Cross-layer scheduling and resource allocation for heterogeneous traffic in 3G LTE

Musabe, Richard; Larijani, Hadi

Research Article

Cross-Layer Scheduling and Resource Allocation for Heterogeneous Traffic in 3G LTE

Richard Musabe and Hadi Larijani

*Department of Computer, Communication, and Interactive Systems (CCIS), School of Engineering and Built Environment,
Glasgow Caledonian University, Glasgow G4 0BA, UK*

Correspondence should be addressed to Richard Musabe; rmusab10@gmail.com

3G long term evolution (LTE) introduces stringent needs in order to provide different kinds of traffic with Quality of Service (QoS) characteristics. The major problem with this nature of LTE is that it does not have any paradigm scheduling algorithm that will ideally control the assignment of resources which in turn will improve the user satisfaction. This has become an open subject and different scheduling algorithms have been proposed which are quite challenging and complex. To address this issue, in this paper, we investigate how our proposed algorithm improves the user satisfaction for heterogeneous traffic, that is, best-effort traffic such as file transfer protocol (FTP) and real-time traffic such as voice over internet protocol (VoIP). Our proposed algorithm is formulated using the cross-layer technique. The goal of our proposed algorithm is to maximize the expected total user satisfaction (total-utility) under different constraints. We compared our proposed algorithm with proportional fair (PF), exponential proportional fair (EXP-PF), and U-delay. Using simulations, our proposed algorithm improved the performance of real-time traffic based on throughput, VoIP delay, and VoIP packet loss ratio metrics while PF improved the performance of best-effort traffic based on FTP traffic received, FTP packet loss ratio, and FTP throughput metrics.

1. Introduction

The third generation partnership project (3GPP) is a standards-developing body that specifies the 3G universal terrestrial radio access (UTRA) and global system for mobile communication (GSM) systems [1]. 3GPP projected 3G LTE technology to be the next technology standard with high data rates. The data rates provided by this technology include: 100 Mbps when transmitting from the eNodeB to the user equipment (downlink) and 50 Mbps when transmitting from the user equipment to the eNodeB (uplink), while supporting up to 20 MHz bandwidth [2]. Different versions of this technology has been released, the first release in 2008 (Release 8) was later modified to a new version (Release 9) and LTE-Advanced (release 10) [3]. LTE applies Orthogonal Frequency-Division Multiple Access (OFDMA) radio technology while moving in downlink direction. This radio technology helps LTE to provide control of wide carriers having high data rates at a low cost. This radio technology also allow LTE to accomplish the main objectives which

include providing spectrum flexibility. However, this radio technology has got one main disadvantage which is that it has a high Peak-to-Average-Power Ratio (PAPR). In order to solve this issue, LTE employs another radio technology called Single-Carrier Frequency-Division Multiple Access (SCFDMA) when moving in the uplink direction [4].

The main components of an LTE network include eNodeB (eNB) and different User Equipments (UEs). eNB is used to control the network core using different standard protocols. 3G LTE employs Physical Resource Blocks (PRB) to transmit resources. PRBs is composed of frequency and time domain phases [5]. eNB has got a specific amount of PRBs based on the assigned bandwidth, it also has the responsibility to distribute these PRBs constantly at each Transmission Time Interval (TTI) [2]. General packet scheduling can be employed by the network operator in the UEs and eNBs in either the uplink and downlink. The main issue is that there is no firm provisions that are set by the 3GPP for controlling the packet scheduling mechanisms. This has become an

open subject and in order to provide QoS requirements for different traffic types, different scheduling algorithms have been discussed in the literature.

Our contributions in this paper are the following.

(i) Formulate the problem of scheduling using utility function optimization by extending the proposed approaches in [3, 6] to include best-effort and real-time traffic metrics.

(ii) Use this technique to theoretically analyze our proposed algorithm in [7, 8] based on transmission rate, queue delay, and queue length parameters. It should be noted that we only developed the algorithm in [7]; it was later updated and mobility features added in [8]. Currently, the algorithm is updated again adding different features at different network layers and analyzing how this proposed algorithm handles different heterogeneous traffic in 3G LTE.

