IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 3, 2020, accepted April 21, 2020, date of publication April 27, 2020, date of current version May 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2990381

A Hybrid Downlink Scheduling Approach for Multi-Traffic Classes in LTE Wireless Systems

MOUSTAFA M. NASRALLA, (Member, IEEE)
Department of Communications and Networks Engineering, Prince Sultan University, Riyadh 11586, Saudi Arabia
e-mail: mnasralla@psu.edu.sa

ABSTRACT The developments in wireless technology and applications in recent years have increased the interest in downlink scheduling and resource allocations among researchers. Moreover, fair scheduling and balanced Quality of Service (QoS) delivery for various forms of traffic are needed for Long-Term Evolution (LTE) wireless systems. This paper proposes hybrid QoS-aware downlink scheduling approaches that aim to address different traffic classes and balance the QoS delivery with improvements to the overall system performance under channel and bandwidth constraints. Moreover, this research introduces a taxonomy that classifies the scheduling algorithms into four main classes: delay aware, queue aware, target bit-rate aware and hybrid aware. The latter class is the scheduling class that is proposed in this paper; it considers channel, queue and delay parameters in its scheduling metric. Using simulations, we compare and analyze different downlink scheduling rules for their network-centric performance metrics, e.g., average packet loss ratio, average throughput, average packet delay, system fairness, and system spectral efficiency. The simulation results show that the queue-aware and delay-aware scheduling rules deliver the best QoS performance for video traffic classes, whereas our proposed hybrid scheduling rules deliver balanced QoS for various types of traffic classes. Employing QoS balancing scheduling rules in an LTE downlink is suggested to provide high QoS delivery for different traffic classes.

INDEX TERMS Packet scheduling algorithms, resource allocation, long term evolution, quality of service, real-time, non-real time.

I. INTRODUCTION

Several challenges have to be overcome to support multimedia services and applications over wireless networks. These challenges are mostly caused by heterogeneities and constraints, for example, random time-varying channel conditions, limited bandwidth, different protocols and standards, limited battery power, and differing QoS requirements. In response to these challenges and the increasing demand for network applications, which have varied requirements, for example, mobile TV, teleconferencing and multimedia messaging, the Third Generation Partnership Project (3GPP) introduced Long-Term Evolution (LTE). LTE is a promising mobile technology that permits the transfer of multimedia applications with high network capacity and utility. LTE employs Orthogonal Frequency Division Multiple Access (OFDMA) as a radio access technology in the downlink channel, which provides greater flexibility and optimal network performance, as it contiguously uses sections of the spectrum. To provide the necessary bandwidth and acceptable delays, resource allocation algorithms are implemented by LTE to distribute radio resources. One disadvantage of LTE is that the transmission order is affected by the poorly defined scheduling algorithm problem of distributing radio resources to users. One of the main objectives of 4G LTE radio access networks is to deliver high QoS. Therefore, the performance of the existing radio resource algorithm is reduced under prioritized conditions due to the minimum data rate employed to establish the order of transmission. This paper addresses the problem of scheduling multi-traffic classes to more than one user on the downlink of a wireless network. Two main classifications can be applied regarding scheduling algorithms: QoS-aware/QoS-unaware schedulers and content-aware schedulers. A recent comprehensive survey of downlink content-aware scheduling algorithms is provided in [1]. However, this study lacked information related to the classification of QoS-aware and QoS-unaware schedulers and their performance analysis.

The use of QoS is a network-centric approach to evaluating performance that involves assessing numerous network performance parameters, such as the end-to-end packet delay, average throughput of the system, efficiency of the system,

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

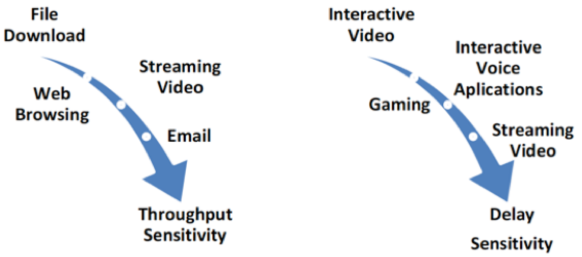


FIGURE 1. Traffic classes based on various QoS requirements.

and fairness and packet loss rate [2]. QoS-aware scheduling approaches consider these parameters and apply an evaluation to optimize the efficiency of wireless systems. These approaches also reliably schedule packets and deliver robust network performance parameters, which translate into excellent QoS to end users. Well-tailored methods exploit the variability in the wireless channel over time and across users. The highest proportion of available resources is assigned to users with excellent channel quality because they are able to handle higher data rates while concurrently ensuring that fairness is sustained across multiple users. Two traffic classes exist: Real Time (RT) and Non-Real Time (NRT), with further classification into Guaranteed Bit Rate (GBR) and Non-Guaranteed Bit Rate (non-GBR). Whether a radio bearer is tagged with GBR or non-GBR depends on the QoS requirements of the flows that they convey. The RT class contains Voice over IP (VoIP) services, video gaming and video conferencing, while the NRT class comprises what are known as best-effort services (e.g., email, browsing the internet, FTP and video streaming). Figure 1 illustrates the different QoS requirements of these applications with regard to throughput and delay. The diagram indicates that RT services are more sensitive to delay, while NRT services are more sensitive to throughput.

None of the scheduling strategies mentioned in the subsequent related work section consider the inherent conflict between QoS-unaware scheduling and QoS-aware scheduling and the possibility of enhancing the QoS of the delivered services (i.e., RT and NRT traffic) by managing this trade-off. Furthermore, the literature lacks proposals of QoS-aware fair scheduling algorithms considering simultaneous transmission of RT and NRT traffic. Hence, a proposal for designing downlink scheduling approaches to balance the QoS metrics for simultaneous transmission of multi-traffic classes over future wireless systems is needed. This paper provides the following three main contributions:

1) Proposing QoS-aware downlink scheduling approaches that balance the QoS parameters for simultaneous transmission of multi-traffic classes. These types of schedulers make scheduling decisions based on varying objective functions (e.g., Head-of-Line (HoL) delay, flow queue sizes, and Channel State Information (CSI)). Their main aim is to simultaneously balance the QoS for RT and NRT users under bandwidth and channel constraints.
2) Proposing a taxonomy that classifies the scheduling algorithms into four key categories: delay-aware, queue-aware, target bit-rate-aware, and hybrid aware strategies. The latter class is the scheduling class proposed in this paper, which considers channel, queue and delay parameters in its scheduling metric. Moreover, the proposed scheduling approaches in the hybrid class contribute to balancing the QoS between both traffic classes.
3) Lastly, providing a benchmark and QoS performance analysis of the existing and proposed downlink scheduling strategies for multi-traffic classes.

The paper's structure is as follows. In Section II, relevant literature is discussed and compared. Section III introduces and classifies the QoS-unaware and QoS-aware scheduling strategies and introduces conceptual description and mathematical equations. The simulation setup and comparative performance analysis are discussed in Section IV, where the system model is elaborated and a description of the 3GPP LTE system is provided. The performance analysis of different scheduling strategies in the designed scenarios is also reported in this section, and a discussion of the relevant performance metrics, numerical results, simulation environment and traffic models are provided. The paper is concluded in Section V.

II. RELATED WORK

Downlink scheduling algorithms are responsible for assigning the physical resource blocks (PRBs) among flows at the Medium Access Control (MAC) layer of Evolved NodeB (eNodeB). For instance, several algorithms for packet scheduling have been proposed with the aim of supporting RT and NRT services over LTE to provide efficient QoS delivery. The LTE scheduling algorithms are designed to handle different types of traffic by considering different QoS parameters, such as delay, packet loss and target rates. Hence, the authors in [1], [3]–[14] proposed diverse scheduling approaches to support either one type of traffic or mixed types of traffic. These algorithms are further investigated in Section III. The authors in [15] evaluated how well commonly employed packet scheduling algorithms performed in downlink LTE systems, namely, the Modified Largest Weighted Delay First (M-LWDF), Proportional Fair (PF) and Exponential Proportional Fair (EXP-PF) schedulers. These algorithms are designed to provide a single QoS-based objective improvement and address one type of traffic every time. The authors in [16], [17] conducted a survey of the downlink QoS-aware and QoS-unaware scheduling approaches in LTE networks, including a performance comparison in terms of QoS provisioning between the most well-known scheduling rules. On the other hand, the authors in [18] conducted a survey of opportunistic scheduling approaches in wireless communications. They classified the opportunistic rules into a taxonomy, where they reported the rules that enhance the total network capacity and those that enhance the QoS objectives, for example, fairness and throughput. These surveys reveal a lack of QoS-aware scheduling algorithms that fairly and simultaneously handle RT and NRT traffic and achieve balanced QoS performance improvement (i.e., multiple objective enhancement).

