BS) and the availability of dedicated bandwidth UEs inside cell may have implications on the apriori defined resource allocation strategies of the other UEs.

INDEX TERMS QoS, differentiated services, mobile networks, experiments.

I. INTRODUCTION

In mobile networks, quality-of-service (QoS) refers to the capability of providing better service for the selected traffic types of a network under the same underlying technology. One of the main reasons of using QoS techniques for Mobile Network Operators (MNOs) is to detect and differentiate the prioritized services. Therefore, preferential services such as real-time applications can have higher priority over other services in MNO infrastructure. The emerging wireless applications in 5G require ultra-low latencies with ultra-high reliability [1]. Some users that demand to differentiate for the same type of services creates the problem. Inevitably, this will create an opportunity for extra income for the service provider [2]. Although 5G architecture offers promising key technologies to revolutionize the spectrum usage and energy efficiency, QoS support makes great sense for the more efficient use of 5G radio resources [3]. QoS allows the user equipment (UE) services to run according to their importance where the most critical services can be served first. Inside the infrastructure, QoS identifiers are used as marks that define the tolerable packet loss rate, packet delay budget, etc. of a mobile network service. The mobile network is aware of the services that are marked with QoS differentiation and handles them differently in accordance with the QoS deployment strategy of the MNO. This differentiation can only be done for the services used by certain UEs that are pre-configured in Core Network (CN), transport network and radio access network (RAN).

In QoS deployment, MNOs simply assign a priority level to the type of traffic that is to be managed and then specify how the RAN will behave for these different types of traffic

The associate editor coordinating the review of this manuscript and approving it for publication was Petros Nicopolitidis.

based on QoS deployment strategy. In this case, the main focus of MNOs is to select the most suitable techniques to manage the underlying network resources efficiently. Thus, there is a trade-off between simplifying the complexity of network and providing the best QoS support to the UE services. An accurate QoS deployment strategy should prevent unequal distribution of mobile network's resources to UEs while satisfying the requirements.

In this paper, we discuss two main QoS deployment strategies that can be performed in the nodes of a real mobile network. The first one is to provide more scheduling rates in the RAN side for the UE services based on their QoS values. This method is based on assigning a different and higher QoS value to the UE services demanded by the prioritized UEs so that this QoS value can be scheduled at higher rates in the Media Access Control (MAC) scheduler of the Base Station (BS). UEs differentiated with this method are priority UEs. The second one is to run the UE services with a guaranteed bit rate so that their resources do not fall below a certain bandwidth requirement. The QoS requirement in this case is to control the amount of allocated bandwidth for each UE service and allow them to consume bandwidth based on their service requirements and mission. In this paper, we concentrate on experimental validations of these different QoS deployment strategies to analyze the change behaviour of UE Key Performance Indicators (KPIs) when diverse set of QoS requirements co-exist under the same network cell.

The rest of the paper is organized as follows: Section II gives the related works and main contributions of the paper. In Section III, we detail the system model, concepts and different QoS deployment scenarios used throughout the text. In Section IV, we give the details of the experimental components and our experimental results for existence of different QoS types UEs inside the cell coverage as well as point out some of the main outcomes of the conducted experimental scenarios. Finally, in Section V we give the conclusions of the paper.

II. RELATED WORK AND MOTIVATION

We review the state-of-the-art works in three main parts.

A. QoS SUPPORT

There are various works that investigate QoS support for different mobile network deployment scenarios. For QoS Long Term Evolution (LTE) networks, the article in [4] proposes an analytical QoS model in terms of context load, processing load and memory access rate in various elements. QoS provisioning solutions are presented in [8], [9] and the QoS needs of critical communications are detailed in [10] for 5G networks. The study in [14] proposes a resource allocation scheme with content caching in Software-Defined Networking (SDN) based networks. For Network Function Virtualization (NFV) cases, the authors in [15] study the placement of Virtual Network Functions (VNFs) to provide better QoS for MNOs. The article in [16] investigates load balancing solutions to optimize the QoS of a cloud radio access network. The authors in [17] study user grouping strategies while considering the diverse QoS requirements of users in Non-Orthogonal Multiple Access (NOMA) systems. A game theoretic approach for interference-limited cellular environments is presented in [11]. The QoS support in LTE heterogeneous network (HetNet) is investigated in [5], [6]. In [12], a framework for implementing a QoS-aware energy and jitter efficient scheduling methodologies in downlink for HetNet is developed. For HetNet, the authors in [13] exploit the network cooperation and propose two joint radio resource management schemes with energy savings while satisfying the system QoS performance. A cluster-based resource allocation scheme is studied in [7] to resolve the resource allocation problem with QoS guarantees for ultra dense networks. All these studies have specialized focus on how QoS structures will take place in the future. However, sufficient practical information on how MNOs will follow a strategy in new generation mobile networks is still missing in those works.

The 3rd Generation Partnership Project (3GPP) standardization has also defined several UE categories for LTE [20]. It is basically defined to categorize both uplink and downlink capabilities of UE. Depending on the UE capability, BSs can connect to each UE more effectively. In LTE standardization and networks, QoS between UE and packet data network (PDN) gateway is applied using "bearers" which represent a set of network configurations so that prioritization of the traffic is handled based on the desired level of QoS guarantees. Moreover, in LTE networks QoS Class Identifiers (QCI) is defined that consists of basic classes which are classified as "default", "expedited forwarding" and "assured forwarding". For example, QCI for Guaranteed Bit Rate (GBR) bearers is between 5 and 9 and for non-GBR bearers is between 1 and 4. QCIs simply deal with UEs that are requesting different services so that the LTE schedulers can set the priorities among them. The UE categories are designed for specific use cases. In 5G networks, QoS model is based on QoS Flows and a standardized 5G QoS Identifier (5QI) to QoS characteristics mapping is given in 3GPP Release 15 [31]. The complexity of UE categories are also defined using Transmission Mode (TM) in 3GPP [21]. For example in 5G networks, in general low numbered UE categories are designed for massive Machine-Type of Communications (mMTC) use cases, whereas high numbered UE categories are especially designed for considering Enhanced Mobile Broadband (eMBB) or Ultra-Reliable Low-Latency Communication (URLLC) use cases. In a mMTC scenario, massive number of UEs with NB-IoT capabilities accessing cellular network are simulated together using a QoS-aware priority-based scheduling strategy in [18]. Based on 3GPP QoS rule, an algorithm design that prioritizes GBR and non-GBR bearers of difference QCIs is designed in [19]. Different from the traditional UEs that access the network, our study provides a hint of the QoS deployment strategy for MNOs and how the management of UEs with differentiated QoS services can be accomplished for next generation mobile services with core network configuration assistance for different user types.

B. SCHEDULER DESIGN

The problem of QoS support using various scheduling algorithms has been proposed for cellular networks with different objectives, such as throughput, latency, fairness, energy, etc in [22]–[24]. A throughput maximizing method using max-weight based scheduling algorithm is proposed in [23]. Out of different studied scheduling algorithms, Proportional Fair (PF) scheduling algorithm has attracted higher interest among the academic as well as industrial community and is widely adopted. The authors in [22] have extended PF scheduling to assign higher priority to Resource Blocks (RBs) that yield above average spectral efficiency. The authors in [24] have compared both QoS-aware and QoS-non-aware scheduling algorithms for multi-tenant SDN-based infrastructure of cellular networks. However, these studies have mostly focused on either simulations or theoretical works of demonstrating the benefits of scheduling to maximize a given objective under given constraints. In this work, we formulate the QoS deployment in MNO environment using an optimization problem. In addition, we also propose a scheduling algorithm as a solution to the defined optimization problem when there are various kinds of differentiated UEs with QoS prioritization. We have also investigated the existence of dedicated bandwidth UEs in the network environment to observe the QoS behaviour of all types of UEs.

C. EXPERIMENTAL TRIALS

There are also various works on the experimental performance analysis of QoS in the literature. In [27], the authors study the QoS performance of LTE networks under different load scenarios. In [25], the focus is to analyze the correlation between UE position, network load and QoS performances for video specific services. For 5G case, a heterogeneous QoS-driven resource allocation policy for mmWave in massive Multiple Input Multiple Output (MIMO) is presented in [28]. Experimental evaluation of a utility based decision-making approach for wireless mobile broadband networks is investigated in [26]. In [18], the authors aim to analyze the performance of the QoS aware Narrow Band IoT (NB-IoT) networks. QoS changes with different physical configurations can also be provided and an example case is the electrical tilt [29]. However, different from those studies, in this paper we evaluated the QoS deployment methods that can be implemented in MNOs infrastructure and showed what QoS strategy can be selected in a real network when commercial UEs exist rather than focusing on validations based on simulation environments. In our previous work [30], we have provided an experimental work that enables differentiated QoS for different types of LTE users. In this paper, we extend this analysis by adding dedicated bandwidth UEs into the experimental setup to observe the end-to-end QoS change with different types of UEs, namely, normal, priority UEs with higher scheduling rates and priority UEs with dedicated bandwidth.

D. OUR CONTRIBUTIONS

The motivation of this paper stems from the fact that most of the literature work mentioned above does not observe the co-existence of UEs with different set of QoS requirements in real operational network. As a matter of fact, there are various efforts to quantify the QoS improvements of various schedulers on LTE users. For example in LTE standardization efforts, differentiated QoS has been considered and 9 standardized QCIs have also been characterized (and 12 more QCIs are added for 5G networks [10], [31]). However, none of the previous works have observed the effects of scheduling and priority weighting on the performance of priority and normal UEs active throughput when real-users in a real network operation scenario is activated and deployed over MNOs infrastructure. Moreover, most of the previously available research works have concentrated on validations via simulations but not using real world experimental trial. Our contributions in this paper can be summarized as follows: (i) We build a real-world test network environment to observe the end-to-end KPI performance values. The experiments were run with three different UE types (namely normal as well as two prioritized UEs with dedicated bandwidth requirements and higher scheduling rates) that are created inside the LTE network infrastructure. (ii) We detail some of the characteristics, limitations and benefits of utilizing prioritized UEs with dedicated bandwidth requirements and higher scheduling rates inside cellular infrastructure. (iii) We have shown experimentally that deploying proposed QoS strategies for diverse set of UEs requires careful network capacity and coverage planning, which are detailed in discussions and main takeaways section of the paper. As a summary, Table 1 provides a summary comparison between various techniques discussed above and the proposed approach in this paper.

Notations: Throughout the paper, bold letters represent vectors, i.e., $\mathbf{x}$ is a vector, and its i-th element is denoted by x_i. The sets are denoted by upper case calligraphic symbols. 0_M is the all-zeros column vector of size M. $||x||_1$ denotes the L1 norm of vector x.