(iii) We investigate how our proposed scheduling algorithm is affected by both best-effort traffic (FTP) and real-time traffic (VoIP).

(iv) Using simulations, we study the performance of the proposed algorithm by comparing it with other algorithms in the literature that is, PF and EXP-PF in [9].

(v) Having incremented the utility function from [3], we also compared VoIP throughput metric of our proposed algorithm with that of U-delay algorithm in [3].

(vi) Based on the simulation analysis, our proposed algorithm performed better than PF and EXP-PF when real-time traffic is transmitted over an LTE network. However, PF performed better than our proposed algorithm when best-effort traffic is transmitted over an LTE network. Our proposed algorithm also had better VoIP throughput than that of U-delay in [3].

This paper is structured as follows. Section 2 presents some related work; Section 3 analyzes the heterogeneous traffic types used in this paper. Section 4 provides the the system model, where we discuss the general problem formulation and the extended version of the problem formulation. Section 5 discusses our proposed algorithm. Section 6 presents the simulation details, where PRB characteristics and scenario setup are discussed; we also provide the simulation results in this section. Section 7 reviews the main conclusions and future work.

2. Related Work

Guaranteeing QoS requirements to different traffic types in 3G LTE has been a major issue and different scheduling algorithms have been proposed in the literature to solve this issue. The proposed algorithms include the following.

A scheduling algorithm which employs a utility based game theoretic concept is proposed in [3]. The fact that LTE does not apply any rules to control user contentment while taking care of different traffic types limitations led to this proposal. This proposed algorithm is made up of two parts.

(1) Part One:

At this level, the algorithm assigns physical resource blocks between different LTE traffic types with dissimilar QoS objectives.

(2) Part Two:

At the second level, the delay based algorithm controls the delay plan needs for different LTE traffic types.

In this algorithm, the authors used a cooperative game concept in order to control how bandwidth is distributed between different LTE traffic services. This concept makes sure that LTE traffic services with low priority are assigned PRBs. These kinds of concepts are normally applied in scheduling algorithms which do not include fairness factor as a function of traffic congestion in specific LTE traffic types. This game concept makes sure that it assigns resources among different LTE traffic types by operating below the exact packet scheduling layer (network layer of the OSI model). This concept works as follows: the first LTE traffic types (inter-traffic types) are arranged to assign PRBs then after that, users inside each LTE traffic type (intratraffic type) are aligned based on the delay plan criteria for spectrum access.

To manage the distribution of PRBs among different LTE traffic types, the game concept is implemented at the eNodeB. This helps to make sure that high priority traffic type does not starve low priority traffic. So the gaming concept maximizes the utility function in order to assign PRBs between different LTE traffic types.

Suppose $F \in Z^+$ is the number of LTE traffic types and the PRBs allocated among LTE traffic type is $N = \{1, 2, \ldots, n\}$. The data rate of each traffic type with respect to the PRBs assigned to traffic type users is $d = \{1, 2, \ldots, D_F\}$. So to calculate what each user (player) n_i in a specific LTE traffic type can get is done from the sum utility function for that traffic type. This is done as follows:

$$R_i(d_i) = f_i * U_i(t_i, r_i, d_i), \tag{1}$$

where, R_i is the sum of utility function for traffic type d_i. f_i is the total number of flows or connection in the traffic type where n_i user (player) belongs to. t_i is the priority type of the LTE traffic types of a user, r_i is the essential resource needs of the LTE traffic type of a user, and U_i is the utility function.

Suppose the whole network is considered including all users in all traffic types; this can be recalculated as follows:

$$P = \sum_{i=1}^{F} R_i(d_i) = \sum_{i=1}^{F} f_i * U_i(t_i, r_i, d_i), \tag{2}$$

where P is the total network profit.

So the overall network optimization can be calculated from maximizing both single user profit $R_i(d_i)$ and whole network profit (P). This maximization will result in maximizing the entire system throughput. So the overall maximization framework can be seen as

$$\max \sum_{i=1}^{F} R_i(d_i) = \max \sum_{i=1}^{F} f_i \cdot U_i(t_i, r_i, d_i) \tag{3}$$

with

$$\sum_{i=1}^{F} f_i * (d_i) = F * d_i^T \le Cp, \qquad (4)$$

where Cp is the network total available capacity.