The authors in [12], [19]–[26] proposed QoS-based downlink packet scheduling schemes for multiple users over wireless networks. The authors in [19] and [20] proposed a multi-service QoS guaranteed scheduling algorithm over wireless private networks. The design of this approach is based on sacrificing the performance of low-priority users to guarantee satisfying the QoS requirements of high-priority users. Note that this approach tends to support one traffic type that has high priority while penalizing the QoS provision to lower priority flows, such as NRT flows. The authors in [21], [22] proposed QoS-aware scheduling algorithms to improve the satisfaction of users in OFDMA systems. These approaches are designed to ensure guaranteed QoS delivery in both uplink and downlink directions, regardless of the traffic type being served, which causes a degradation of the quality of the served traffic due to unbalanced scheduling. The authors in [12], [23] proposed a novel downlink resource allocation algorithm for downlink LTE networks considering the standardized QoS Class Identifier (QCI) requirements for the different types of service. The algorithm provides a way to fulfill each radio bearer's demands; it distributes the available bandwidth such that the bearer traffic and bandwidth requirements are satisfied, and the total throughput in the system is not compromised. Hence, this algorithm utilizes the solution proposed by the 3GPP standard by using the different types of radio bearers: GBR and non-GBR [27], [28]. These approaches are designed to address the issue of satisfying the QoS requirement of VoIP and video traffic (i.e., RT traffic only) by employing the features of QCI and QoS parameters of traffic when allocating RBs. However, these approaches fail to balance the QoS parameters for the simultaneous transmission of heterogeneous traffic.

A QoS-aware downlink scheduling algorithm was proposed in [24] with the objective of improving the QoS experience of edge users in an LTE mobile network. This approach is designed to reduce the effect of low throughput and high delay experienced by these users. Hence, the main goal is to provide edge users with a better QoS experience while simultaneously preventing large losses in the throughput and QoS of the system as a whole. The authors in [25], [26] proposed delay-based and QoS-aware packet scheduling for multimedia services in LTE downlink systems. This approach is designed to guarantee QoS for heterogeneous traffic over (4G) mobile networks under different speed conditions. This approach's principal objective is to investigate the effect of delay on improving the QoS for RT flows in different speed scenarios while giving less importance to other QoS parameters and guaranteeing the minimum QoS for other traffic types, such as NRT. Therefore, we propose a scheduling approach that balances the QoS parameters for multi-traffic classes.

In our previous studies [1], we carried out a comprehensive review of existing content-aware strategies. In addition, we classified content-aware scheduling strategies into three categories as follows: 1) quality-driven scheduling approaches; 2) proxy-driven radio resource allocation approaches; and 3) client-driven approaches. In this paper, we achieve advances by classifying the content-unaware scheduling strategies, i.e., QoS-based scheduling strategies, into QoS-aware and QoS-unaware scheduling categories. Within each category, there are different classes with different purposes. This classification has helped identify a need for network operators by proposing scheduling strategies that would fall under our proposed scheduling class (referred to as the hybrid scheduling strategy class), consisting of strategies for scheduling that have been designed especially for offering simultaneous services to multi-traffic classes. When they make scheduling decisions, these schedulers consider various objective functions (e.g., HoL delay, flow queue sizes, and CSI), where the main goal of these schedulers is to simultaneously balance the QoS for RT and NRT users under bandwidth and channel constraints.

This study proposes two scheduling strategies that fall under the proposed hybrid scheduling class, namely, the Queue-HoL-MLWDF rule and Modified-EXP-rule. To facilitate comprehension of the work that has been carried out, the proposed taxonomy classifies the QoS-based downlink scheduling rules in the literature. This taxonomy consists of our proposed hybrid scheduling class, which includes scheduling strategies that balance the QoS for different traffic classes. We present the results, compare the classified state-of-the-art strategies, and provide a benchmark and QoS performance analysis of different scheduling algorithms for multi-traffic classes. To the best of our knowledge, there is a lack of proposals for designing downlink scheduling approaches with the aim of balancing the QoS metrics for simultaneous transmission of multi-traffic classes over future wireless systems in the literature. Moreover, most of the existing studies mentioned earlier focus on a single QoS-based objective improvement. Note that QoS balancing scheduling strategies should be employed in LTE and beyond wireless systems to offer high QoS delivery for different traffic classes.

III. QoS-AWARE AND QoS-UNAWARE DOWNLINK SCHEDULING STRATEGIES

Following recommendations in recent studies [1], [10], radio resource management and packet scheduling solutions for QoS-based scheduling strategies can be categorized into two broad groups: QoS-aware strategies and QoS-unaware strategies. As illustrated in the proposed taxonomy in Figure 2, the QoS-aware strategy category is divided into four classes and the QoS-unaware strategy category consists of three main scheduling rules, namely, the PF, round robin, and max-rate rules. Moreover, the former category contains four classes: delay-aware strategies, queue-aware strategies, target bit-rate-aware strategies, and the proposed hybrid strategies.

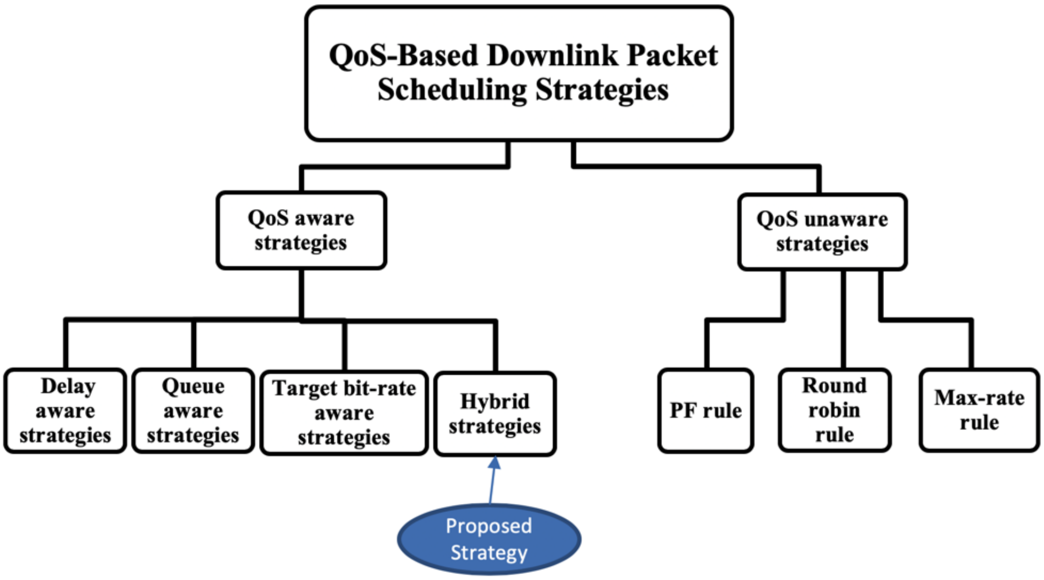


FIGURE 2. QoS-aware and QoS-unaware downlink packet scheduling approach classification.

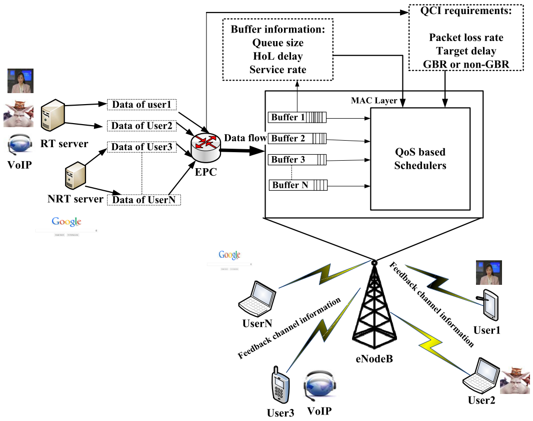


FIGURE 3. QoS-based scheduling strategies for multi-traffic classes [1].

The following subsections provide more detail about each of these classes.

Figure 3 has been provided to illustrate a network scenario in which end-to-end communication occurs across the various layers of an LTE system. Studying this network will help in understanding how QoS-aware scheduling strategies function. As shown in Figure 3, multi-traffic classes (e.g., Video, VoIP, Best effort) are transmitted to the Evolved Packet Core (EPC) from their corresponding servers. Using the S1 interface, the EPC connects with an eNodeB. Packets are managed and the physical resources are assigned to different flows (i.e., NRT and RT) by the LTE protocol's MAC layer. Examples of the RT and NRT classes are provided in the introduction. QoS-aware strategies take into account buffer-related information and QCI requirements, as shown in Figure 3. The QCI is a mechanism in 3GPP LTE systems that ensures that a suitable QoS is provided for traffic flows by controlling the packet loss rate and latency. In the EPC, a QCI is allocated to each individual traffic flow. The QCI is represented by a scalar, 8-bit header field, which determines the packet forwarding rate for each node, for example, by controlling scheduling weights and admission thresholds at the eNodeB. Each type of traffic has different QoS requirements, and thus, different QCI values are assigned. QCI parameters include the packet error loss rate, class of the flow (GBR or non-GBR), budget for packet delay and priority value for admission control. For instance, a conversational voice service flow carries the following QoS parameters: QCI (1), radio resource type (GBR), priority (2), packet delay budget (100 ms), and packet error loss rate (102).