III. SYSTEM MODEL AND CONCEPTS

Fig. 1 shows a high level general diagram of the considered scenario where there are normal and prioritized UEs distributed around the cells which are connected to core network to provide connectivity services to all UEs inside the coverage area of a cell. In Table 2, the key notations symbols used throughout the paper are summarized. In our considered experimental scenario, we assume that $\mathcal{N} = \{1, 2, \ldots, N\}$ represents the set of multiple UEs with diverse set of QoS requirements with N UEs in a given cell. Denote $\mathcal{N}_p \subset \mathcal{N}$ as the set of prioritized UEs. Moreover, denote $\mathcal{N}_p^d \subset \mathcal{N}$ as the set of prioritized UEs with dedicated bandwidth requirements of K_f RBs with N_p^d prioritized UEs of this type using the carrier frequency f, $\mathcal{N}_p^s \subset \mathcal{N}$ as the set of prioritized UEs

TABLE 1. A summary comparison of various existing QoS techniques and validation approaches with the proposed approach.

	Different Approaches		Proposed Approach	
	Characteristics	Limitations	Advantages	Differences
QoS Support	—Develop QoS models [4], [5], [6], [7]. —Focus on QoS provisioning [8], [9], [10], [11] —Examining energy-efficiency aspects [12], [13] —Different network deployment technologies [14], [15], [16], [17] —Implementation of QCI awareness [18], [19]. —Define new roles for UEs. [20], [21].	—Lack of practical deployment —Focus on future implementations. —Immature technology type. —Complex QCI configurations. —Needs enhancements at UE side.	—Evaluates the performance with UEs with different QoS requirements. —Provides QoS deployment strategy for MNOs. —Real-world implementation	—Management of UEs with differentiated QoS services. —Demonstrate the trade-off between QoS strategies and adjusting capacity resources.
Scheduler Support	—QoS support using various scheduling algorithms [22], [23], [24]. —Demonstrate the benefits to maximize a given objective (throughput, latency, fairness, energy, etc.) [22], [23], [24]	—Focus on either simulations or theoretical works. —Needs new specialized hardware to support the system —Buffer size issues.	—Scheduler design over existing LTE infrastructure —Independent from deployment location. —No additional hardware enhancement needed. —Provides 3GPP compatibility. —Applicable to all current UEs. —Prioritization is preserved at cell-edges.	—A new scheduling algorithm with differentiated UEs with QoS prioritization. —Targets to maximize throughput and fairness. —Hierarchical resource allocation based on QoS level. —Weighted QoS assignment. —Core network configuration assistance for different user types. —Existence of dedicated bandwidth UEs
Experimental Trials	—QoS investigation for different services [25], [26], [18] —QoS policy [27], [28] —Network configuration aspects [29], [30]	—The policy needs to be supported by the tariffs to generate revenue.. —Needs special configuration for each site.	—Real-users in a real network operation scenario deployed over MNOs infrastructure. —Ready to be implemented in an MNO network.	—Detailing limitations and benefits of utilized experimental setup and outcome. —Monitoring cell-center & cell-edge performance characteristics. —Strategy based comparison.

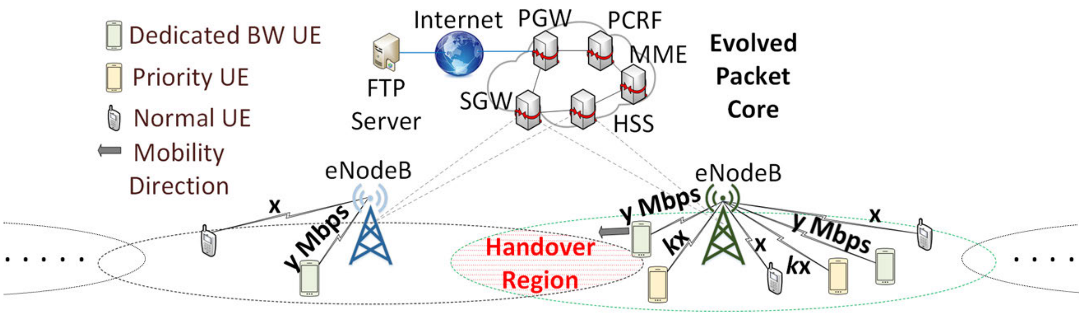


FIGURE 1. Cellular network with eNodeBs providing services for normal, priority and dedicated UEs.

with N_p^s prioritized UEs having k times higher scheduling rate than normal UEs and $\mathcal{N}_n \subset \mathcal{N}$ as the set of normal UEs with N_n normal UEs as shown in Fig. 1. Note that $\mathcal{N}_p = \mathcal{N}_p^s \cup \mathcal{N}_p^d$ and $\mathcal{N} = \mathcal{N}_p \cup \mathcal{N}_n$. Let $\mathcal{T} = \{1, 2, \ldots, T\}$ denote the set of the observation time where T is the duration of observation. Moreover, we also would like to point out that the analyzed problem considers a single shared Serving Gateway (S-GW) resource serving multiple connected BSs as shown in Fig. 1 as well.

We assume that there are M available RBs in a given cell at $t \in \mathcal{T}$. We denote the RB set $\mathcal{M} = \{1, 2, \ldots, M\}$. Let the binary variable $q_{m,n}^t$ indicate whether UE $n \in \mathcal{N}$ is assigned to RB $m \in \mathcal{M}$ or not (i.e., if n-th UE is assigned to m-th RB, then $q_{m,n}^t = 1$ else $q_{m,n}^t = 0$) at $t \in \mathcal{T}$ i.e. during Transmission Time Interval (TTI) (usually equal to 1 millisecond). Hence, each UE $n \in \mathcal{N}$ is assigned to only one RB so that

$$\sum_{m \in \mathcal{M}} q_{m,n}^t = 1, \quad \forall n \in \mathcal{N}. \tag{1}$$

We define $\mathbf{\Delta^t} = [\mathbf{\Psi}_1^t \;\; \mathbf{\Psi}_2^t \;\; \ldots \;\; \mathbf{\Psi}_n^t] = (\mathbf{\Psi}_n^t, \mathbf{\Psi}_{-n}^t)$ as the $M \times N$ UE assignment matrix of all RBs at $t \in \mathcal{T}$. Here $\mathbf{\Psi}_n^t = [q_{1,n}^t \;\; q_{2,n}^t \;\; \ldots \;\; q_{M,n}^t]^T$ denotes n-th UE's $M \times 1$ RB assignment vector and $\mathbf{\Psi}_{-n}^t$ denotes the assignment vector of all UEs other than the n-th UE. Denote $\mathbf{S}_n^t = [s_{1,n}^t, s_{2,n}^t, \ldots, s_{M,n}^t]^T$ as $M \times 1$ vector of the obtained Modulation Coding Scheme (MCS) index values $s_{m,n}^t$ of n-th UE and m-th RB at $t \in \mathcal{T}$. Let us also denote $\mathbf{R}_n^t = [r_{1,n}^t, r_{2,n}^t, \ldots, r_{M,n}^t]^T$ as $M \times 1$ vector of the obtained Transport Block Size (TBS) values $r_{m,n}^t$ of n-th UE and m-th RB at $t \in \mathcal{T}$. Hence, $\mathbf{R^t} = [\mathbf{R}_1^t \;\; \mathbf{R}_2^t \;\; \ldots \;\; \mathbf{R}_n^t]$ and $\mathbf{S^t} = [\mathbf{S}_1^t \; \mathbf{S}_2^t \; \ldots \; \mathbf{S}_n^t]$ are the $M \times N$ matrix of TBS and MCS index values at $t \in \mathcal{T}$ respectively. Define $\Upsilon_n^t = (\mathbf{\Psi}_n^t)^T \times \mathbf{R}_n^t$ as the achieved data rate at $t \in \mathcal{T}$ for UE $n \in \mathcal{N}$.

Note that UEs exchange information with its corresponding eNodeB in a particular region with an assigned TBS value in a given TTI. The maximum value of assigned TBS for a given eNodeB for each UE is identified by an integer value of α in this paper. For this reason, each UE can get at most TBS value of α.

A. PROBLEM FORMULATION

The problem definition can be described as follows: Given a network state $\mathbf{S} = (\mathbf{\Psi}_n^t, \mathbf{\Psi}_{-n}^t)$ where $(\mathbf{\Psi}_n^t, \mathbf{\Psi}_{-n}^t)$ is a combination of each RB assignment in the set of $\mathcal{M}$ to each UEs in the set $\mathcal{N}$, we look for the optimal values of assignments to minimize a cost function $f(\mathbf{\Psi}_n^t, \mathbf{\Psi}_{-n}^t)$:

$$f(\mathbf{\Psi}_n^t, \mathbf{\Psi}_{-n}^t) = -\sum_{n \in \mathcal{N}} U_n^t, \tag{2}$$

TABLE 2. Symbols used throughout the paper.

Symbol	Meaning
N, $\mathcal{N}$	Number of UEs , UE set
N_p^d, $\mathcal{N}_p^d$	Number of prioritized UEs with dedicated bandwidth requirements of K_f RBs, corresponding prioritized UE set
N_p^s, $\mathcal{N}_p^s$	Number of prioritized UEs with k times higher scheduling rate than normal UEs, corresponding prioritized UE set
N_n, $\mathcal{N}_n$	Number of normal UEs , Normal UE set
M , $\mathcal{M}$	Number of RBs , RB set for all UEs
$\mathcal{M}'$, $\mathcal{M}''$, $\mathcal{M}'''$	the set of RBs assigned to dedicated, priority and normal UEs respectively
$\mathbf{\Delta^t}$	$M \times N$ RRH assignment matix of all UEs over all RBs
$\mathbf{\Psi}_n^t$	$M \times 1$ RB assignment vector for n-th UE at $t \in \mathcal{T}$
$q_{m,n}^t$	$= \begin{cases} 1; & \text{if m-th RB is assigned to n-th UE at TTI t} \\ 0; & \text{else} \end{cases}$
$\mathbf{R}_n^t$	$M \times 1$ vector of the obtained TBS values of n-th UE in M different RBs
$r_{m,n}^t$	TBS value obtained at m-th RB for n-th UE.
$\mathbf{S}_n^t$	$M \times 1$ vector of the obtained MCS index values of n-th UE in M different RBs
$s_{m,n}^t$	MCS index value obtained at m-th RB for n-th UE.
α	Maximum achievable TBS value by each UE
$U_{n,m}^t$	utility metric obtained for the n-th UE using the m-th resource
λ_n^t	the average data rate of the n-th UE at time t.
$R_{m,n}^t$	obtained instantaneous data rate of the n-th UE at m-th resource and time t.
k	the scheduling constant to provide higher data rate for UEs
τ	time constant of smooth filter in PF scheduler of LTE network
Δt	allocation interval in PF scheduler of LTE network
K_f	maximum number of RBs allocated by prioritized UEs with dedicated bandwidth requirements on carrier frequency f
Υ_n^t	the achieved data rate at $t \in \mathcal{T}$ for UE $n \in \mathcal{N}$ and is equal to $(\mathbf{\Psi}_n^t)^T \times \mathbf{R}_n^t$

where U_n^t is the utility of the n-th UE at $t \in \mathcal{T}$. In order to accomplish this, each UE's utility needs to be maximized by choosing appropriate RB assignments. Using assigned TBSs as the maximization parameter, the utility function of n-th UE is expressed as follows:

$$U_n^t = \sum_{m \in \mathcal{M}} U_{m,n}^t = \sum_{m \in \mathcal{M}} \left(q_{m,n}^t \times r_{m,n}^t \right), \quad \forall n \in \mathcal{N}, \quad \forall t \in \mathcal{T}, \tag{3}$$

where the term $U_{m,n}^t = q_{m,n}^t \times r_{m,n}^t$ is the obtained TBS value of the UE $n \in \mathcal{N}$ for RB $m \in \mathcal{M}$ at $t \in \mathcal{T}$ [1]. Then, the optimization problem can be described as follows: Our goal is to maximize the sum of total TBS utility of all UEs with the decision variables: (i) *Assignment problem:* the assignment of UE to each RB is represented by the variables $q_{m,n}^t$. (ii) *UEs types QoS satisfaction problem:* the resource allocation between two different prioritized UEs and normal UEs are characterized by Υ_n^t and dedicated BW UEs by $||\mathbf{\Psi}_n^t||_1$.