In [6], Huang et al. addressed the gradient-based scheduling issue in downlink OFDM systems. The authors analyzed different features such as resource allocation, various sub-channelization criteria, signal-noise-ratio constraints, and channel estimation errors. In every time slot a group of users are scheduled and the available resource blocks as well as power are assigned to them. Using the gradient based approach, the authors analyzed this problem as an optimization problem which can be solved in every time slot. Applying the dual formulation. The authors were also able to provide an optimal algorithm for this problem with several users being able to share resources. Their approach motivated us to further address the problem of gradient-based scheduling and resource allocation. We specifically focused on best-effort (FTP) and real-time (VoIP) traffic when transmitted over an LTE network. In our approach we considered various parameters provided in the channel state information such as transmission rate. We also used the parameters provided at the MAC layer that is, (queue length).

In [9, 10], the authors proposed a Proportional Fair algorithm (PF). The main aim of this algorithm is to maximize the total network utility so that it can improve the system throughput and to ensure fairness between traffic flows. It allocates radio resources based on channel quality as well as past user throughput [11]. This scheduler uses the metric which is defined as the ratio between the immediate present data rate and the average past rate in accordance to the jth traffic flow in the ith traffic flow subchannel. This can be depicted as obtained from [9]:

$$m_{i,j} = \left(\frac{r_{i,j}}{R_{i,j}} \right), \qquad (5)$$

where $m_{i,j}$ is the transmission rate, $R_{i,j}$ is the estimated average data rate, and $r_{i,j}$ is the instantaneous available data rate. $r_{i,j}$ is calculated by the adaptive modulation code (AMC) module which analyzes the channel quality indicator (CQI) feedback that is sent by the UE hosting the jth traffic flow to the ith subchannel. It should be also noted that i and j are sub channel flows. This scheduling algorithm is more suitable for best-effort traffic.

In [9, 10, 12], the authors proposed an Exponential/ Proportional Fair Algorithm (EXP-PF). This algorithm is basically aimed at taking control of real-time traffic compared to best-effort traffic, which has their head-of-line packet delay approaching the delay threshold [11]. The real-time traffic metrics are computed in [9] using the following equations:

$$m_{i,j} = \exp\left(\frac{\alpha_i D_{\mathrm{HOL},i} - X}{1 + \sqrt{X}} \right). \qquad (6)$$

The variable X in (6) can be obtained as follows:

$$X = \frac{1}{N_{r,t}} \sum_{i=1}^{N_{r,t}} \alpha_i D_{\mathrm{HOL},i}, \qquad (7)$$

where $N_{r,t}$ is the active downlink real-time traffic flows. α_i in (7) can be described as the probability for $D_{\mathrm{HOL},i}$ to exceed the delay threshold. If we consider the packet threshold to be T_i, then α_i in (7) can be calculated as follows:

$$\alpha_i = -\frac{\log \alpha_i}{T_i}. \qquad (8)$$

Equations (7) and (8) proposed in [9] calculate the average total of the entire downlink real-time traffic flows based on the probability that the initial packet to be conveyed in the queue surpasses the delay threshold. This helps to prioritize downlink real-time traffic flows.

In [13], the authors provided an overview on the key issues that arise in the design of a resource allocation algorithm for LTE networks. The authors analyzed most aspects of LTE networks with the main focus being set on the downlink channel under frequency division duplex configuration. This paper also provided the performance comparisons of the most well-known scheduling schemes, with particular focus on QoS provisioning capabilities. This comparison was performed through system-level simulations. The authors also highlighted different pros and cons of these scheduling schemes. The compared scheduling schemes were divided into five groups.

(i) *Channel-Unaware*: channel unaware strategies are based on the assumption of time-invariant and error-free transmission media. While their direct application in LTE is not realistic, they are typically used jointly with channel-aware approaches to improve system performance. The schemes that apply this strategy include First in First out (FIFO), Round Robin, Blind Equal Throughput, Resource Preemption, Weighted Fair Queuing, and Guaranteed Delay.