Figure 3 shows the second block of information, which comprises the packet latency or HoL delay, size of the queue and service rate from the buffer at the MAC layer. The scheduler receives this information with the QCI requirements and uses it to determine the flows' scheduling weights. Different schedulers, such as the packet loss fair strategy and delay-aware strategy, exist, and each scheduler is configured differently according to the various QoS requirements. LTE systems must fulfill these QoS requirements by achieving an optimal balance between utilization of the service and fairness for all the users. As an example, the delay-aware strategy utilizes the HoL delay (which is obtained from the buffer) and information about the target delay to establish how resources should be allocated to the various flows. Using a similar method, other strategies utilize information related to the objective of the scheduling rule, which is also extracted from the buffer. Scheduling rules aim to satisfy the flow requirements of the end user but can do so through different means, e.g., a target throughput, packet delivery delay bounds, the packet loss rate or any combination of these.

Further elaboration of the application of the two main categories, i.e., QoS-aware strategy and QoS-unaware strategy, is provided in the following subsections.

A. QoS-UNAWARE SCHEDULING STRATEGIES

The QoS-unaware strategies are designed to employ strategies that focus on parameters related to the fairness of the system, including the CSI and average data rate. Their objective is to allocate radio resources and schedule packets for wireless network users. This category incorporates the max-rate rule, round-robin rule and PF rule, as discussed in [4], [29] and illustrated in our proposed taxonomy in Figure 2. The max-rate rule is only aware of the instant bit rate of the users according to their instantaneous channel conditions (in contrast to the target bit rate in QoS-aware strategies, which take into account flow rates). The schedulers in this type of strategy have several objectives: 1) the throughput is maximized in the max-rate rule, but fairness among users is not ensured; 2) fairness is achieved by the round robin rule, but the throughput is not maximized; and 3) PF achieves maximum throughput and fairness.

One of the most important scheduling rules in this category is the Proportional Fair (PF) scheduler [4]. With this scheduler, PRBs are assigned by considering the experienced channel quality and estimated user average data rate, as calculated in Equation 1. PF schedulers are appropriate for NRT traffic because the scheduling matrix does not impose limitations on the flows, such as queue size or target delay. They aim to concurrently provide maximum network throughput and flow fairness. Equation 1 provides the metric utilized to represent the PF scheduler:

$$W_{i,j}(t) = \frac{r_{i,j}(t)}{\bar{R}_i(t)} \tag{1}$$

where $r_{i,j}(t)$ is the instantaneous data rate, calculated based on the Adaptive Modulation and Coding (AMC) module, which is chosen in relation to the Channel Quality Indicator (CQI) feedback sent by the User Equipment (UE). This feedback represents the channel quality (e.g., Signal-to-Interference Plus Noise-Ratio (SINR)) of the j-th sub-channel associated with the i-th flow. In addition, $\bar{R}_i(t)$ is represented in Equation 2.

$$\bar{R}_i(t) = (1-\beta)\bar{R}_i(t-1) + \beta r_i(t) \tag{2}$$

where $r_i(t)$ is the data rate achieved by the i-th flow in the current scheduling epoch t (i.e., total number of transmitted bits over the entire Physical Resource Block (PRB)s allocated to i-th flow per Transmission Time Interval (TTI)), $\bar{R}_i(t-1)$ is the estimated average data rate achieved by the i-th flow in the previous TTI, $\bar{R}_i(t)$ is the estimated average data rate achievable by the i-th user in the current TTI, and β is a moving average to smooth the system performance and control system fairness; the lower β is, the higher the system fairness.

B. QoS-AWARE SCHEDULING STRATEGIES

Following the previously mentioned discussion, QoS-aware strategies are divided into four classes (as shown earlier in Figure 2). These classes are determined by considering the different system and application parameters that contribute to the scheduling decision. These scheduling classes are listed as follows: 1) delay-aware strategies, 2) queue-aware strategies, 3) target bit-rate-aware strategies, and 4) proposed hybrid strategies. The applications and specifications of each of these classes are further discussed next.

1) DELAY-AWARE STRATEGIES

This class consists of the scheduling approaches most suitable for RT traffic types, for example, video gaming. The LTE QoS architecture in [30] specifies a packet delivery target delay and packet loss ratio thresholds for real-time and non-real-time traffic types. Since QoS constraints mainly depend on the application type, video traffic has stringent QoS requirements; therefore, one of the principal objectives of mobile networks is to ensure QoS provision by guaranteeing packet delivery within a target delay and with a constrained packet loss rate. According to citeQoSsurveyPiro, M-LWDF [7], Exponential Proportional Fair (EXP/PF) [7], and Exp-rule [8] are capable of meeting the video streaming requirements in terms of packet delivery delay bounds. However, these rules do not provide balanced QoS delivery when serving a mixture of traffic types, such as RT and NRT traffic. For simulation and performance analysis, the extensively deployed scheduling strategies from this class are discussed as follows:

- The purpose of the *M-LWDF* scheduler is to support multiple RT data users [7] (derived in [31]). PRBs are assigned by the scheduler to different RT flows, taking into consideration the properties of the classical PF rule and the HoL packet delay parameter. Moreover, the scheduler assigns PRBs to NRT data users with the PF rule. While the PF scheduler is considered to be more appropriate for NRT flows, both RT and NRT services can be supported by M-LWDF. Equation 3 illustrates the metric employed to represent the M-LWDF scheduler:

$$W_{i,j}(t) = \begin{cases} \alpha_i D_{HoL,i}(t)(\frac{r_{i,j}(t)}{\bar{R}_i(t)}), & \text{if } i \in RT \\ \frac{r_{i,j}(t)}{\bar{R}_i(t)}, & \text{if } i \in NRT \end{cases} \tag{3}$$

  where α_i is given by:

$$\alpha_i = \frac{-log\delta_i}{\tau_i} \tag{4}$$

  where the probability δ_i is defined as the maximum probability that the Head-of-Line (HoL) packet delay $D_{HoL,i}(t)$ (i.e., delay of the first packet that resides in the buffer to be transmitted) exceeds the target delay (τ_i). Hence, α_i is employed to ensure the delay constraints of users and is dependent on the choice of flow class from the 3GPP QoS Class Identifier (QCI) table in [30]. Therefore, if packets that belong to a RT service exceed the target delay while waiting at the MAC buffer, then they will be discarded. For definitions of $r_{i,j}(t)$ and $\bar{R}_i(t)$, refer to Equation 1.

- The purpose of the *EXP/PF* scheduler is to exponentially raise the priority of RT flows *w.r.t* NRT ones when their HoL packet delays are close to the target delay. For RT flows, parameters sensitive to delay and the PF rule are used to formulate the EXP/PF scheduler. This will prioritize the delay parameters over the flow channel conditions, which results in greater support for RT flows. However, the PF rule is used on its own to schedule packets among NRT users. The below equation illustrates the metric employed to represent the EXP/PF scheduler:
$$W_{i,j}(t)=\begin{cases} exp\left(\frac{\alpha_i D_{HoL,i}(t)-h(t)}{1+\sqrt{h(t)}}\right)\frac{r_{i,j}(t)}{\bar{R}_i(t)}, & \text{if } i\in RT \\ \frac{r_{i,j}(t)}{\bar{R}_i(t)}, & \text{if } i\in NRT \end{cases} \tag{5}$$
where $h(t)$ is given by:
$$h(t) = \frac{1}{N_i}\sum_{i=1}^{N_i}\alpha_i D_{HoL,i}(t), \text{ for } i \in RT \tag{6}$$
where the parameters are defined the same as in Equation 3. $h(t)$ refers to the average head-of-line delay (system head-of-line delay), and N_i is the number of active downlink RT flows. This approach aims to limit the delays of all the RT flows.
- The *EXP-rule* scheduler is adopted from the EXP/PF rule, which has been optimized to produce greater throughput for RT users. The PF rule is used here for scheduling NRT flows, whereas the EXP-rule along with the PF rule are used for scheduling RT flows. The below equation illustrates the metric employed to represent the EXP-rule scheduler:
$$W_{i,j}(t) = \begin{cases} exp\left(\frac{\alpha_i D_{HoL,i}(t)}{1+\sqrt{h(t)}}\right)\frac{r_{i,j}(t)}{\bar{R}_i(t)}, & \text{if } i \in RT \\ \frac{r_{i,j}(t)}{\bar{R}_i(t)}, & \text{if } i \in NRT \end{cases} \tag{7}$$
where $h(t)$ is given by:
$$h(t) = \frac{1}{N_i}\sum_{i=1}^{N_i} D_{HoL,i}(t), \text{ for } i \in RT \tag{8}$$
where the parameters are defined the same as in Equations 3 and 5. However, the maximum (δ_i) in α_i is set to either 6 or 10, as in [8] (as this delivers good performance results and $h(t)$ does not consider the selected value of α_i).