[1]Note that TBS values are extracted from a table mapping obtained using MCS index and number of RBs values according to 3GPP specification to determine how many bits can be transmitted per one TTI for Physical Downlink Shared Channel (PDSCH) [32]. The TBS index, $i \in \{0, \ldots, 26\}$, is a function of modulation and coding scheme as given in Table 8.6.1-1 of [32]

For the considered shared mobile architecture, we use the following formulation for our optimization problem:

$$\underset{\mathbf{\Delta^t}}{\text{minimize}} \quad f(\mathbf{\Psi}_n^t, \mathbf{\Psi}_{-n}^t) \tag{4}$$

$$\text{subject to} \quad \sum_{m \in \mathcal{M}} q_{m,n}^t = 1, \quad \forall n \in \mathcal{N}, \forall t \in \mathcal{T}, \tag{4a}$$

$$\{q_{m,n}^t\} \in \{0, 1\}, \quad \forall m \in \mathcal{M}, \forall n \in \mathcal{N}, \tag{4b}$$

$$||\mathbf{\Psi}_n^t||_1 = K_f, \quad \forall n \in \mathcal{N}_p^d, \tag{4c}$$

$$\Upsilon_i^t = k \times \Upsilon_j^t, \quad \forall i \in \mathcal{N}_p^s, \forall j \in \mathcal{N}_n, \forall t \in \mathcal{T}, \tag{4d}$$

$$r_{m,n}^t \leq \alpha, \quad \forall n \in \mathcal{M}, \forall n \in \mathcal{N}, \forall t \in \mathcal{T}. \tag{4e}$$

where the objective function given in (4) is to maximize the total sum of TBS values over all UEs, RBs and observation duration $t \in \mathcal{T}$. Constraint in (4a) illustrates that each UE is assigned to only one RB, (4b) denotes the binary value constraint of UE assignment per RB, (4c) gives the dedicated bandwidth requirements of prioritized UEs at carrier frequency f, (4d) gives the constraint imposed for prioritized UEs with k times higher scheduling rate than normal UEs and (4e) yields the maximum achievable TBS value by each UE at a given TTI $t \in \mathcal{T}$.

The time scale of the operation of the maximization operate at faster scale at the BSs, than the various gateways at the core or transport networks due to dynamic nature of the propagation environment. The optimization problem can be solved for each TTI in a given time frame depending on the use cases and the requirements. Note that in the optimization problem defined above, the main decision process at the scheduling layer is to decide the number of RBs assigned to each UEs based on their time-frequency locations, MCS and transmission power. In our optimization problem, we are dealing with a binary optimization problem. The optimization variables are binary, hence this problem is an Integer Linear Problem (ILP) [33] (and also c.f. Problem (P1) in [34]). However, ILPs are difficult to solve and known to be NP-hard and with exponential execution time, preventing solutions even for reasonable problem sizes. On the other hand, there are also several approximations (e.g. continuous methods) for binary optimization in the literature [35]. One approach is to recognize that after relaxation, the problem is a linear program (LP) and enforce the binary constraints after solving the LP.

Moreover, solving (4) is challenging because (i) of the existence of coupling behaviour in UE assignments for each RB problem and large-scale QoS satisfaction requirements problem for different UE types and carrier frequency, (ii) the achievable maximum TBS value per TTI depends on many factors such as wireless link, channel bandwidth, number of UEs, locations of UEs, interference, that may not be controllable, (iii) the globally optimal UE assignment per RB decision solution for given demands depends on QoS requirements of multiple number and types of UEs. This is non-tractable in large-scale and has high computational complexity.

Therefore, in the following section, we will discuss a heuristic approach to the optimization problem given in (4). In our scheduling design methodology, UEs-RBs assignments are done while providing the necessary QoS guarantees for different UE types. This scheduler design is also used throughout our experiment trial to observe its implications in real-life scenarios.

B. HEURISTIC SCHEDULER DESIGN

To find a solution for the optimization problem defined in (4), we study a scheduler design and propose an algorithm in this section. Schedulers are one of the core components of LTE systems utilized in eNodeBs for resource management among UEs and network performance optimization. In its basic functionality, schedulers allocate resources to UEs based on their Channel Quality Indicator (CQI) and QoS requirements which can be defined by MNOs.

Some key challenges of providing end-to-end QoS support for LTE users are: First, to map different QoS classes across different domains (such as RAN, transport and core networks). Second, to provide the appropriate scheduling methodologies that can enable QoS differentiation among users. In eNodeBs of 4G systems, one of the main scheduler methodology is using PF scheduling algorithm. It provides a balance between fairness and overall spectrum efficiency. The performance metric of PF algorithm for the n-th UE can be written as

$$R^t_{m,n} = \frac{U^t_{m,n}}{\lambda^t_n}, \tag{5}$$

where $U^t_{m,n} = q^t_{m,n} \times r^t_{m,n}$ is the instantaneous achievable data rate (or TBS) of the n-th user, at m-th RB and time $t \in \mathcal{T}$, λ^t_n denotes the average data rate of the n^{th} UE until time $t \in \mathcal{T}$ and it can be calculated by

$$\lambda^t_n = \left(1 - \frac{1}{\tau}\right) \times \lambda^{(t-\Delta t)}_n + \sum_{\forall m \in \mathcal{M}} \frac{q^{(t-\Delta t)}_{m,n} \times U^{(t-\Delta t)}_{m,n}}{\tau}, \tag{6}$$

where $\tau > 1$ denotes the time constant of smooth filter which controls the system latency and is the past window length, Δt is the TTI, which is the period of allocation. Note that window length value τ gives a trade-off between throughput and latency. Higher window value results in higher throughput since the scheduler waits for the large peaks. This in turn increases the latency. Lower window value indicates low waiting period for throughput peaks, whereas it decreases the latency [36]. To solve the above optimization problem in an experimental set-up, we perform configuration updates on the existing scheduler methods of eNodeBs. In our experimental trials, we have used the following utility metric,

$$\bar{R}^t_{m,n} = w^t_n R^t_{m,n}, \tag{7}$$

where w^t_n is the weight assigned to each user $n \in \mathcal{N}$ based on their priority status. For prioritized UEs with k times higher scheduling rate than normal UEs $w^t_i / w^t_j = k, \ \forall i \in \mathcal{N}^s_p, \forall j \in \mathcal{N}_n, \forall t \in \mathcal{T}$. Without loss of generality, we assign $w^t_i = 1, \forall i \in \mathcal{N}_n$. Hence during experimental trials, the utility metric in (7) becomes,

$$\bar{R}^t_{m,n} = \begin{cases} R^t_{m,n}, & \text{if } n \in \mathcal{N}_n, \mathcal{N}^d_p \\ k \times R^t_{m,n}, & \text{if } n \in \mathcal{N}^s_p \end{cases} \tag{8}$$

Algorithm 1 Pseudo Code of the Heuristic Scheduling Algorithm Used in Experimental Setup

Input: $\mathbf{S^t}$, k, $\mathcal{M}$, $\mathcal{N}$, K_f
Output: RBs-UEs $M \times N$ assignment matrix, $\mathbf{\Delta^t}$ at $t \in \mathcal{T}$
Initialization: $\mathcal{M}' = \emptyset$, $\mathcal{M}'' = \emptyset$, $\mathcal{M}''' = \emptyset$, $\mathbf{\Psi}^t_n = \mathbf{0}_M$.

1: **procedure** SCHEDULE
2: **Foreach** (UE-n $\in \mathcal{N}$)
// Calculate RB assignment vector for n //
3: **Compute:** $U^t_{m,n}$ and λ^t_n, $\forall m \in \mathcal{M}$ using $\mathbf{S^t}$
4: **if** $n \in \mathcal{N}^d_p$ **then** ▷ Iterate until min. bandwidth
5: **while** ($|\mathcal{M}'| < K_f$) **do**
// Allocate free RBs for Dedicated UEs //
6: **Run:**($\mathbf{\Psi}^t_n$, *FLAG*) = *ALLOCATE*($\mathcal{M}'$, $\mathcal{M}$, $\mathcal{M}''$, $\mathcal{M}'''$)
7: **if** !FLAG **then**
8: **break** // Requirement satisfied //
9: **end if**
// Reallocate RBs of Normal UEs //
10: **Run:** ($\mathbf{\Psi}^t_n$, *FLAG*) = *REALLOCATE*($\mathcal{M}'$, $\mathcal{M}'''$)
11: **if** !FLAG **then**
12: **break** // Requirement satisfied //
13: **end if**
// Reallocate RBs of Priority UEs //
14: **Run:** ($\mathbf{\Psi}^t_n$, *FLAG*) = *REALLOCATE*($\mathcal{M}'$, $\mathcal{M}''$)
15: **if** !FLAG **then**
16: **break** // Requirement satisfied //
17: **else** //Share RBs between dedicated UEs//
18: **Run:** $\mathbf{\Psi}^t_n$ = *RB_ALLOCATION* ($\mathcal{M}'$, $U^t_{m,n}$, λ^t_n, k, $\mathcal{N}$)
19: **break** // Requirement satisfied //
20: **end if**
21: **end while**
22: **else** //Calculate RB assignment vector for// //normal and priority UEs//
23: **Run:** $\mathbf{\Psi}^t_n$ = *RB_ALLOCATION* ($\mathcal{M}$, $U^t_{m,n}$, λ^t_n, k, $\mathcal{N}$)
24: **end if**
25: **end foreach**
26: **end procedure**

Algorithm 1 and corresponding function calls in Algorithm 2 summarizes the pseudo code of the RB allocation strategy for each UE with different QoS requirements. Note that in line #18 of Algorithm 1, dedicated BW UEs uses (8) similar to normal UEs after obtaining its required RBs.