(ii) *Channel-Aware/QoS-Unaware*: this strategy uses CQI feedbacks, which are periodically sent (from UEs to the eNB) using ad hoc control messages. The scheduler can estimate the channel quality perceived by each UE using these CQI feedbacks; hence, it can predict the maximum achievable throughput. The schemes that apply this strategy include Proportional Fair Scheduler, Joint Time and Frequency domain schedulers, and Buffer-aware schedulers.

(iii) *Channel-Aware/QoS-Aware*: QoS differentiation is handled by associating a set of QoS parameters to each flow. Knowing the values of such parameters, the scheduler can treat data to guarantee some minimum required performances, either in terms of guaranteed data rates or of delivery delays. it is important to note that QoS-awareness does not necessarily mean QoS provisioning, since it consists of taking allocation decision depending on the requirements of each flow, without necessarily guaranteeing the meeting of such requirements, because it could be unfeasible if procedures for admission control are not implemented. The schemes that apply this strategy include Schedulers for Guaranteed Data-Rate, Schedulers for Guaranteed Delay Requirements, and Dynamic Schedulers for VoIP support.

(iv) *Semipersistent for VoIP Support*: semipersistent allocation solutions aim at increasing the VoIP capacity of the network in terms of maximum number of contemporary supported VoIP calls. They are not specifically conceived for improving spectral efficiency or for reducing packet delay and PLR. They can be considered in practice as channel-unaware approaches. Anyway, using them it is possible to indirectly improve performance as the number of scheduled users increases.

(v) *Energy-Aware*: energy saving solutions can be applied to both eNB and UE. For end-user devices, power consumptions can be limited through discontinuous reception (DRX) procedures and the persistent allocation, which is the only allocation strategy able to meet this goal.

All these algorithms proposed in the literature have not solved all the problems of scheduling and resource allocation for different traffic types. Most of the research on this issue tends to focus on improving the QoS of real-time traffic without unambiguously taking into account the best-effort traffic. Even though the trend of real-time application development and use is on a rise, best-effort applications are still expected to be among the prevailing percentage of the internet traffic. So in this paper, we are going to analyze the effect of our proposed algorithm on both real-time and best-effort traffic. Our proposed algorithm extends the work in [6], where the problem of scheduling and resource allocation was formulated using the utility function optimization approach. We introduce the best-effort and real-time traffic metrics to this approach in place of power allocation metrics which was considered in [6]. We also extended the utility function used in [3] and introduce the channel state information such as transmission rate at the physical layer as well as the queuing state information, i.e., queue length, at the MAC layer in place of using the gaming concept used in [3]. Using simulations we analyze the performance of our proposed algorithm by comparing it with other algorithms in [9], PF, and EXP-PF. This comparison was done for both best-effort traffic (FTP) as well as real-time traffic (VoIP) and the metrics used in this comparison are throughput, delay, packet loss, traffic sent, and traffic received. The simulation results were generated using the open source LTE system simulator called LTE-SIM [9].

3. Heterogeneous Traffic Types

Before analyzing different scheduling algorithms proposed in the literature, let us first present a brief description of the two traffic types that will be used in this paper.

3.1. Best-Effort Traffic (BE). A network with best effort traffic means that network users receive best-effort services with no guaranteed variable bit rate or delivery time. Users receive services based on the available network traffic load. The main best-effort traffic metrics include throughput and responsiveness [14]. In this paper, the best-effort traffic is generated by TCP-based flows and it includes File Transfer Protocol (FTP) traffic services.

3.2. Real-Time Traffic (VoIP). In this paper we mainly focused on VoIP traffic as our real-time traffic. VoIP is a method of conveying traffic in form of data packets across an IP network. Voice traffic needs to be changed into digital signals before being compacted into a sequence of packets. These sequences of packets can then be added up together and decoded at the receiver. This process can occur before or simultaneously with packetization [15]. VoIP standards have risen efficiently because of providing advantages which include cheap calls, the use of VoIP, and other best-effort traffic on the same network [16]. Normally voice as well as data traffic would be conveyed over two different networks (circuit-switched network). However, the existence of VoIP standards has led these two different traffic types to be conveyed using the same network (packet-switched network) [17]. This VoIP technology comes with some issues and the main one is the QoS. QoS offered by this method is far worse than the QoS offered by the previous method when data and voice traffic were transmitted by separate networks. These issues are brought about by the fact that VoIP traffic is influenced by different factors which include delay, jitter, and packet loss. These factors heavily affect the quality of voice [18].