2) QUEUE-AWARE STRATEGIES

These strategies employ parameters that determine fairness, in particular queue size. They allocate radio resources and schedule packets for different wireless network users, as proposed in [6], [32]. Therefore, these strategies are employed to optimize the throughput for RT traffic types and provide minimum rate guarantees for NRT traffic types. For instance, one of the extensively applied schedulers in this class is referred to as Virtual Token Modified Largest Weighted Delay First (VT-M-LWDF) (i.e., the virtual token scheduling rule in [6]). This rule is a modified version of the M-LWDF rule, and it employs the parameters highlighted in Equation 3, apart from the HoL packet delay parameter. The VT-M-LWDF scheduler replaces the HoL packet delay parameter with the queue size parameter for RT users. This parameter indicates to the scheduler the magnitude of the flow available in the buffer. Conversely, the M-LWDF scheduling decision is based largely on HoL packet delays. The principal objective of the VT-M-LWDF rule is to improve QoS performance metrics for multimedia services, such as VoIP, and offer minimum throughput guarantees for NRT services. This approach affects the QoS requirements for NRT services. Therefore, a balance is needed in the delivery of QoS requirements for RT and NRT services, which inevitably involves a trade-off. Equation 9 illustrates the metric employed to represent the VT-M-LWDF scheduler:
$$W_{i,j}(t) = \begin{cases} \alpha_i Q_i(t)(\frac{r_{i,j}(t)}{\bar{R}_i(t)}), & \text{if } i \in RT \\ \frac{r_{i,j}(t)}{\bar{R}_i(t)}, & \text{if } i \in NRT \end{cases} \tag{9}$$
where the $\alpha_i(t)$, $r_{i,j}(t)$ and $\bar{R}_i(t)$ parameters are defined the same as in Equation 3. $Q_i(t)$ refers to the queue size of the i-th flow at a particular scheduling epoch (t) when serving RT users. On the other hand, the PF scheduler is employed to make scheduling decisions for NRT users.

3) TARGET BIT-RATE-AWARE STRATEGIES

This class comprises strategies that are aware of the bit rate of the flows in the scheduling buffers. The scheduling takes into account non-GBR and GBR classes for the flows, as proposed in [33]–[35]. These schedulers aim to maximize the throughput of the total system and deliver the minimum/maximum target bit rate for a mixture of RT and NRT flows. However, these rules do not provide balanced QoS delivery when serving a mixture of traffic types, such as RT and NRT traffic.

4) THE PROPOSED HYBRID SCHEDULING APPROACH

This class is our proposed class, which is suitable for simultaneously providing balanced QoS for RT and NRT traffic types. This scheduling class is different from the previous scheduling classes in the way that it handles and prioritizes the packets. To achieve the QoS balancing goal for RT and NRT traffic types, the designed priority metric considers a mixture of four important flow parameters: channel conditions, HoL packet delay, flows queue sizes, and flow type. Moreover, the goal of this proposed scheduling class is to maintain an acceptable Mean Opinion Score (MOS) quality level for video applications and increase the system capacity and fairness. More details and a performance analysis of one of our proposed hybrid-based scheduling algorithms is provided in [3]. On the other hand, it is important to mention that

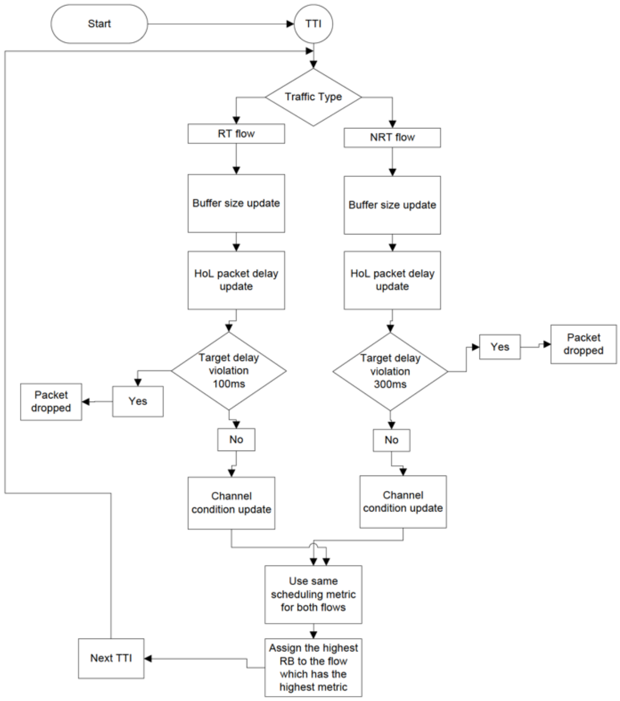


FIGURE 4. Flow chart for the proposed hybrid scheduling approach.

a few studies (such as [36]–[39]) proposed hybrid scheduling approaches for traditional multi-carrier systems, such as WiFi and WiMAX systems, considering different priority parameters that are not suitable for RT and NRT flows. Figure 4 shows the flow chart for the implementation of the proposed hybrid scheduling approach. The hybrid scheduling metrics are further elaborated as follows:

- The *Queue-HoL-M-LWDF* scheduler in [3] is proposed to enhance the performance of existing M-LWDF and VT-M-LWDF scheduling algorithms. This scheduler is a combination of the effective components and main principles of the schedulers mentioned above. We adopt consideration of the queue size of the buffer, traffic types, and HoL packet delay parameters in the VT-M-LWDF and M-LWDF rules. This decision was made in order to enhance the scheduler's performance when serving both RT and NRT services. In particular, the scheduler aims to deliver better performance metrics for video services while sustaining an acceptable and balanced QoS delivery for the other network services. Equation 10 illustrates the metric employed to represent the Queue-HoL-M-LWDF scheduler:

$$W_{i,j}(t)=\alpha_i D_{HoL,i}(t) Q_i(t)\Big(\frac{r_{i,j}(t)}{\bar{R}_i(t)}\Big), \text{ for} i \in RT/NRT \tag{10}$$

  where the parameters are defined the same as in Equations 3 and 9.
- The proposed *Modified-EXP-rule* scheduler is introduced to improve the existing EXP-rule scheduling rule presented in Equation 7. The proposed rule considers the awareness of the channel, traffic type, and HoL delay parameters for both RT and NRT users simultaneously. This fusion of metrics in the EXP-rule improves the performance of the system, efficiently utilizes the system's radio resources, and balances the QoS delivery for both RT and NRT traffic. It is important to highlight that the NRT flows are delay-tolerant services, i.e., delayed packets are not discarded and their delay tolerance threshold is higher than RT traffics as stated in the 3GPP QCI table [30]. However, the RT flows are delay sensitive services, where packets belonging to this kind of services are discarded, if the HoL packet delay exceeds the target delay. As a result, this will help maintain the balance and satisfy both RT and NRT admitted users by providing: a higher system spectral efficiency, higher system throughput, and lower system PLR, guaranteeing a satisfactory level of fairness, and supporting a higher number of subscribers. The below equations illustrate the metric employed to represent the proposed Modified-EXP-rule scheduler:

$$W_{i,j}(t) = exp\left(\frac{\alpha_i D_{HoL,i}(t)}{1+\sqrt{h(t)}}\right)\frac{r_{i,j}(t)}{\bar{R}_i(t)}, \text{ for } i \in RT/NRT \tag{11}$$

where $h(t)$ is given by:

$$h(t) = \frac{1}{N_i}\sum_{i=1}^{N_i} D_{HoL,i}(t), \text{ for } i \in RT/NRT \tag{12}$$

where the parameters are defined the same as in Equation 9, whereas the i in this rule corresponds to both RT and NRT users.

To highlight more on the complexity level of the proposed scheduling strategy, QoS aware schedulers with packet priority results in a less complex system, as such scheduling rules are simple to implement. Low complexity scheduling rules are very important in LTE mainly because of the short scheduling interval of 1 ms. Complexity analysis of the proposed scheduler is highlighted. We analyze the complexity of the considered strategy in terms of the maximum number of required iterations. In this strategy, the scheduler computes $I \cdot M_{\text{PRB}}$ metrics per scheduling epoch, where I is the number of flows and M_{PRB} is the number of PRBs. The scheduling function for the proposed strategy requires the computation of HoL delay which comprises the recording of packets' arrival time at the eNodeB. However, this does not affect the number of iterations required to compute the scheduling rule. Therefore, the per-PRB scheduling rule has a linear dependency on the number of PRBs and flows for this strategy. In other words, the proposed algorithm allocates a PRB to the user after performing a linear search among all the active users thus the computation complexity is the product of the number of users

TABLE 1. QoS and channel parameters used by QoS-aware and QoS-unaware downlink scheduling approaches.