Algorithm 2 Utilized RB Assignment Functions

1: **procedure** ALLOCATE($\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$)
2:     **while** ($\mathcal{B} \setminus \{\mathcal{A} \cup \mathcal{C} \cup \mathcal{D}\}$ is $\{\emptyset\}$ and *FLAG*) **do**
3:         **Set:** *FLAG* = True
4:         **Select:** item e from set $\mathcal{B} \setminus \{\mathcal{A} \cup \mathcal{C} \cup \mathcal{D}\}$
5:         **Update:** $\mathcal{A} \leftarrow \mathcal{A} \cup \{e\}, \mathcal{B} \leftarrow \mathcal{B} \setminus \{e\}$.
6:         **if** $|\mathcal{A}| > K_f$ **then**
7:             *FLAG* = False
8:             **Update:** Ψ_n^t using $\mathcal{A}$ and $\mathcal{B}$
9:         **end if**
10:     **end while**
11:     **Return:** (Ψ_n^t, FLAG)
12: **end procedure**

13: **procedure** REALLOCATE($\mathcal{A}, \mathcal{B}$)
14:     **while** ($\mathcal{B}$ is $\{\emptyset\}$ and *FLAG*) **do**
15:         **Set:** *FLAG* = True
16:         **Select:** item e from set $\mathcal{B}$
17:         **Update:** $\mathcal{A} \leftarrow \mathcal{A} \cup \{e\}, \mathcal{B} \leftarrow \mathcal{B} \setminus \{e\}$.
18:         **if** $|\mathcal{A}| > K_f$ **then**
19:             *FLAG* = False
20:             **Update:** Ψ_n^t using $\mathcal{A}$ and $\mathcal{B}$
21:         **end if**
22:     **end while**
23:     **Return:** (Ψ_n^t, FLAG)
24: **end procedure**

25: **procedure** RB_ALLOCATION($\mathcal{M}, U_{m,\underline{n}}^t, \lambda_n^t, k, \mathcal{N}$)
26:     **Find:** $(m^*, n^*) = \arg\max_{m \in \mathcal{M}, n \in \mathcal{N}} \left\{ \bar{R}_{m,n}^t \right\}$ using (8)
27:     **Update:** Ψ_n^t using m^*
28:     **Return:** (Ψ_n^t, FLAG)
29: **end procedure**

In summary, we propose a simple heuristic scheme to solving the optimization problem at hand with low computational complexity that can work on the desired time scale. The algorithm is simple, but an essentially greedy scheme (not necessarily the optimal one). It is based on time-level optimization. Each RB can only be allocated to a single user in any TTI. At any given $t \in \mathcal{T}$, the first priority is to assign the available resources to priority UEs (i.e. dedicated BW UEs) in order to satisfy the minimum throughput requirements of dedicated BW priority UEs. No extra RBs are given to dedicated BW UEs. When the number of RBs were enough for dedicated BW UEs, the remaining RBs would be used to maximize the throughput of remaining priority UEs and normal UEs. After this requirement is satisfied, we try to maximize the remaining priority UEs' (with higher scheduling rates) throughput or equivalently maximize the number of scheduled priority UEs in the current $t \in \mathcal{T}$. (8) tries to enforce that if the priority UEs with higher scheduling rates are scheduled, they should transmit k times more data bits than the normal UEs. In the case of infeasibility, i.e. the problem does not lead into any optimal solution (no resources available for additional priority users (i.e. dedicated BW UEs) at any time $t \in \mathcal{T}$), the heuristic solution serves as many dedicated BW UEs as possible, excluding the remaining UEs from RB allocations. Hence, the algorithm always converges even though it may not be feasible for some UEs.

C. COMPLEXITY ANALYSIS

The heuristic scheme presented in Algorithm 1 is motivated by the complexity of the schemes in the optimal solution. With $M \times N$ variables, the complexity of the optimization in (4) is $\mathcal{O}(2^{M \times N})$. In Algorithm 1, the worst case would be when dedicated BW request cannot be satisfied with allocation of all RBs. So the number of RBs are not enough for the remaining UEs (i.e. priority users with higher scheduling rates and normal users). In this case, lines #5-21 in Algorithm 1, #2-10, and #14-22 in Algorithm 2 are executed. When M and N are large, the complexity is due to the two while iterations (Line #5 in Algorithm 1 and line #2 in Algorithm 2) and also finding "arg max" (line #26 in Algorithm 2. The complexity for performing arg max is proportional to number of values being sorted. The overall time complexity of Algorithm 1 with Algorithm 2 is therefore $\mathcal{O}(M^2 \times N)$.

D. PRACTICAL SETTINGS AND CHARACTERISTICS OF QoS DEPLOYMENTS

In Fig. 1's architecture, to provide end-to-end QoS support for all types of UEs, RAN, transport and core network equipment need to be configured appropriately for each type of defined UEs with a given set of QoS requirements. End-to-end QoS deployment strategy should be assured and managed by entities that have a global mobile network topology view. The basis of QoS support in LTE networks is accomplished via Evolved Packet System (EPS) bearers [37]. An EPS bearer builds a logical channel between UE and PDN. EPS bearers can be classified as GBR and non-GBR bearers depending on scheduling and queue management policy. In GBR bearers, a permanent network resource is allocated whereas in non-GBR this does not exist. In our experimental scenarios, prioritized UEs are created by defining different non-GBR QCI profile at Home Subscriber Server (HSS) where QCI levels are assigned statically. This assigned QCI gives higher priority to prioritized UEs with high scheduling rate than normal UEs. During operation in first step, UE service requests QoS value from the CN via Control Plane (CP) signaling during service initiation. Then, the CN assigns the pre-defined QoS value that will be used in the User Plane (UP) session of the UE service (such as voice, eMBB). Therefore, QoS values for UE services in MNO environment are assigned by the CN. RAN equipment is just executing the QoS policies (e.g. via scheduling) in radio access segment depending on the QoS assignments done at CN.

For our experimental scenarios, infrastructure provider plans to provide a dedicated wireless resource allocation for its customers (e.g. to a national bank that has

TABLE 3. Comparisons of QoS deployment strategies to provide cellular services.

QoS Deployment Strategy	Characteristics	Limitations	Advantages/ Benefits
Higher Scheduling Rates	— Prioritize UEs by assigning scheduling rates. — Define different non-GBR QCI profile in HSS. — Prioritized UE connects with higher QCI which is assigned with higher scheduling rate.	— Minimum bandwidth requirements can not be satisfied due to increased number of UEs. — Bandwidth stability can be affected more when handover between eNodeBs occurs. — Must be applied considering the service types (e.g. mobile broadband services do not need to be prioritized.)	— Can be configured for more UEs. — Better for high mobility cases. — Has little effect on overall bandwidth capacity. — Better for protecting non-critical services. — Relatively easy to configure and optimize.
Dedicated Bandwidth	— Priority users are statically assigned to a predefined QCI level. — Pre-defined QCI level is assigned with a minimum bandwidth value. — Scheduling rate to provide is adapted based on all connected UEs to base station.	— Consumes bandwidth faster, with increased number of dedicated bandwidth UEs. — Detrimental effect over normal and priority UEs. — Cannot be assigned in cases where too many UEs are in base station's coverage (a disadvantage for critical services). — Careful RAN planning required (cell-edge positioning severely degrades overall cell performance).	— Effective in conditions where channel is less time-variant. — Great benefit for critical services. — Better performance for slow-moving UEs. — Better to satisfy bandwidth requirements.

nationwide branches). The intended use case is to migrate Automated Teller Machines (ATMs) from g.shdsl [38] which is an old technology using fixed line access into a wireless access device. For this reason, the required Upload (UL) RB allocation for an ATM is selected to be K_f. In this case, the location, received signal quality and number of the ATMs inside the coverage cell are quite important so that BSs can service to those ATMs with fixed bandwidth wireless access while also having minimum impact on existing normal UEs. For this reason before selecting the experimental fields and corresponding cells, prior studies on how many dedicated bandwidth ATMs can be accommodated in a certain region need to be studied via appropriate network planning tools while considering the expertise of network planning experts. Other prioritized UEs can have higher scheduling rates than normal UEs. These UEs are generally the public and private enterprise customers of MNOs. To give insights into different QoS deployment strategies used in our experimental set-up, we provide Table 3, which gives a summary of the characteristics, limitations and benefits of the experimented different QoS deployment strategies, namely both dedicated and higher scheduling rate policies that provide cellular services to UEs.

Note that in our analysis results given in the next section, we have not run simulations of the proposed heuristic scheduling algorithm to compare it with the optimal solution of the original optimization problem given in (8). Our contributions are mainly focusing on experimental analysis and results as opposed to pure simulation-based analysis results.

Note that experimental works that are performed in real live networks of operators on differentiated QoS trials in LTE networks are not common in the literature and is not a trivial task especially when real users are using the existing operational infrastructure. Additionally, other practical restrictions such as the hardware/software limitations, regulatory restrictions (compliance requirements, customized tariffs to different users, etc), security, marketing demands (price per differentiated users, service usage needs, etc.) need to be taken into account in practical QoS deployments and algorithm design. Hence, results and related experiences (main takeaways discussions, lessons learned, trade-off analysis or challenges experienced) on field trials are quite valuable insights into investigation of the achievable performances of different QoS deployment strategies in parallel with UEs having different QoS policies under real conditions and with real equipment limitations.

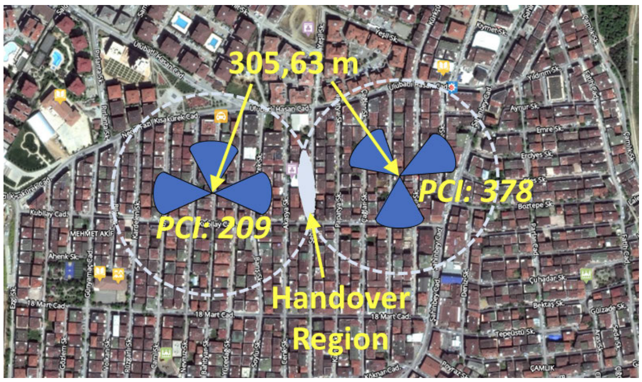


FIGURE 2. The location of the BS, handover region and coverage areas where cell center and cell edge UEs' KPIs (for normal and both prioritized UEs) are observed.

IV. PERFORMANCE ANALYSIS

A. DETAILS OF THE EXPERIMENTAL SETUP

Fig. 2 shows the schematic illustration of the network topology (the locations of LTE networks' cell-edge and cell-center test sites) used throughout the real network experiments in Cekmekoy region of city of Istanbul in Turkey. The experiments were run during different times in two days (ranging between 14:50 to 23:40 local time). Two feature enabled cells namely $PCI-209$ and $PCI-378$ are used for observations. Normally, for Reference Signal Received Power (RSRP) values below −65 dBm LTE users are considered to be at near distance locations to connected eNodeB. For ranges between −75 dBm and −85 dBm LTE users are at middle distance locations to connected eNodeB (practical value is around −80 dBm or slightly better and this is also good for $f = 800$ MHz (low bandwidth) conditions)). For ranges between −100 dBm and −120 dBm, LTE users are considered to be in far (edge) distance to connected eNodeB (practical value is −105 dBm or −110 dBm where noise limited conditions exist for f = 1800 MHz (high bandwidth)).