The conversation VoIP traffic in LTE can be assumed as the two state Markov models containing an appropriate voice activity factor (VAF). Different open source Codecs can be used in LTE but the most common used codec according to [19] is Adaptive Multirate (AMR). AMR offers 32-bytes voice stack in 20 milliseconds while talking and 7-bytes payload every 160 millisecond while silent. The VoIP protocol stack is made up of three protocols which are Real-Time Transport Protocol (RTP), User Datagram Protocol (UDP), and Internet Protocol (IP). The addition of all these protocols results into a 40 bytes IPv4 header or a 60-byte IPv6 header. This overhead problem can be solved by the use of robust header compression (ROHC) technique. This technique can reduce the header to as little as 2 or 4 bytes [5, 20].

3.2.1. VoIP End-to-End Delay and Capacity. The main characteristic of voice traffic is severe delay requirements [21]; according to [22] the approved highest mouth-to-ear delay for voice traffic is 250 ms. These delay requirements include different assumptions.

(i) Core network delay is approximately 100 ms.

(ii) Delay for radio link control (RLC), medium access control (MAC) buffering, scheduling, and detection should be below 150 ms.

(iii) If we take an assumption that only LTE users are in the network, then we can assume that the reasonable delay for buffering and scheduling should be below 80 ms.

(iv) In order to budget for unpredictability in network delays, 3GPP performance evaluation has also chosen the delay of 50 ms when moving in the downlink direction [21].

Most of these features can be seen in Figure 1 [2]. Packets will plummet if the error rate and delay exceed the limits of the set latency, while VoIP traffic is transmitted over an LTE
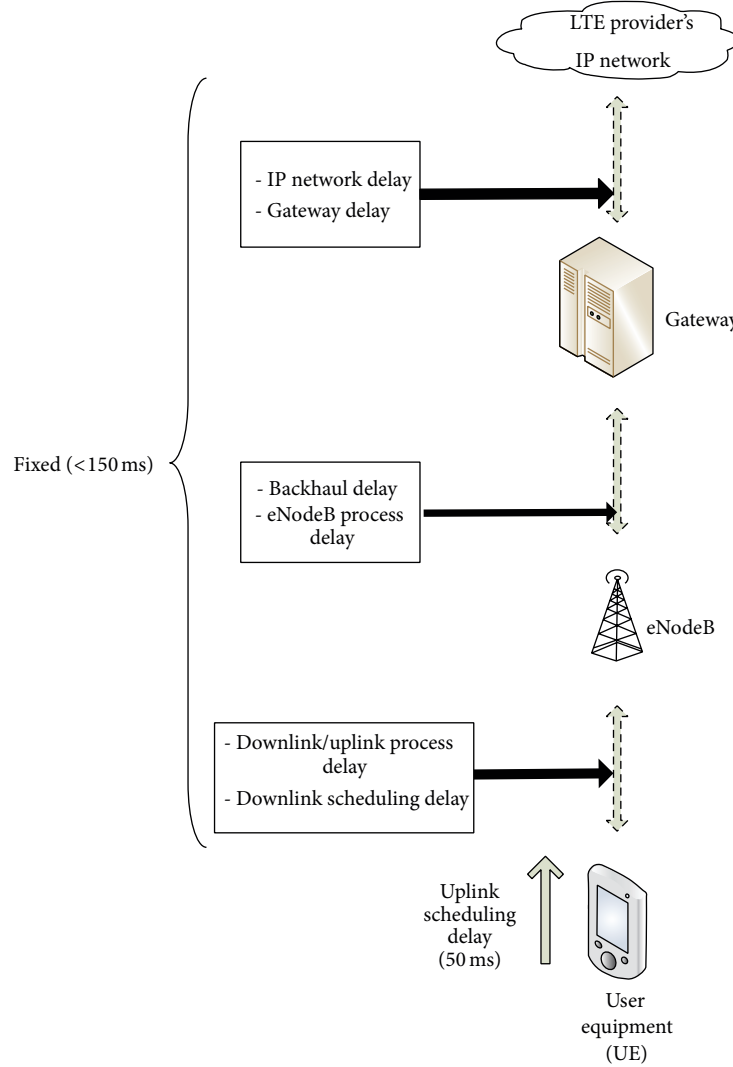
FIGURE 1: VoIP end-to-end delay components in E-UTRAN (LTE).