Scheduling Approaches	**Channel Parameters**		**QoS Parameters**					
	CSI	*Average data rate*	*PLR*	*Target rate*	*Queue size*	*Service aware*	*Target delay*	*HoL delay*
QoS-aware approaches								
Delay-aware approaches								
M-LWDF [5]	X	X				RT	X	X
EXP/PF [7]	X	X				RT, NRT	X	X
DAPS various traffic classes [40]	X					RT, NRT	X	X
EXP-rule [8]	X	X					X	X
Log-rule [8]	X	X					X	X
Queue-aware approaches								
PFPS [32]	X	X			X			
VT-M-LWDF [6]	X	X			X	RT, NRT		
Target bit-rate-aware approaches								
Target bit-rate rule [33]	X	X		X				
Target bit-rate rule [34]	X			X				
Target bit-rate rule [35]	X			X				
Hybrid approaches								
Queue-HoL-M-LWDF [3]	X	X			X	RT, NRT	X	X
Modified-PF [36]	X			X		RT, NRT	X	
Joint channel and queue [41]	X	X			X	RT, NRT	X	X
Hybrid approach [37]	X	X	X	X		RT, NRT	X	
DPS [38]			X	X		RT, NRT	X	
FLS [14]	X	X			X	RT, NRT	X	
QoS-unaware approaches								
PF [4]	X	X						
Maximum-rate [29]	X							
Round robin [29]								

and the number of PRBs in the system (i.e., the computation complexity at the scheduler will be computed as $O(I \cdot M_{\text{PRB}})$. Hence linear complexity enables real-time implementation of the algorithm.

To summarize the scheduling classes, Table 1 reports the important scheduling approaches and their common parameters, which are incorporated into both QoS-aware and QoS-unaware downlink scheduling strategies.

IV. SIMULATION SET-UP AND COMPARATIVE PERFORMANCE ANALYSIS

A. SYSTEM MODEL

A range of factors can affect the QoS in LTE and beyond wireless systems. These factors include the target delays, number of available radio resources, channel conditions and type of services (e.g., sensitive or insensitive to delay). The radio resources allocated to a user are known as a PRB, equal to 180 kHz in the frequency domain. The duration of each resource is 0.5 ms, and each resource contains 6 or 7 Orthogonal Frequency Division Multiplexing (OFDM) symbols in the time domain. A range of channel bandwidths, i.e., 1.4, 3, 5, 10, 15 and 20 MHz, are incorporated into the LTE standard, and each bandwidth comprises a different number of PRBs. A detailed study with recommendation on the selection of the proper bandwidth in terms of system efficiency utilization is carried out in [42], where efficient employment of these bandwidth ranges leads to improved LTE system efficiency. The results reported in this study provide guidelines for combining bandwidth scalability and admission control strategies in LTE networks in order to achieve high system resource utilization and deliver high traffic quality for the LTE users. Therefore, a 10 MHz bandwidth is carefully chosen in our study in order to ensure a proper utilization of the system spectral efficiency. This can be observed in Figure 9 (b) below, where our proposed scheduling algorithms efficiently utilize the available radio resources. The work in this study involves various types of downlink scheduling rules. Therefore, the most relevant downlink parameters were set according to the LTE standard (i.e., resource allocation: every TTI equal to 1 ms; two time slots, where each time slot has 7 OFDM symbols spread over 12 consecutive sub-carriers; Frequency Division Duplexing (FDD) frame composed of 10 TTIs).

In the system, CQI feedback should be reported to the eNodeB by the users in each TTI via uplink control messages over the Physical Uplink Control Channel (PUCCH). This CQI value is representative of the users' instantaneous channel quality/available transmission data rate for each PRB. In the scenario in this work, the feedback mechanism is that the UE sends a single CQI that relates to every PRB to the

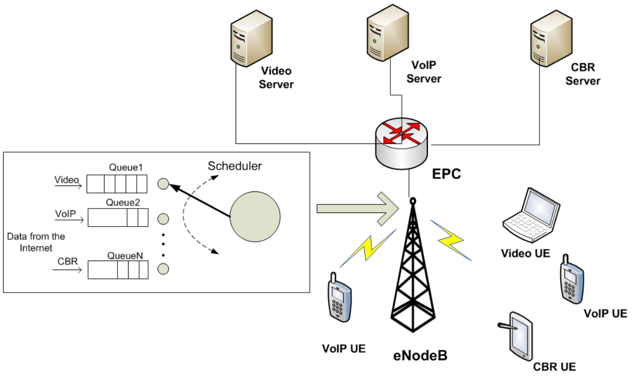


FIGURE 5. System model [12].

eNodeB in the corresponding channel bandwidth. AMC is deployed by mapping the SINR and CQI values in accordance with the mapping tables presented in [43]. Therefore, the selected modulation and coding scheme ensures that the user enjoys high-quality service delivery and communication and that the estimated Block Error Rate (BLER) is less than the target BLER of 10% [43], [44].

The packet scheduler at the MAC layer in the serving eNodeB controls the selection of the available PRBs by allocating them to the active flows with the highest priority value, which vie with each other for resources. At the MAC layer, scheduling of packets and allocation of radio resources are carried out. The scheduling decision in this work is based on parameters such as the size of the buffer, channel conditions and HoL packet delays. Several scheduling algorithms are considered (refer to Section III). Figure 5 illustrates the system model employed in this research. We consider a multi-user network scenario, with VoIP, video and CBR servers as its backbone, and the schedulers implemented at the eNodeB's MAC layer control the allocation of resources between flows.

B. PERFORMANCE ANALYSIS

This subsection discusses the simulation environment, video traffic models, relevant performance metrics, and numerical results.

1) SIMULATION ENVIRONMENT

The simulation involves a single LTE cell with equally distributed users. At the center of the cell is the eNodeB, and a random mobility model is employed to model the users. The users are assigned a mobility speed of 3 km/h to simulate pedestrians, and the LTE-Sim [44] simulator is utilized. This simulator supports several aspects in the approved 3GPP LTE standard, in order to provide researchers the ability to obtain real implementation of an LTE network. The simulator involves the following implementations: 1) multi-cell scenarios; 2) downlink and uplink scheduling strategies; 3) user mobility models; 4) LTE frame structure modes; 5) frequency reuse techniques; 6) Adaptive Modulation and Coding (AMC) module; and 7) LTE protocol stack and various channel models. In this study, the simulator is utilized in order to enable designing and testing different packet scheduling strategies at the eNodeB MAC layer. The traffic classes of the users are assumed to be distributed as follows: 40% are video users, 40% are VoIP users and 20% are Constant Bit Rate (CBR) users.

To achieve reliable and precise results, the simulation is repeated several times, and the performance metrics are averaged. Table 2 reports the simulation parameters.

TABLE 2. Simulation parameters for LTE downlink system.

PARAMETERS	VALUE
Bandwidth	10 MHz
Number of PRBs	50
β	0.01
Frame structure	FDD
Cell radius	1 km
E-UTRAN frequency band	1 (2.1 GHz)
Target delay (τ)	100 ms
Video bit rate	242 kbps
Constant Bit Rate (CBR) bit rate	100 kbps
VoIP bit rate	8.4 kbps
Flow duration	20 sec
Simulation time	30 sec
UE speed	3 km/h
Path loss/channel model	Typical urban (pedestrian A propagation model)
Simulation repetitions per scheduler	10

2) TRAFFIC MODEL

The video flow is a trace-based application that employs realistic video trace files to send packets. These trace files were extracted from [45]. The H.264 encoder is utilized to encode the chosen video flow at a rate of 242 kbps and a frame rate of 25 fps. The maximum transmission unit is set to 500 bytes, as [46] reported this to be efficient.

The voice flow is a bursty application that is modeled with an ON/OFF Markov chain. The G.729 code is employed to encode the generated VoIP flow at a rate of 8.4 kbps.

The BR flow models a constant bit-rate application, such as File Transfer Protocol (FTP), Peer-to-Peer (P2P) and Hyper Text Transfer Protocol (HTTP), with fixed packet size and inter-arrival packet time. The selected CBR application is a non-real-time FTP flow with a constant bit rate of 100 kbps and a packet size and an inter-arrival packet time of 500 bytes and 0.04 s, respectively. Note that NRT traffic is usually implemented under TCP, which ensures that the PLR rate is lower than the rates given in the figures in the following section due to the Transmission Control Protocol (TCP) retransmission control. However, the LTE-Sim simulator [44] only supports User Datagram Protocol (UDP) protocols.

3) PERFORMANCE METRICS

The performance metrics for each scheduling strategy (average packet loss ratio, average packet delay, average throughput, fairness index and system spectral efficiency) were measured for repetitions of the simulation and then averaged. The first three metrics are user-oriented, while the last two metrics are system-oriented.