Our considered scenario for the experimental set-up is as follows: 11 monitored UEs are connecting to eNodeB sequentially in which there are three types of UEs with different QoS requirements. These types are Normal UEs, prioritized UEs with dedicated bandwidth (shortly named as dedicated BW UE) and prioritized UEs (shortly named as priority UE) with high scheduling rate. In addition 11 monitored UEs, there are also normal commercial UEs in the real network connected to this site whose throughput values are not monitored but their presence has direct effect on observed throughput of both priority, dedicated BW and normal UEs.

During our experiments, we have configured prioritized UEs to be $k = 1.5$ times higher scheduling rate than normal UEs. To create priority UEs to be $k = 1.5$ times higher scheduling rate, a different non-GBR QCI profile is defined in HSS and those users are statically assigned to this QCI level. This QCI value has a higher priority than the QCI value of normal UEs. In RAN, eNodeB is configured by applying a weight ratio of $w_i^t/w_j^t = 1.5, \ \forall i \in \mathcal{N}_p^s, \forall j \in \mathcal{N}_n, \forall t \in \mathcal{T}$ resource allocation inside PF scheduler when E-UTRAN Radio Access Bearer (E-RAB) is established. Therefore, resource allocations considering the QCI values are also taken into account in eNodeB at the same time. As part of experimental UE equipment, we have used 10 QCI-6 SIM cards assigned as dedicated bandwidth UEs, 10 QCI-7 SIM cards assigned as priority UEs, 1 QCI-8 SIM card assigned as normal UEs with 11 identical brand UE terminals.

In all the experimental tests, eNodeBs are configured to operate in f = 1800 MHz carrier frequency with 20 Mhz bandwidth at cell centers and in f = 800 Mhz carrier frequency in cell edges with 10 MHz bandwidth. Therefore, for requirements of dedicated BW UEs, $K_f = 20$ RBs is selected for f = 1800 Mhz and $K_f = 10$ RBs is selected for f = 800 Mhz. During experiment, UEs download data via FTP server as shown in Fig. 1. Then in all test scenarios, while normal UEs are downloading data, $N_p = 10$ priority test UEs (depending on UE type) enter into the cell area and download data simultaneously under the same connected eNodeB. After a certain amount of test period, priority UEs quit the cell. Depending on the test scenario, those normal and priority UEs may be on the cell center or cell edge. During our experiments, we have configured dedicated BW UEs bandwidth requirements to be $K_f = 20$ RBs. All experiments are done in full-buffered traffic mode to force the scheduler of eNodeB work in full performance capacity. Hence, high-traffic areas are selected for experiments. Content size of 5 Gbytes for UL are used via FTP for demonstrating the UL throughput variations. In summary, the system level parameters used throughout the experiments are detailed in Table 4.

B. EXPERIMENTAL RESULTS

In this section, we present some of the experimental evaluation results to provide end-to-end QoS support for LTE UEs.

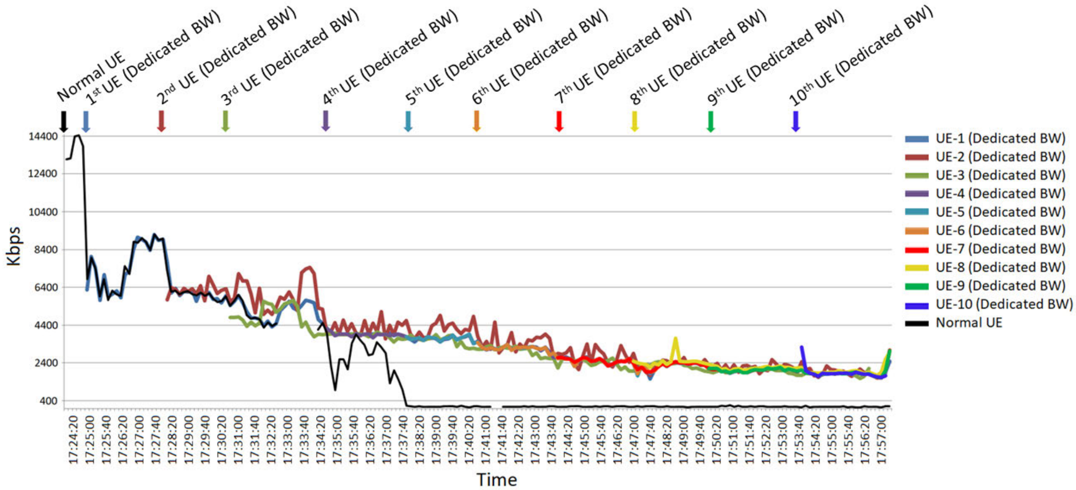
TABLE 4. Experimental parameters and their corresponding values.

Parameter	Value
N	≥ 21
N_p^d	10
N_p^s	10
N_n	≥ 1
M	100
K_f	20 RBs (f=1800 Mhz) 10 RBs (f=800 Mhz)
α	75,326 (bits)
k	1.5
T	1000 msec
τ	50 msec
Δt	1 msec
f (Carrier Freq.)	f = 800 MHz f = 1800 MHz
System Bandwidth	20 MHz
Cells	PCIs: 209 and 378
LTE Duplex Mode	FDD
eNodeB Max. Power	46 dBm
Min. RSRP (Reference Signals Received Power	-130 dBm
Inter-cell distance	305.63 m

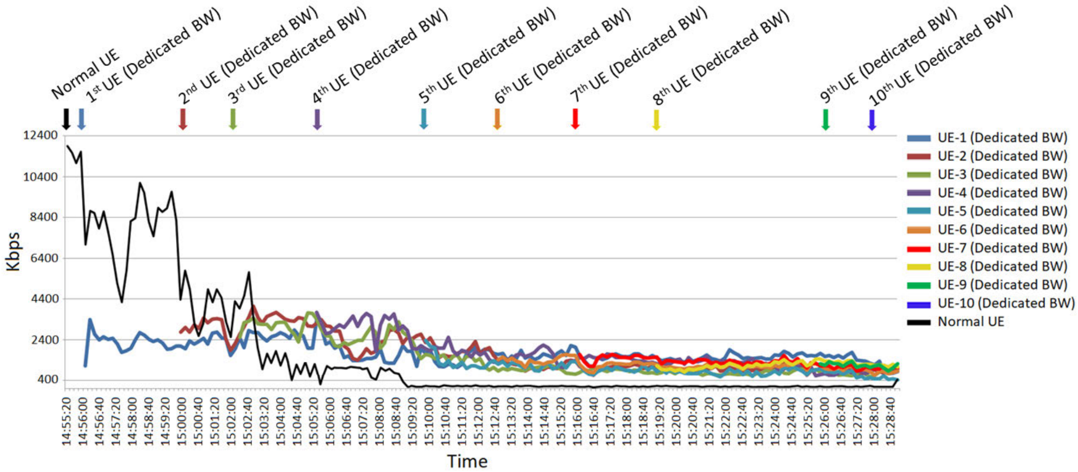
In our experiments, we only show achieved experimental throughput results in UL direction without loss of generality using the scheduling algorithm in Algorithm 1.

Fig. 3 shows the UL performance of UEs in two different experimental scenarios to enable QoS support in live LTE networks. In Fig. 3a, all monitored UEs are in cell center whereas in Fig. 3b normal UE is in cell center and $N_p^d = 10$ dedicated BW UEs are in cell edge. In all of two scenarios, at first while normal UE is generating traffic, 10 more dedicated BW UEs are starting to generate traffic one-by-one sequentially. In Fig. 3a, it is observed that all of dedicated BW UEs' UL throughput values stay above 4 Mbps until the arrival of sixth dedicated BW UE (where no free RBs can be obtained from normal UEs since $K_f = 20$ RBs) whereas normal UE throughput values diminish to zero value after fifth dedicated BW UE enter into the coverage area. Note that dedicated BW UEs get lower RBs than minimum required RB of $K_f = 20$ due to non-availability of free RBs allocated by normal UEs. For this reason, dedicated BW UEs share RBs with other dedicated BW UEs as observed in Fig. 3a. Most of the throughput values of all dedicated BW UEs are slightly lower than 2.4 Mbps at the end of experimental observation period in Fig. 3a where each dedicated BW UE obtains $K_f = M/N_p^d = 10$ RBs since $N_p^d = 10$. In Fig. 3b, we can observe that after all dedicated BW UEs are put into cell edge, the amount of dedicated resources to normal UEs diminishes fast. This is due to low signal quality dedicated BW UEs entering into the coverage area of the cell and suppressing RBs utilization of normal UE. Normal UEs' throughput value decreases after arrival of fifth dedicated BW UE into the cell.

The mobility of users is one of the dimensions that can have an impact on the QoS of UEs. High mobility scenarios can make resource allocations more challenging and cause



(a)



(b)

FIGURE 3. **UL performance of users (a) All users are in cell center (both $f = 1800$ MHz) (b) Normal user is in cell center ($f = 1800$ MHz), users with minimum dedicated bandwidth are in cell edge ($f = 800$ MHz) [Figures are best viewed on colors].**