network. This may not affect the voice quality if the packet loss is less than the set threshold [21]. The set limit implies that the error rate for VoIP users is kept below 2%. This gives us the actual idea that the highest VoIP capacity for LTE is given by the specific limit which is described in TR 25.814 [5] and was later updated in R1-070674 [23]. We can finally describe VoIP capacity in LTE as the highest number of VoIP users that the network can sustain while still in the threshold limits and 95% of all VoIP users should be within the proposed confines [2].

4. System Model

4.1. General Problem Formulation. In [6] the authors considered sending packets from the eNodeB to UE in a network which had the eNodeB and number of UEs. K was considered to be the maximum number of available users such that $K = \{1, \ldots / K\}$. So in every time slot, assigning of resources was done by choosing the rate vector $r_t = (r_{1,t}, \ldots, r_{K,t})$ from $R_{e_t} \subseteq R_+^K$, with e_t being the channel quality. In short,

the general problem is to find $r_t \epsilon R(e_t)$ that can maximize the system utility function $U(W_t) := \sum_{i=1}^{K} U_i(W_{i,t})$, with $U_i(W_{i,t})$ being the utility function of user i's average throughput $W_{i,t}$.

4.2. Our Extended Version of the Problem Formulation

4.2.1. Utility Function. Before extending the problem formulation in [6] to include physical/MAC layer parameters and other real-time traffic (VoIP) parameters, let us first describe the utility function. Utility functions can be useful in cross layer optimization as they can map network resources utilized by users into real numbers. The utility function can also indicate the level of satisfaction of the user which in turn helps in balancing the efficiency and fairness between the users. In 3G LTE, just as most wireless communication technologies, consistent transmission rate is the main factor that can determine the level of satisfaction of the user. So if we take m_j to be the transmission rate vector, then the utility $U(m_t)$ should be a nondecreasing function of the transmission rate m_j.

We adopted the utility function calculation from [3] and used it for the transmission rate m_t as

$$U\left(m_j\right) = X_j \left\{ \frac{1}{1 + e^{-p_j(m_j - R_j)}} - Y_j \right\}, \qquad (9)$$

with

$$X_j = \frac{1 + e^{p_j R_j}}{e^{p_j R_j}},$$
$$Y_j = \frac{1}{1 + e^{p_j R_j}}, \qquad (10)$$

where $U(m_j)$ is the utility function of user j with respect to their transmission rate. p_j is the priority tag assigned to VoIP users. R_j is the available resource blocks. X_j and Y_j are constants used to normalize the utility function.

4.2.2. Problem Formulation. The main aim of this problem formulation is to map the network resources of each user to their corresponding utility values. After that, the established utility function is optimized. Let K, indexed by j, be the maximum number of available users such that $j \in \{1, \dots, K\}$. If we consider the utility function of user j to be $U_j(\cdot)$, then if user j has the transmission rate of m_j, we can say that the utility of user j is $U_j(m_j)$. Again if we let Q_l be the length of user j's queue and Q to be the total number of queues for user j.

Then, the total utility function of user j is calculated from the utility function of user j's queue

$$\text{total-utility}_j = Q_l * U_j\left(p_j, R_j, m_j\right), \qquad (11)$$

where $U_j(p_j, R_j, m_j)$ can be equal to $U(m_j)$ in (9).

If we take all the user j's queues in the network, then

$$\text{Total-utility}_j = \sum_{j=1}^{Q} Q_l * \sum_j U_j\left(p_j, R_j, m_j\right)$$
$$= \sum_{j=1}^{Q} Q_l * \sum_j U_j\left(m_j\right). \qquad (12)$$

So, our problem is to find a user that can maximize the total utility with respect to transmission rate, as well as the user queue with packets having higher waiting time:

$$\max \sum_{j=1}^{Q} \text{total-utility}_j = \max \sum_{j=1}^{Q} Q_l * \text{Max} \sum_j U_j\left(m_j\right). \quad (13)$$

However, the fact that we are dealing with real-time applications such as VoIP means that we need some constraints to control the QoS requirements. So, the above optimization objective equations should be subject to $m_j \leq$ NC and $Q_d \leq D_{\max}$, where NC is the total available network capacity, Q_d is the queuing delay, and $D_{\max}$ is the maximum allowable mouth to ear delay.