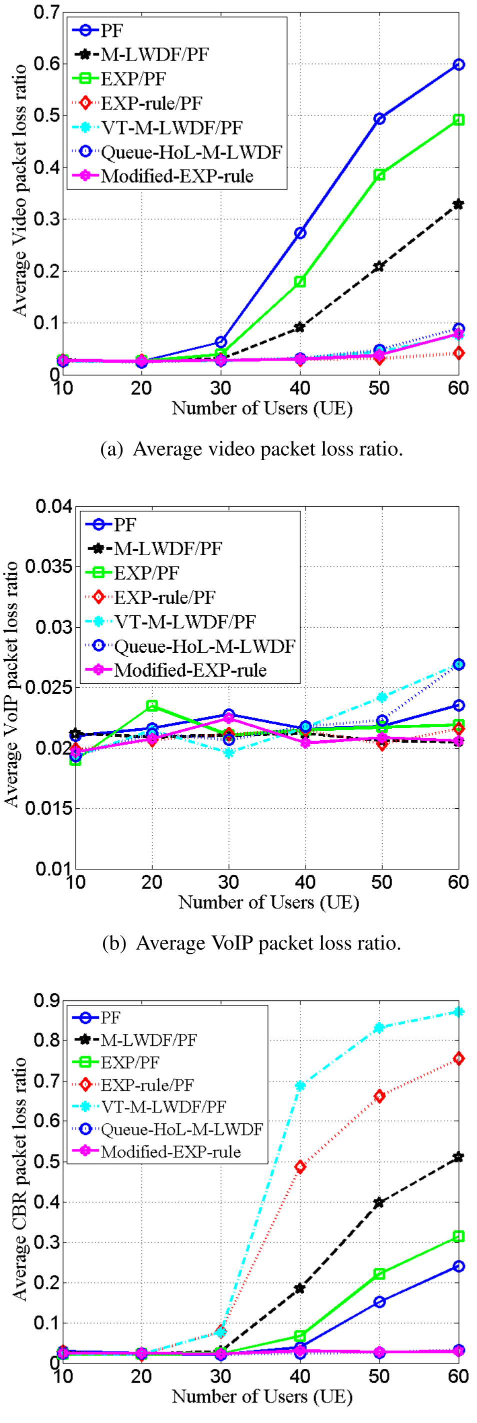
- The *packet loss ratio* is calculated by dividing the difference between the transmitted and received packets by the number of transmitted packets.
- The *average packet delay* parameter is calculated by dividing the sum of the received packet delays by the number of packets received.
- The *average throughput* parameter is calculated by dividing the number of successfully received bits by the duration of the flow.
- The *fairness index* parameter follows Jain's fairness index criterion [47]. It illustrates how fairness is being maintained in the assignment of a fair share of system resources to users. Equation 13 represents the fairness index metric:

$$F(x_1, x_2, \ldots, x_n) = \frac{(\sum_{i=1}^{n} x_i)^2}{n \cdot \sum_{i=1}^{n} x_i^2} \tag{13}$$

  where n is the number of served users, and x_i is the throughput for the *i*-th connection. The maximum value is 1, which is achieved when the same amount of resources are shared or received by all users, indicating fairness of the system.
- The *system spectral efficiency* parameter is calculated by dividing the total number of received bits per second by the size of the bandwidth.

4) NUMERICAL RESULTS

The performance evaluation conducted in this work is presented with respect to the scheduling rules mentioned in Section III. The analyzed results are presented in terms of the performance metrics defined in Subsection IV-B.3 for video, VoIP and CBR users with increasing number of users. The schedulers applied in our test are denoted PF, M-LWDF/PF, EXP/PF, EXP-rule/PF, VT-M-LWDF/PF, Queue-HoL-MLWDF, and Modified-EXP-rule. The following simulation results show the balance of QoS delivery among the different flows produced by our proposed hybrid scheduling rules: Queue-HoL-MLWDF and Modified-EXP-rule.



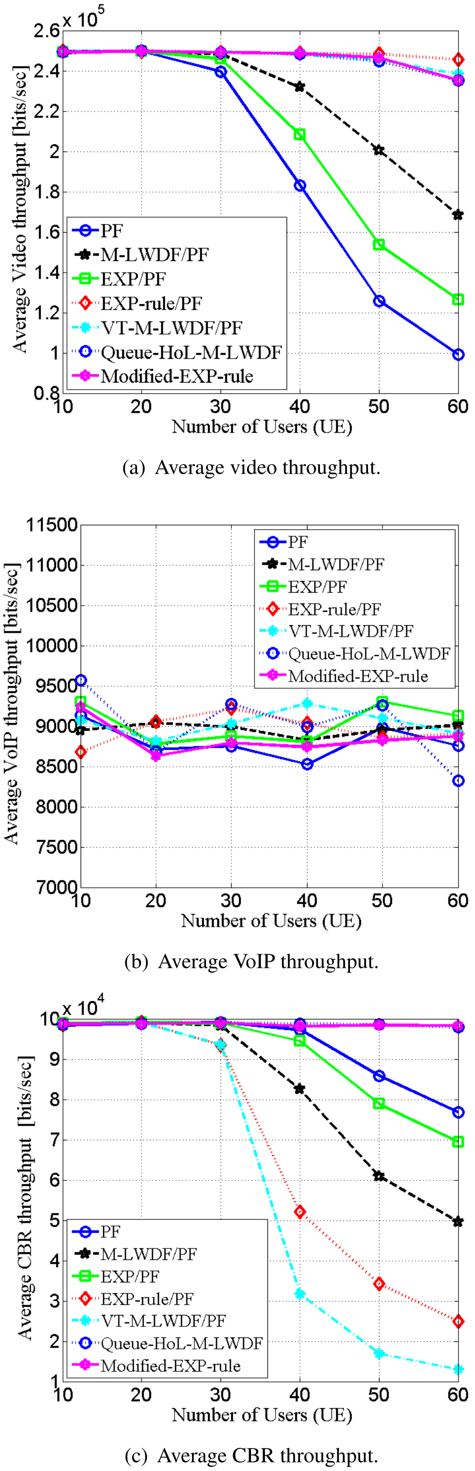
(a) Average video packet loss ratio.

(b) Average VoIP packet loss ratio.

(c) Average CBR packet loss ratio.

FIGURE 6. Average packet loss ratios for video, VoIP, and CBR flows.

The average packet loss ratios (PLRs) for video, VoIP, and CBR flows are presented in Figure 6. As the number of users rises, the quality decreases. VT-M-LWDF, EXP-rule/PF, the proposed Queue-HoL-MLWDF, and the proposed Modified-EXP-rule maintain low PLRs for video and VoIP flows with rising user number. However, the PLR for CBR flows is high with PF, M-LWDF/PF, EXP/PF, EXP-rule/PF, VT-M-LWDF/PF, whereas the PLR is kept low with the proposed Queue-HoL-M-LWDF and proposed Modified-EXP-rule. Moreover, note that the PF rule is suitable for NRT flows, as mentioned in Subsection III-A, where it shows relatively low PLR for CBR application compared with other existing scheduling rules. To highlight the significance of the proposed hybrid scheduling rules, it is observed that some of the delay-aware and queue-aware scheduling rules (i.e., the VT-M-LWDF rule and EXP-rule/PF rule) show very low PLR for RT flows only, whereas the proposed hybrid scheduling rules also maintain very low PLR for RT and NRT flows. Hence, this observation indicates that our proposed hybrid scheduling approaches (i.e., Queue-HoL-M-LWDF and Modified-EXP-rule) balance the QoS for all traffic classes.

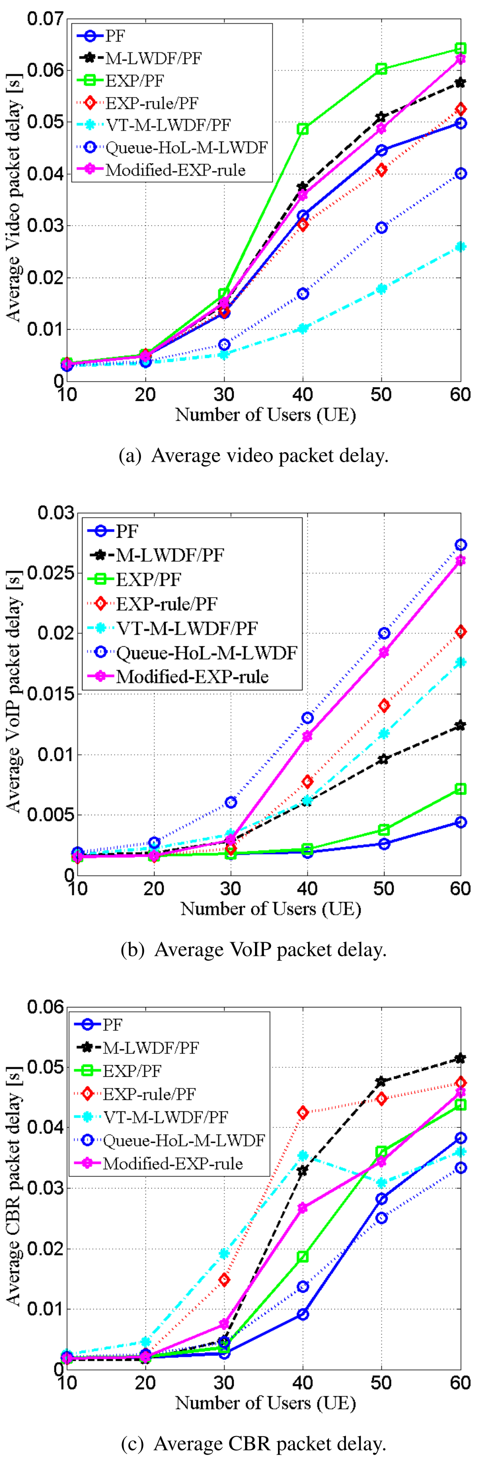


(a) Average video throughput.

(b) Average VoIP throughput.

(c) Average CBR throughput.