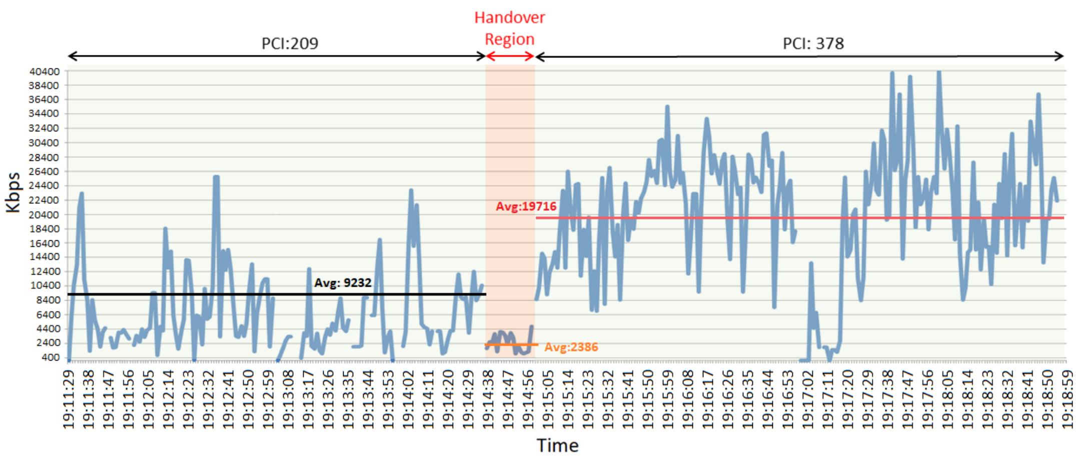
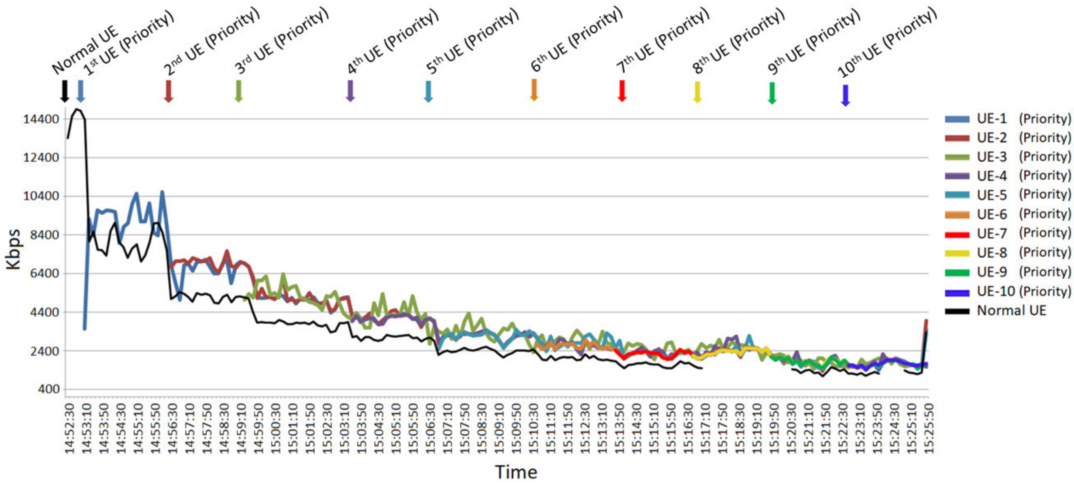
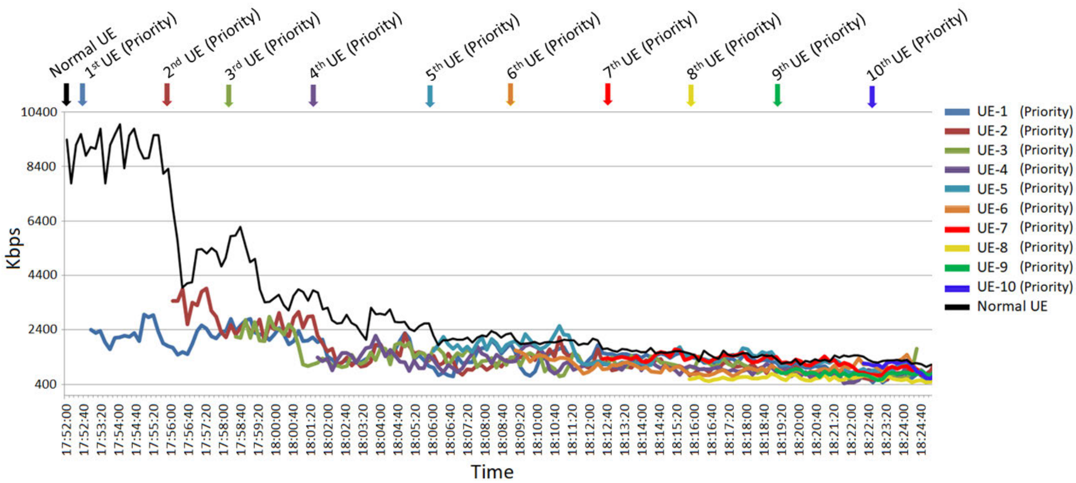


FIGURE 4. **Throughput change of a mobile dedicated BW UE during inter-eNodeB handover ($f = 1800$ MHz).**

service disruption when providing real-time services for some users. Our experimental results also take into account the mobility of the users (mostly pedestrian UEs) during real-world deployment of the QoS services. To experiment with mobility behaviour of dedicated BW UEs, Fig. 4 shows the throughput variation of a dedicated BW UE during inter



(a)



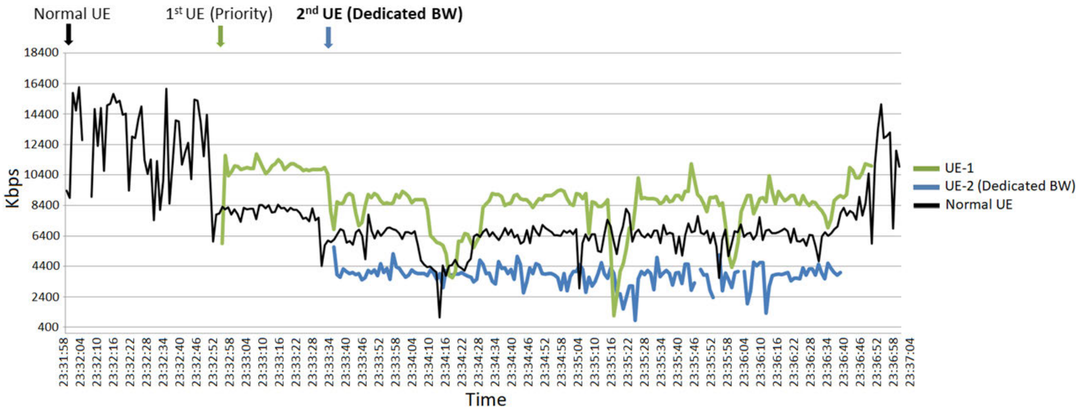
(b)

FIGURE 5. **UL performance of users (a) All UEs are in cell center (both $f = 1800$ MHz) (b) Normal UE is in cell center ($f = 1800$ MHz), priority users are in cell edge ($f = 800$ MHz) [Figures are best viewed on colors].**

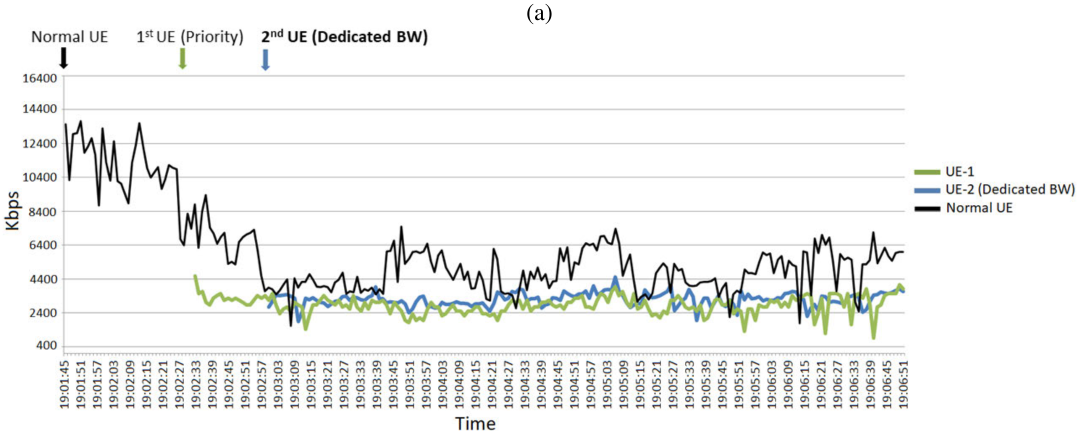
eNodeB handover between two feature-enabled BSs. Dedicated BW UE can obtain more than two times the throughput in $PCI - 378$ cell (with average throughput value of 19716 kbps) compared to $PCI - 209$ cell (with average throughput value of 9232 kbps). This is due to the fact that there are high number of dedicated BW UEs in $PCI - 209$ cell in addition to normal UEs, which increase the amount of traffic in buffer of the scheduler. In comparison, $PCI - 378$ cell has less number of dedicated BW UEs, so that dedicated BW UEs can get higher dedicated bandwidth. On the other hand, the throughput of dedicated BW UE diminishes to average value of 2386 kbps during handover where the dedicated RB allocation requirement is violated significantly. These momentary changes in dedicated BW UE throughput values have demonstrated how performance of mobile dedicated UE between cells can differ as a result of the availability of UE load with different QoS deployment strategies inside the cell.

Fig. 5 shows the experimental scenario where there are one normal UE and $N_p^s = 10$ priority UEs that obtain throughput values based on (8). Similar to previous scenario, while normal UE is generating traffic, 10 higher priority UEs are starting to generate traffic sequentially in time. Fig. 5a shows the scenario where all UEs are in cell center whereas Fig. 5b shows the scenario when normal UE is in cell center and 10 priority UEs are in cell edge. From Fig. 5a, it is observed that the expected theoretical 40% to 60% throughput ratio split between normal and priority UEs respectively has been achieved in this experimental set-up. From Fig. 5b, we can observe that after all priority UEs are on cell-edge, normal UEs' throughput values remain higher than the rest of the UEs due to proximity to eNodeB. The throughput values of priority UEs have diminished, hence the expected 40% to 60% throughput ratio split has not been achieved. These results indicate that even though theoretical throughput split can be achieved in cell-center scenarios, due to poor channel conditions in cell-edge, the designed scheduler performance cannot achieve a successful throughput split between normal and priority UEs.

Fig. 6 shows the amount of throughput generated by three UE types namely normal, priority and dedicated BW UE



(a)



(b)

FIGURE 6. **UL performance of UEs (a) All UEs are in cell center (both $f = 1800$ MHz) (b) Normal UE is in cell center ($f = 1800$ MHz), 1 priority UE and 1 dedicated BW UE are in cell edge ($f = 800$ MHz) [Figures are best viewed on colors].**

over the experiment duration. All UEs are located in the cell center in Fig. 6a whereas normal UE is in cell center and other UEs (namely dedicated BW and priority ones) are located at cell edge in Fig. 6b. First of all, it can be observed that once the priority UE enters into the cell, the throughput values of the normal UE drop accordingly as shown in both figures of Fig. 6. As expected in Fig. 6a, the UL throughput values of priority UE are higher than normal UE. The expected throughput split of 40% to 60% ratio among normal and prioritized UE (UE-1) respectively has been partially achieved where deep fades in normal UE also effects priority UEs' throughput values in consecutive time intervals due to adaptation of scheduler allocations. Transmit buffer size of the different types of UEs can also impact the obtained throughput values. Additionally, dedicated UE (UE-2) has achieved less throughput compared to normal and priority UE due to poor channel conditions even though it can obtain target of 20 RBs utilization.

In Fig. 6b, both dedicated BW and priority UEs have obtained approximately the same UL throughput values which is below 4 Mbps. On the other hand, normal UE has obtained the highest throughput value since priority UEs are now located at cell edge and normal UE is in cell center. Moreover, cell center UE is using $f = 1800$ MHz whereas cell edge UEs are now using $f = 800$ MHz (due to higher coverage potential at large distances) which also has an effect in throughput reductions. This scenario is planned by network operations experts to provide connectivity rather than higher data rates at far locations to BSs. Even though the scheduler is configured to distribute resources 3 to 2 ratio among priority and normal UEs respectively, the obtained throughput values in Fig. 6b differ due to low RSRP values for priority UE. Note that the eNodeB scheduler considers different metrics including MCS index values. prior obtained throughput values, etc. during resource allocation in addition to statically assigned weight configuration defined in Section III-B. Moreover, dedicated BW UE has achieved the same minimum required RBs utilization of $K_f = 20$. However, due to poor channel conditions Fig. 6b's throughput values are low compared to Fig. 6a throughput values.