Having seen that the main metrics in our problem formulation procedure are transmission rate (m_j) and queue length (Q_l), it is very important to know how these two metrics are obtained. This is described below.

4.2.3. Finding the Transmission Rate (m_j). During every transmission process, the user sends the instantaneous achievable signal to noise ratio (SNR) to their respective eNodeB. This value keeps on changing depending on different factors like mobility, selective fading channels, and so forth. So according to [3], user j' transmission rate can be calculated as

$$m_j(t) = \frac{n_{\text{bits}}}{\text{symbols}} * \frac{n_{\text{symbols}}}{\text{slot}} * \frac{n_{\text{slots}}}{\text{TTI}} * \frac{n_{\text{subcarriers}}}{\text{RB}}, \qquad (14)$$

where n_{bits}, n_{symbols}, n_{slots}, and $n_{\text{subcarriers}}$ are, respectively, the number of bits, number of symbols, number of slots, and number of subcarriers according to the PRB characteristic described earlier.

These PRB characteristics are affected by path loss and fading channels but their values are kept constant for the entire PRB transmission time. According to [24], the channel gain of user j on a PRB at time t as a function of loss is calculated as

$$\text{CN}_{(\text{gain}_j)}(t) = 10^{\text{pathloss}/10} * 10^{\text{fading}/10}. \qquad (15)$$

Employing this channel gain (CN_{gain}), the user knows the instantaneous SNR to send to eNodeB. Again according to [25], this SNR can be calculated as a function of CN_{gain}:

$$\text{SNR}_j(t) = \frac{P_{\text{total}} * \text{CN}_{\text{gain}}(t)}{R\left(N_o + I\right)}, \qquad (16)$$

where P_{total} is the power with which the eNodeB transmits, R is the total number available PRBs, I is neighboring interference, and N_o is the thermal noise measure.

4.2.4. Finding the Queue Length (Q_j). In order to obtain queue length metric, we adopted the queuing method in the LTE-SIM simulator. In this method, different traffic generators were developed. Using the application class, we were able to generate the packets and deliver them to the network. Once the packets reach the network, they are forwarded to the user-plane protocol stack to add protocol headers. Then, the packets are placed in the queue by the MAC queue class at the MAC layer before being sent to the destination. The MAC queue object has got a counter which increases or decreases when the packet is inserted or removed from the queue, respectively.

If we consider $Q_j[l]$ to be packets in the queue of user j at time T_s, so if during the time slot n, the eNodeB attends to user j at rate $r_j[n]$, the queue length of user j at time $(n+1)T_s$, $Q[n+1]$ is deliberate as

$$Q_j[n+1] = Q_j[n] - r_j[n] T_s + a_j[n], \qquad (17)$$

where $a_j[n]$ is the arrival bits in time slot n.

4.2.5. Solution Approach. In order to solve our optimization problem in (13) that will maximize the network utility, we used the dual decomposition approach with Lagrange multipliers. Solving this equation determines the user to be scheduled. This same user will be assigned resource blocks

based on transmission rate (m_j) and queue values (Q_l, Q) parameters subject to $m_j \leq NC$ and $Q_d \leq D_{max}$ constraints. Writing up the optimization problem as a Lagrange dual function, it becomes

$$L\left(m_j, Q_d, \lambda, \mu\right) = \sum_j U m_j + \lambda \left(NC - \sum_j m_j \right)$$
$$+ \mu \left(D_{max} - \sum_j Q_d \right). \tag{18}$$

The corresponding dual function can be written as

$$L\left(\lambda, \mu\right) = \max_{m_j, Q_d} L\left(m_j, Q_d, \lambda, \mu\right). \tag{19}$$

The inequality constraints in the optimization problem are put under consideration by augmenting the objective function with a weighted sum of the constraint function. Therefore, λ is called the lagrangian multiplier associated with $m_j \leq NC$ constraint and μ is the lagrangian multiplier associated with $Q_d < D_{max}$ constraint.