FIGURE 7. Average throughputs for video, VoIP, and CBR flows.



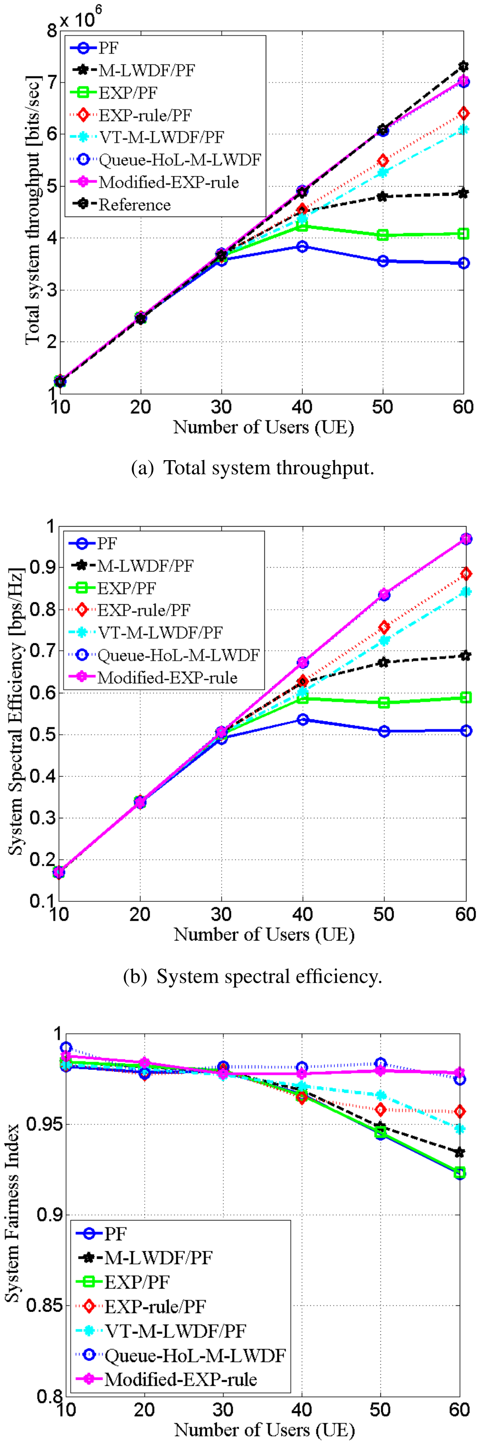
(a) Average video packet delay.

(b) Average VoIP packet delay.

(c) Average CBR packet delay.

FIGURE 8. Average packet delays for video, VoIP, and CBR flows.

The average throughputs for video, CBR and VoIP flows are reported in Figure 7. Similar to the previously observed performance metric, the average throughputs achieved by our proposed hybrid scheduling rules show significant improvement and balanced QoS for both RT and NRT flows. This finding is also evident when the network is loaded with 60 users. Furthermore, note that the throughput magnitude is affected by the value of the PLR, which is affected by an increase in the user number. In addition, the average packet delay is presented in Figure 8. As previously stated in Table 2, the target delay of the services is set to 100 ms; hence, the results show that the QoS-based scheduling approaches satisfy the QoS requirements.



(a) Total system throughput.

(b) System spectral efficiency.

(c) System fairness.

FIGURE 9. Total system throughput, system spectral efficiency, and system fairness index.

The total system throughput, system spectral efficiency, and system fairness for a network occupied with video, VoIP and CBR flows are reported in Figure 9. The total system throughput shows the cumulative throughput for the three traffic types with increasing number of users. The reference black curve in Figure 9(a) shows the actual throughput received by users in a network without wireless errors (e.g.,approximately 7.3 Mbps for a system load of 60 users). We observe that our proposed hybrid scheduling approaches, which are designed for QoS balancing, are near the reference curve. This finding signifies that the radio resource management and channel utilization achieved by our proposed schedulers are efficiently controlled, as shown in Figure 9(b). Figure 9(b) reports the system spectral efficiency, which shows the utilization of the resources. According to the simulation parameters, the average system capacity is approximately 23 Mbps (2.33 bits/sec/Hz considering a 10 MHz bandwidth). The figure also shows that the cell is not saturated for the proposed hybrid scheduling rules because the schedulers can still serve more users in the network with a reasonable packet loss ratio until the saturation point. The saturation point is the point at which the cell is loaded with a high number of users and managing the allocation of resources becomes difficult. The system fairness is presented in Figure 9(c). This metric shows how fairly users are being served by the system via a fair assignment of system resources. The figure shows that our proposed hybrid schedulers achieve better fairness for different types of services.

V. CONCLUSION

This paper proposes hybrid QoS-aware downlink scheduling approaches that aim to address different traffic classes and balance the QoS parameters with an improvement in the overall system performance. Moreover, the paper proposes a taxonomy that classifies the scheduling algorithms into four main classes: delay aware, queue aware, target bit-rate aware, and hybrid aware. The latter class is the proposed scheduling class, which considers channel, queue and delay parameters in its scheduling metric. Using simulations, we compare and analyze different downlink scheduling rules in terms of network-centric performance metrics, such as the average packet loss ratio, average throughput, average packet delay, system fairness, and system spectral efficiency. According to the simulation results, queue-aware and delay-aware scheduling rules perform best in terms of QoS performance for video traffic classes, whereas our proposed hybrid scheduling rules (i.e., Queue-HoL-MLWDF and Modified-EXP-rule) deliver balanced QoS among the different types of traffic classes and maximize the system performance in terms of system throughput, packet loss, spectral efficiency, delay, and fairness. Therefore, QoS balancing scheduling rules appear to be the most attractive strategies for an LTE downlink, and they should therefore be employed to offer high QoS delivery for different traffic classes. In future, an extension to this study related to developing novel scheduling strategies over 5G systems will be considered.

ACKNOWLEDGMENT

Dr. Moustafa M. Nasralla would like to acknowledge the management of Prince Sultan University (PSU) and the Renewable Energy Lab for their valuable support and provision of research facilities that were essential for the completion of this work.

REFERENCES

[1] M. M. Nasralla, N. Khan, and M. G. Martini, "Content-aware downlink scheduling for LTE wireless systems: A survey and performance comparison of key approaches," *Comput. Commun.*, vol. 130, pp. 78–100, Oct. 2018.

[2] ITU-T Recommendation, "Definitions of terms related to quality of service," Int. Telecommun. Union, Geneva, Switzerland, Tech. Rep., 2008.

[3] M. M. Nasralla and M. G. Martini, "A downlink scheduling approach for balancing QoS in LTE wireless networks," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, London, U.K., Sep. 2013, pp. 1571–1575.

[4] J.-G. Choi and S. Bahk, "Cell-throughput analysis of the proportional fair scheduler in the single-cell environment," *IEEE Trans. Veh. Technol.*, vol. 56, no. 2, pp. 766–778, Mar. 2007.

[5] P. Ameigeiras, J. Wigard, and P. Mogensen, "Performance of the M-LWDF scheduling algorithm for streaming services in HSDPA," in *Proc. IEEE 60th Veh. Technol. Conf. (VTC-Fall)*, Los Angeles, CA, USA, Sep. 2004, pp. 999–1003.

[6] M. Iturralde, T. Ali Yahiya, A. Wei, and A.-L. Beylot, "Performance study of multimedia services using virtual token mechanism for resource allocation in LTE networks," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, San Francisco, CA, USA, Sep. 2011, pp. 1–5.

[7] R. Basukala, H. A. M. Ramli, and K. Sandrasegaran, "Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system," in *Proc. 1st Asian Himalayas Int. Conf. Internet*, Kathmandu, Nepal, Nov. 2009, pp. 1–5.

[8] B. Sadiq, R. Madan, and A. Sampath, "Downlink scheduling for multiclass traffic in LTE," *EURASIP J. Wireless Commun. Netw.*, vol. 2009, no. 1, pp. 1–18, Dec. 2009.

[9] S. Shakkottai and A. L. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule," *Translations Amer. Math. Soc.-Series 2*, vol. 207, pp. 185–202, Dec. 2002.

[10] M. M. Nasralla, M. Razaak, I. U. Rehman, and M. G. Martini, "Content-aware packet scheduling strategy for medical ultrasound videos over LTE wireless networks," *Comput. Netw.*, vol. 140, pp. 126–137, Jul. 2018.

[11] I. Rehman, M. Nasralla, and N. Philip, "Multilayer perceptron neural network-based QoS-aware, content-aware and device-aware QoE prediction model: A proposed prediction model for medical ultrasound streaming over small cell networks," *Electronics*, vol. 8, no. 2, p. 194, 2019.

[12] M. M. Nasralla and I. U. Rehman, "QCI and QoS aware downlink packet scheduling algorithms for multi-traffic classes over 4G and beyond wireless networks," in *Proc. Int. Conf. Innov. Intell. for Informat., Comput., Technol. (3ICT)*, Nov. 2018, pp. 1–7.