Fig. 7 shows the experimental scenario when there are one normal UE, $N_p^s = 5$ priority UEs and $N_p^d = 5$ dedicated BW UEs inside cell coverage. As before, 5 priority UEs and 5 dedicated BW UEs start to upload traffic inside the considered cell sequentially in time. Fig. 7a demonstrates the scenario when all UEs are in cell center with good RF conditions whereas Fig. 7b shows the scenario where normal UE is in cell center, 5 priority UEs and 5 dedicated BW UEs

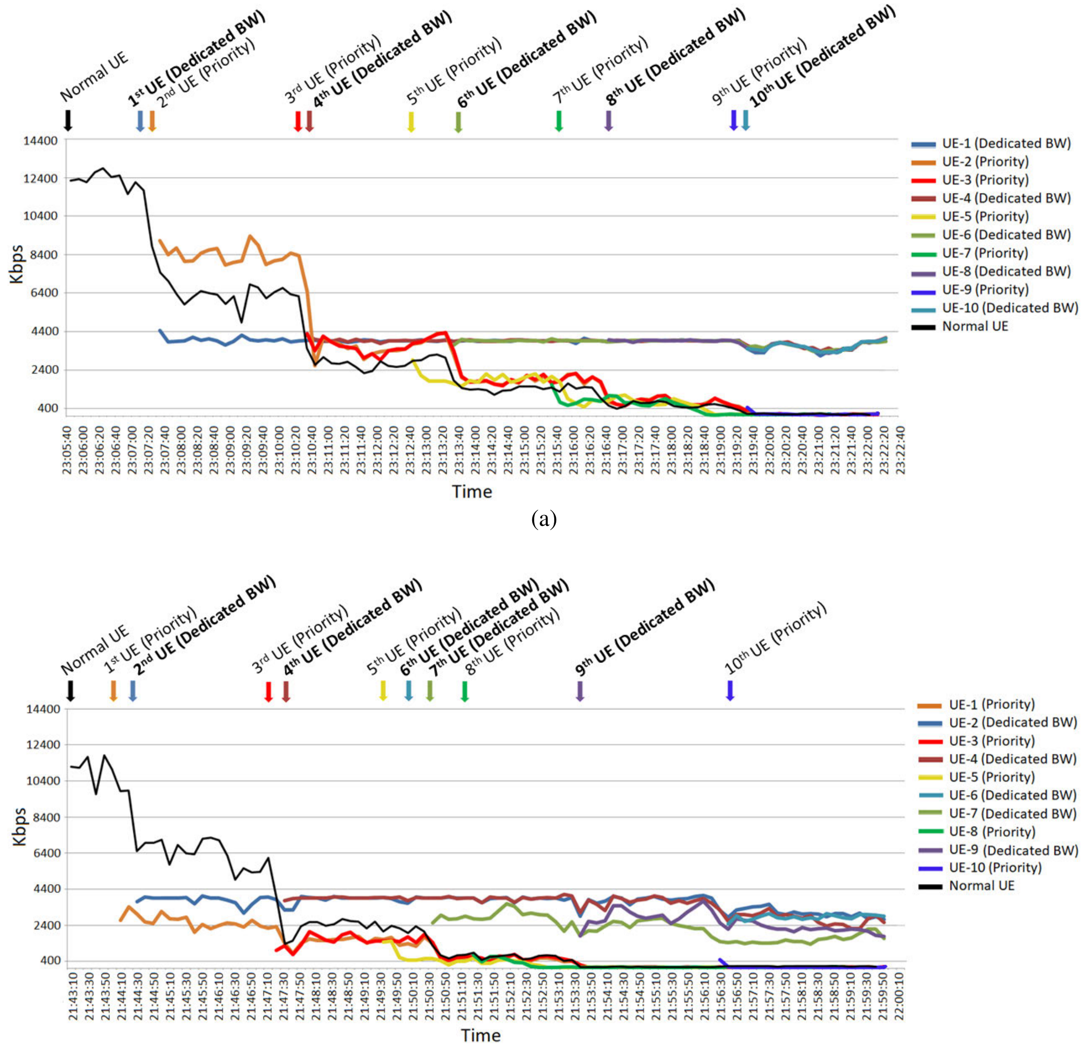


FIGURE 7. UL performance of UEs (a) All UEs are in cell center (both $f = 1800$ MHz) (b) Normal UE is in cell center ($f = 1800$ MHz), 5 priority UEs and 5 dedicate BW UEs are in cell edge ($f = 800$ MHz) [Figures are best viewed on colors].

are at cell edge. From Fig. 7a, we can observe that throughput values of dedicated BW UEs are around 4 Mbps. We can also observe that as number of UEs increases in time and more specifically after UE-6 (dedicated BW UE) enters into cell, normal and priority UEs's throughput values start to diminish and become zero value after dedicated BW UE-10 (fifth dedicated UE) enters into coverage area. This is due to lack of resources, i.e. unavailability of RBs, which signifies that LTE eNodeB cannot schedule resources for these normal and priority UEs. Hence, the expected throughput split between normal and priority UEs of 40% to 60% ratio respectively has not been achieved in this experimental scenario even though the channel quality is good for all the UEs. Fig. 7b shows that after all priority and dedicated BW UEs are located at cell-edge, dedicated BW UEs start to obtain less throughput values (e.g. after UE-7 (dedicated BW) enters). After UE-10 (priority) enters, throughput of all dedicated BW UEs start to diminish together. On the other hand, the performance of normal and priority UEs becomes worse than the scenario in Fig. 7a where the data rates diminish significantly after UE-5 (priority) enters. After this point, normal and priority UEs obtain no resources as UE-9 (dedicated BW) is connected to the same cell. In summary, the results in Fig. 7 indicate that the diminishing effect of lack of radio resources can be observed directly on priority UEs and normal UEs rather than dedicated BW UEs.

C. MAIN OBSERVATIONS AND TAKEAWAYS

In summary, we have tested three main types of experimental scenarios using the experimental set-up. First one is with priority and normal UEs, second one is with dedicated BW and normal UEs and final one is with priority, dedicated BW and normal UEs. Our experimental observations have revealed that location of the UE with respect to BS and the availability of dedicated BW UEs inside cell may have implications on the apriori defined resource allocation strategies of the other UEs. For example, the results in Fig. 5 indicate that even though theoretical throughput split can be achieved in cell-center scenarios, due to poor channel conditions in cell-edge, the designed scheduler performance cannot achieve a successful throughput split between normal and priority UEs even though dedicated BW UE have obtained lower throughput values. This is also true in poor channel conditions of priority UE as observed from experimental results of Fig. 6b. These results signify that before deploying critical or non-critical services in a cellular network, network planning and optimization should consider the available number of diverse set of UEs with different QoS requirements in the surrounding area meticulously to avoid reaching limitations on eNodeB capacity.

Another important observation to consider is that dedicated BW UEs should be placed in cell-center locations after careful network and coverage planning. In case dedicated BW UEs are in cell-edge areas, they can have diminishing performances on dedicated BW UEs, but will also have huge impact on normal and priority UEs that are located in cell center regions by allocating their RBs in exchange of poor throughput values. Dedicated BW UEs is primarily designed for static UEs such as ATMs of a bank. In case, this feature is enabled for mobile dedicated BW UEs, they can have detrimental impact on new cells in case normal and priority UEs exist. Moreover, dedicated BW UEs can experience severe throughput decrements during handover period due to not completed RB allocation strategies. If a mission-critical network service is running over mobile dedicated BW UEs, these handover interruptions can disrupt the service continuity. It has also been observed in UL traffic tests that the BS is scheduling rate for a UE depends on the transmit buffer size of the UE. Therefore, UEs with high buffer size will be scheduled further. The difference between the theoretically calculated maximum UE throughput amount and the throughput that the UE can actually practically achieve is created by the transmit buffer size of the UE. Furthermore, if the buffer sizes of the priority and normal UEs are different, this would be an advantage for UEs with high buffer size in terms of obtained throughput values and the targeted 60% to 40% ratios would not be provided.

If there were no dedicated BW UEs in the system design and only UEs with different priorities were present, the resource allocation problem would be solved by prioritizing these UEs and their services simply based on priority ordering during scheduling interval. However, the problem arises on MNO's concern about UEs with dedicated bandwidth requirements. To provide dedicated bandwidth, MNO assigns the dedicated BW UE to a higher priority QoS, i.e. QCI priority higher than the priority UEs. Although this is positive in terms of providing dedicated bandwidth, the dedicated BW UEs can deplete resources of the other UEs. Therefore, an upper bound limitation on number of connected dedicated BW UEs is needed by careful radio network capacity planning at each site. However, this situation cannot be prevented in case there is a single-tier scheduler in the system. If a multi-tier scheduler was present, each UE type could be scheduled on its own tier. However, in this case the complexity of the system would increase. The network slicing concept that comes with 5G networks can actually be a suitable solution for the problem presented in this paper. Dedicated slicing, a deployment method of network slicing, can be implemented according to different QoS strategies since it can have a separate scheduler for each slice type.

V. CONCLUSION AND FUTURE WORK

In this paper, we have investigated a potential solution of providing better QoS support for differentiated types of UEs to improve the performance of the services provided by MNOs. For this, we first formalized the optimization problem in a formal manner. Later, we described our experimental set-up where experimental evaluation of different QoS deployment strategies are performed in a real operational network in Turkey. We have also analyzed the necessary network planning in detail for deploying the proposed QoS strategies. Our real-world experiments on a LTE network indicated that prioritized UEs with dedicated bandwidth have higher precedence compared to prioritized UEs with high scheduling rate and normal UEs. Moreover, the theoretical 40% to 60% ratio of throughput split between prioritized UEs with high scheduling rate and normal UEs can only be achieved as long as the amount of dedicated resources allocated to dedicated BW UEs is carefully planned during network capacity optimization stage. This signifies that before deployment of QoS policies for any critical or non-critical services, MNOs need to perform extensive experimental trials to find the best configuration and optimization parameters to serve all UEs based on their assigned QCI levels. As a result, MNOs can only gain major benefits by differentiating priority UEs and diversifying the network services provided for their UEs via appropriate network planning. For future study, the field KPIs and service use cases can be analyzed with Machine Learning (ML) and QoS deployment strategy can be changed dynamically. Moreover, noting that we have only considered a single shared S-GW in our both optimization problem and experimental scenario, a more general setting in large-scale deployments could be that multiple such S-GWs can be utilized for sharing similar resources among differentiated UEs. This extension of the model would lead to a multi-layer problem and would be an interesting future work direction.

REFERENCES

[1] The 5G Infrastructure Public Private Partnership (5G PPP). (2015). *5G Vision: The 5G Infrastructure Public Private Partnership: The Next-Generation of Communication Networks and Services*. Accessed: Sep. 2020. [Online]. Available: https://bit.ly/347XjDx

[2] C. Joe-Wong, L. Zheng, S. Ha, S. Sen, C. W. Tan, and M. Chiang, *Smart Data Pricing in 5G Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017, pp. 478–500.

[3] V. W. S. Wong *et al.*, "Overview of new technologies for 5G systems," in *Key Technologies for 5G Wireless Systems*. 2017, p. 1.

[4] W. Diego, I. Hamchaoui, and X. Lagrange, "Cost factor analysis of QoS in LTE/EPC mobile networks," in *Proc. 13th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2016, pp. 614–619.

[5] R. Zhang, M. Wang, X. Shen, and L.-L. Xie, "Probabilistic analysis on QoS provisioning for Internet of Things in LTE–A heterogeneous networks with partial spectrum usage," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 354–365, Jun. 2016.

[6] Y. Qi and H. Wang, "QoS-aware cell association based on traffic prediction in heterogeneous cellular networks," *IET Commun.*, vol. 11, no. 18, pp. 2775–2782, Dec. 2017.

[7] W. Li and J. Zhang, "Cluster-based resource allocation scheme with QoS guarantee in ultra-dense networks," *IET Commun.*, vol. 12, no. 7, pp. 861–867, Apr. 2018.

[8] X. Zhang, W. Cheng, and H. Zhang, "Heterogeneous statistical QoS provisioning over 5G mobile wireless networks," *IEEE Netw.*, vol. 28, no. 6, pp. 46–53, Nov. 2014.

[9] X. Zhang and J. Wang, "Statistical QoS-driven power adaptation for distributed caching based mobile offloading over 5G wireless networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 760–765.

[10] M. Hoyhtya, K. Lahetkangas, J. Suomalainen, M. Hoppari, K. Kujanpaa, K. Trung Ngo, T. Kippola, M. Heikkila, H. Posti, J. Maki, T. Savunen, A. Hulkkonen, and H. Kokkinen, "Critical communications over mobile Operators' networks: 5G use cases enabled by licensed spectrum sharing, network slicing and QoS control," *IEEE Access*, vol. 6, pp. 73572–73582, 2018.

[11] A. S. M. Z. Shifat, M. Z. Chowdhury, and Y. M. Jang, "Game-based approach for QoS provisioning and interference management in heterogeneous networks," *IEEE Access*, vol. 6, pp. 10208–10220, 2018.

[12] K. Hammad, A. Moubayed, S. L. Primak, and A. Shami, "QoS-aware energy and jitter-efficient downlink predictive scheduler for heterogeneous traffic LTE networks," *IEEE Trans. Mobile Comput.*, vol. 17, no. 6, pp. 1411–1428, Jun. 2018.

[13] G. H. S. Carvalho, I. Woungang, A. Anpalagan, and E. Hossain, "QoS-aware energy-efficient joint radio resource management in multi-RAT heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6343–6365, Aug. 2016.

[14] J. Liu, X. Xu, W. Chen, and Y. Hou, "QoS guaranteed resource allocation with content caching in SDN enabled mobile networks," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC Workshops)*, Jul. 2016, pp. 1–6.

[15] P. Vizarreta, M. Condoluci, C. M. Machuca, T. Mahmoodi, and W. Kellerer, "QoS-driven function placement reducing expenditures in NFV deployments," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[16] M. Khan, R. S. Alhumaima, and H. S. Al-Raweshidy, "QoS-aware dynamic RRH allocation in a self-optimized cloud radio access network with RRH proximity constraint," *IEEE Trans. Netw. Service Manage.*, vol. 14, no. 3, pp. 730–744, Sep. 2017.

[17] F. Guo, H. Lu, D. Zhu, and H. Wu, "Interference-aware user grouping strategy in NOMA systems with QoS constraints," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr. 2019, pp. 1378–1386.

[18] X. Chen, Z. Li, Y. Chen, and X. Wang, "Performance analysis and uplink scheduling for QoS-aware NB-IoT networks in mobile computing," *IEEE Access*, vol. 7, pp. 44404–44415, 2019.

[19] P. Ameigeiras, J. Navarro-Ortiz, P. Andres-Maldonado, J. M. Lopez-Soler, J. Lorca, Q. Perez-Tarrero, and R. Garcia-Perez, "3GPP QoS-based scheduling framework for LTE," *EURASIP J. Wireless Commun. Netw.*, vol. 2016, no. 1, p. 78, Dec. 2016.

[20] *User Equipment (UE) Radio Access Capabilities*, Standard TS 36.306 V16.0.0 (Release 16), 3GPP Technical Specification Group Radio Access Networks, 2020. Accessed: Jul. 2020. [Online]. Available: https://bit.ly/2S0EFaZ

[21] J. Wannstrom. (2013). *LTE-Advanced*. Accessed: Jan. 2020. [Online]. Available: https://bit.ly/2uEINFq

[22] C. Deniz, O. G. Uyan, and V. C. Gungor, "On the performance of LTE downlink scheduling algorithms: A case study on edge throughput," *Comput. Standards Interface*, vol. 59, pp. 96–108, Aug. 2018.

[23] X. Wang and L. Cai, "Limiting properties of overloaded multiuser wireless systems with throughput-optimal scheduling," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3517–3527, Oct. 2014.

[24] O. Narmanlioglu and E. Zeydan, "Software-defined networking based network virtualization for mobile operators," *Comput. Electr. Eng.*, vol. 57, pp. 134–146, Jan. 2017.

[25] M. Solera, M. Toril, I. Palomo, G. Gomez, and J. Poncela, "A testbed for evaluating video streaming services in LTE," *Wireless Pers. Commun.*, vol. 98, no. 3, pp. 2753–2773, Feb. 2018.

[26] V. Karyotis, M. Avgeris, M. Michaloliakos, K. Tsagkaris, and S. Papavassiliou, "Utility decisions for QoE-QoS driven applications in practical mobile broadband networks," in *Proc. Global Inf. Infrastruct. Netw. Symp. (GIIS)*, Oct. 2018, pp. 1–5.

[27] T. D. Assefa. (2015). *QoS Performance of LTE Networks With Network Coding: Master Thesis, Norwegian University of Science and Technology*. Accessed: Nov. 2019. [Online]. Available: https://bit.ly/3mMY5yi

[28] X. Zhang, J. Wang, and H. V. Poor, "Heterogeneous statistical-QoS driven resource allocation over mmWave massive-MIMO based 5G mobile wireless networks in the non-asymptotic regime," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 12, pp. 2727–2743, Dec. 2019.

[29] N. Dandanov, S. R. Samal, S. Bandopadhaya, V. Poulkov, K. Tonchev, and P. Koleva, "Comparison of wireless channels for antenna tilt based coverage and capacity optimization," in *Proc. Global Wireless Summit (GWS)*, Nov. 2018, pp. 119–123.

[30] Y. Turk, E. Zeydan, and C. A. Akbulut, "An experimental analysis of differentiated quality of service support for LTE users," in *Proc. 5th Int. Conf. Electr. Electron. Eng. (ICEEE)*, May 2018, pp. 408–412.

[31] *System Architecture for the 5G System (Release 15)*, Standard TS 23.501 version 15.2.0, 3GPP Technical Specification Group Services and System Aspects, 2020. Accessed: Jul. 2020. [Online]. Available: https://bit.ly/3jcnbop

[32] *LTE E-UTRA Physical Layer Procedures*, Standard TS 36.213 version 12.3.0 (Release 12), 3GPP Technical Specification Group Radio Access Networks, 2014. Accessed: Jul. 2020. [Online]. Available: https://bit.ly/3j1nhil

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

[34] M. Mohseni, S. A. Banani, A. W. Eckford, and R. S. Adve, "Scheduling for VoLTE: Resource allocation optimization and low-complexity algorithms," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1534–1547, Mar. 2019.

[35] G. Yuan and B. Ghanem, "Binary optimization via mathematical programming with equilibrium constraints," 2016, *arXiv:1608.04425*. [Online]. Available: http://arxiv.org/abs/1608.04425

[36] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Commun. Lett.*, vol. 9, no. 3, pp. 210–212, Mar. 2005.

[37] *GPRS Enhancements for E-UTRAN Access*, Standard TS 23.401 V8.12.0 (Release 8), 3GPP Technical Specification Group Services and System Aspects, 2010. Accessed: Feb. 2017. [Online]. Available: https://bit.ly/33Tx7wy

[38] D. Dean and C. Riley, "Data services over G. SHDSL transport infrastructure," U.S. Patent 11 412 455, Nov. 1, 2007.

ENGIN ZEYDAN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Department of Electrical and Electronics Engineering, Middle East Technical University, Ankara, Turkey, in 2004 and 2006, respectively, and the Ph.D. degree from the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ, USA, in February 2011. He has worked as a Research and Development Engineer for Avea, a Mobile Operator in Turkey, from 2011 to 2016. He was also a part-time Instructor with the Electrical and Electronics Engineering Department, Ozyegin University, from 2015 to 2018. He was with Turk Telekom Labs working as a Senior Research and Development Engineer from 2016 to 2018. He is currently a Senior Researcher with the Communication Networks Division, Centre Tecnológic de Telecomunicacions de Catalunya (CTTC). His research interests include telecommunications and big data networking. He received the Best Paper Award from the Network of Future Conference in 2017.

JOSEP MANGUES-BAFALLUY received the degree and the Ph.D. degree in telecommunications engineering from UPC, in 1996 and 2003, respectively. He was the Vice-Chair of the IEEE WCNC, Barcelona, in 2018. He was also a Researcher and an Assistant Professor with UPC. He is currently a Senior Researcher and the Head of the Communication Networks Division, Centre Tecnológic de Telecomunicacions de Catalunya (CTTC), Barcelona. He has participated in various roles (including leadership) in several public funded and industrial research projects, such as 5GPPP 5Growth, 5G-Transformer, or Spanish 5G-REFINE. His research interests include NFV applied to mobile networks and autonomous network management.

OMER DEDEOGLU received the B.S. degree in electrical and electronics engineering from Bilkent University, in 2001, and the M.S. degree in electrical and computer engineering from New Mexico University, in 2003. He worked for Research and Development projects and made Radio NW investment plans at Turkcell for about six years. Since 2011, he has been working with Türk Telekom as the Radio Network Planning Manager.

YEKTA TURK received the B.Sc. degree in electrical and electronics engineering from Anadolu University, Turkey, in 2005, the M.Sc. degree in telecommunications and computer networks from George Washington University, Washington, DC, USA, in 2007, and the Ph.D. degree from the Department of Computer Engineering, Maltepe University, Istanbul, Turkey, in 2018. He is currently a Mobile Network Architect-Based in Istanbul, Turkey. His research interests include mobile radio telecommunications and computer networks.