[13] J.-H. Rhee, J. M. Holtzman, and D.-K. Kim, "Scheduling of real/non-real time services: Adaptive EXP/PF algorithm," in *Proc. 57th IEEE Semiannu. Veh. Technol. Conf. (VTC -Spring)*, Apr. 2003, pp. 462–466.

[14] G. Piro, L. A. Grieco, G. Boggia, R. Fortuna, and P. Camarda, "Two-level downlink scheduling for real-time multimedia services in LTE networks," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 1052–1065, Oct. 2011.

[15] H. A. M. Ramli, R. Basukala, K. Sandrasegaran, and R. Patachaianand, "Performance of well known packet scheduling algorithms in the downlink 3GPP LTE system," in *Proc. IEEE 9th Malaysia Int. Conf. Commun. (MICC)*, Kuala Lumpur, Malaysia, Dec. 2009, pp. 815–820.

[16] A. H. Ali and M. Nazir, "Radio resource management with QoS guarantees for LTE—A systems: A review focused on employing the multi-objective optimization techniques," *Telecommun. Syst.*, vol. 67, no. 2, pp. 349–365, Feb. 2018.

[17] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 678–700, 2nd Quart., 2013.

[18] A. Asadi and V. Mancuso, "A survey on opportunistic scheduling in wireless communications," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 4, pp. 1671–1688, Jan. 2013.

[19] P. Ma, Y. Lu, Y. Hou, L. Li, X. Zhao, L. Zhang, and H. Zhu, "A multiservice QoS guaranteed scheduling algorithm for TD-LTE 230 MHz power wireless private networks," in *Proc. 12th Int. Symp. Antennas, Propag. EM Theory (ISAPE)*, Dec. 2018, pp. 1–4.

[20] F. Asadollahi and R. Dehdasht-Heydari, "Introduction of a novel hybrid weighted exponential logarithm-maximum throughput (HWEL-MT) scheduler for QoS improvement of LTE/4G cellular networks," *Wireless Pers. Commun.*, vol. 98, no. 1, pp. 91–104, Jan. 2018.

[21] F. H. C. Neto, E. B. Rodrigues, D. A. Sousa, T. F. Maciel, and F. R. P. Cavalcanti, "QoS-aware scheduling algorithms to enhance user satisfaction in OFDMA systems," *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 10, p. e3165, Oct. 2017.

[22] S. Chaudhuri, I. Baig, and D. Das, "A novel QoS aware medium access control scheduler for LTE-advanced network," *Comput. Netw.*, vol. 135, pp. 1–14, Apr. 2018.

[23] M. Mamman, Z. M. Hanapi, A. Abdullah, and A. Muhammed, "Quality of service class identifier (QCI) radio resource allocation algorithm for LTE downlink," *PLoS ONE*, vol. 14, no. 1, 2019, Art. no. e0210310.

[24] O. G. Uyan and V. C. Gungor, "QoS-aware LTE–A downlink scheduling algorithm: A case study on edge users," *Int. J. Commun. Syst.*, vol. 32, no. 15, p. e4066, Oct. 2019.

[25] N. K. M. Madi, Z. M. Hanapi, M. Othman, and S. K. Subramaniam, "Delay-based and QoS-aware packet scheduling for RT and NRT multimedia services in LTE downlink systems," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, pp. 1–21, Dec. 2018.

[26] F. G. C. Rocha and F. H. T. Vieira, "A channel and queue-aware scheduling for the LTE downlink based on service curve and buffer overflow probability," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 729–732, Jun. 2019.

[27] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Transmission and Reception (Release 8)*, document TS 36.101, 3GPP, 2009.

[28] *3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and Charging Control Architecture (Release 8)*, document TS 23.203, 3GPP, 2010.

[29] H. A. M. Ramli, K. Sandrasegaran, R. Basukala, R. Patachaianand, and T. S. Afrin, "Video streaming performance under well-known packet scheduling algorithms," *Int. J. Wireless Mobile Netw.*, vol. 3, no. 1, pp. 25–38, Feb. 2011.

[30] *3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Policy and Charging Control Architecture (Release 10)*, document TS 23.203, 3GPP, Mar. 2012.

[31] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.

[32] W. Shih-Jung, "A channel quality-aware scheduling and resource allocation strategy for downlink LTE systems," *J. Comput. Inf. Syst.*, vol. 8, no. 2, pp. 695–707, Jan. 2012.

[33] G. Monghal, K. I. Pedersen, I. Z. Kovacs, and P. E. Mogensen, "QoS oriented time and frequency domain packet schedulers for the UTRAN long term evolution," in *Proc. VTC Spring-IEEE Veh. Technol. Conf.*, Singapore, May 2008, pp. 2532–2536.

[34] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, "Multi-QoS-aware fair scheduling for LTE," in *Proc. IEEE 73rd Veh. Technol. Conf. (VTC Spring)*, Yokohama, Japan, May 2011, pp. 1–5.

[35] D. N. Skoutas and A. N. Rouskas, "Scheduling with QoS provisioning in mobile broadband wireless systems," in *Proc. Eur. Wireless Conf. (EW)*, Lucca, Italy, Apr. 2010, pp. 422–428.

[36] P. Svedman, S. K. Wilson, and B. Ottersten, "A QoS-aware proportional fair scheduler for opportunistic OFDM," in *Proc. IEEE 60th Veh. Technol. Conf. (VTC-Fall)*, Los Angeles, CA, USA, Sep. 2004, pp. 558–562.

[37] L. Yanhui, W. Chunming, Y. Changchuan, and Y. Guangxin, "Downlink scheduling and radio resource allocation in adaptive OFDMA wireless communication systems for user-individual QoS," in *Proc. World Acad. Sci., Eng. Technol.*, Mar. 2006, pp. 221–225.

[38] M. Sarkar and H. Sachdeva, "A QoS aware packet scheduling scheme for WiMAX," in *Proc. World Congr. Eng. Comput. Sci.*, San Francisco, CA, USA, Oct. 2010, pp. 857–864.

[39] C. He and R. D. Gitlin, "Application-specific and QoS-aware scheduling for wireless systems," in *Proc. IEEE 25th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Washington, DC, USA, Sep. 2014, pp. 1147–1151.

[40] S. J. Bae, B.-G. Choi, and M. Y. Chung, "Delay-aware packet scheduling algorithm for multiple traffic classes in 3GPP LTE system," in *Proc. 17th Asia Pacific Conf. Commun.*, Sabah, Malaysia, Oct. 2011, pp. 33–37.

[41] K. Sun, Y. Wang, T. Wang, Z. Chen, and G. Hu, "Joint channel-aware and queue-aware scheduling algorithm for multi-user MIMO-OFDMA systems with downlink beamforming," in *Proc. IEEE 68th Veh. Technol. Conf.*, Calgary, AB, Canada, Sep. 2008, pp. 1–5.

[42] M. M. Nasralla, O. Ognenoski, and M. G. Martini, "Bandwidth scalability and efficient 2D and 3D video transmission over LTE networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC)*, Budapest, Hungary, Jun. 2013, pp. 617–621.

[43] *Physical Layer Procedures (Release 8)*, document TS 36.213, 3GPP, 2010.

[44] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE cellular systems: An open-source framework," *IEEE Trans. Veh. Technol.*, vol. 60, no. 2, pp. 498–513, Feb. 2011.

[45] *H.264/AVC and SVC Video Trace Library*. Accessed: Jan. 1, 2020. [Online]. Available: http://trace.eas.asu.edu/

[46] C. T. E. R. Hewage, M. G. Martini, and N. Khan, "3D medical video transmission over 4G networks," in *Proc. 4th Int. Symp. Appl. Sci. Biomed. Commun. Technol. (ISABEL)*, Barcelona, Spain, Oct. 2011, pp. 26–29.

[47] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Digit. Equip. Cooper., Hudson, MA, USA, DEC Res. Rep. TR-301, Sep. 1984.

MOUSTAFA M. NASRALLA (Member, IEEE) received the B.Sc. degree in electrical engineering from Hashemite University, Jordan, in 2010, the M.Sc. degree (Hons.) in networking and data communications from Kingston University London, U.K., in 2011, and the Ph.D. degree from the Faculty of Science, Engineering and Computing (SEC), Kingston University London. His Ph.D. research was based on video quality and QoS-driven downlink scheduling for 2-D and 3-D video over LTE networks. He was a member of the Wireless Multimedia and Networking (WMN) Research Group. He is currently an Assistant Professor with the Department of Communications and Networks Engineering, Prince Sultan University, Riyadh, Saudi Arabia. His research interests include the latest generation of wireless communication systems, such as 5G, LTE-A, LTE wireless networks, M2M, the Internet of Things (IoT), machine learning, OFDMA, and multimedia communications. He is a Fellow of the Higher Education Academy (FHEA). He has served as an Active Reviewer and received several distinguished reviewer awards from several reputable journals, such as the IEEE Transactions on Wireless Communications, the IEEE Transactions on Multimedia, the IEEE Transactions on Vehicular Technology, *Wireless Communications* (Elsevier), and *Computer Networks* (Elsevier). He recently won a funded project named Smart City and Adoption of 5G Technology in Saudi Arabia.