If we divide the above objective function into $|\lambda|$ and $|\mu|$ separate subproblems, then each subproblem can be solved separately if the values of λ and μ are known. the objective function of the dual problem then becomes

$$D\left(\lambda\right) = \max_{m_j \varepsilon NC} L\left(NC, \lambda\right), \tag{20}$$

$$D\left(\mu\right) = \max_{Q_d \varepsilon Q_{max}} L\left(Q_{max}, \mu\right). \tag{21}$$

5. The Proposed Algorithm

Detailed explanations of this algorithm can be found in [7, 8]. In our proposed algorithm, the scheduler assigns resource once every TTI and based on the user's current transmission rate (m_j), queuing delay (Q_d), and queue length (Q_l). The first part of our algorithm is to arrange users to schedule. The arrangement of users to schedule is based on their decreasing order of their queuing delay (Q_d) and queue length (Q_l). Once the arrangement is done, the PRB assignment is made by taking every user and determining the parameters that can maximize the utility of transmission rate (m_t). In order not to starve other applications in the network, we used the procedure employed in [2].

This method provides limits to our proposed scheduling algorithm, which is adaptively changed between a prespecific minimum and maximum according to the ratio of dropped packets. Higher drop ratio means that there are many ongoing VoIP calls, and hence it is necessary to increase the limits to allow more consecutive TTIs to be dedicated to VoIP calls. On the other hand, low drop ratio implies that QoS of VoIP calls are satisfied at decent levels, and thus it is safe to reduce the duration of the algorithm and serve other applications in the network.

5.1. Steps of the Proposed Algorithm

Step 1. Determine the procedure of inserting users/packets into their queues.

Step 2. Scheduling starts at every TTI.

Step 3. Find out if there are any users/packets in the queues.

Step 4. If there are users/packets in the queue, apply our proposed algorithm and go to the next step otherwise apply the normal scheduling algorithm and exit.

Step 5. Sort the users/packets according to the decreasing order of their queuing delay (Q_d) and queue length (Q_l); initialize $j = 1, m_{ext} = m_{max}$, and $Q_{ext} = Q_{max}$, where m_{ext} and Q_{ext} are the extra or remaining transmission rate and queue length values at each stage.

Step 6. Determine if the successive counts of our proposed algorithm are not greater than the provided adaptive limit.

Step 7. If it is not greater than the limits then go to the next step; otherwise apply the normal algorithm, that is, default algorithm such as FIFO and exit.

Step 8. Find the parameter that maximizes the transmission rate (m_j) for user j with $m_j \leq m_{ext}$ and $Q_j \leq Q_{ext}$.

Step 9. Schedule this user.

Step 10. Reduce m_{ext} by m_j and Q_{ext} by Q_j, respectively.

Step 11. If more resources blocks (RBs) and users exist, as well as $m_{ext} > 0, Q_{ext} > 0$ then set $j = j + 1$ and repeat from Step 8. If any of the three checks fails, then exit.

The algorithm flow chart is presented in Figure 2.

6. Simulation Details

6.1. PRB Characteristics and Scenario Setup. Before we go into the details of our simulation setup, lets first give a brief description of PRBs, which are seen as the communication resources in LTE. LTE systems can be analyzed both in time and frequency phases. The time phase can be separated into 1 ms TTI that is made up of 2 slots where each slot is 0.5 ms to make 1 ms subframes. Every subframe has 7 OFDMA symbols. Every TTI has 14 OFDMA symbols, in these 14 OFDMA symbols, 2 of them are used for uplink communications; the remaining symbols are applied to the data and control information transmission. TTIs are seen to be the smallest share within the time phase. However, from the frequency phase perspective, this minimum allocation unit can be described as the PRB. This PRB has 12 subcarriers of 15 kHz bandwidth each. The amount of OFDMA symbols in a PRB is determined by the cyclic prefix used. Figure 3 shows all these procedures. As far as VoIP packets are concerned, they must be conveyed in every TTI [18].
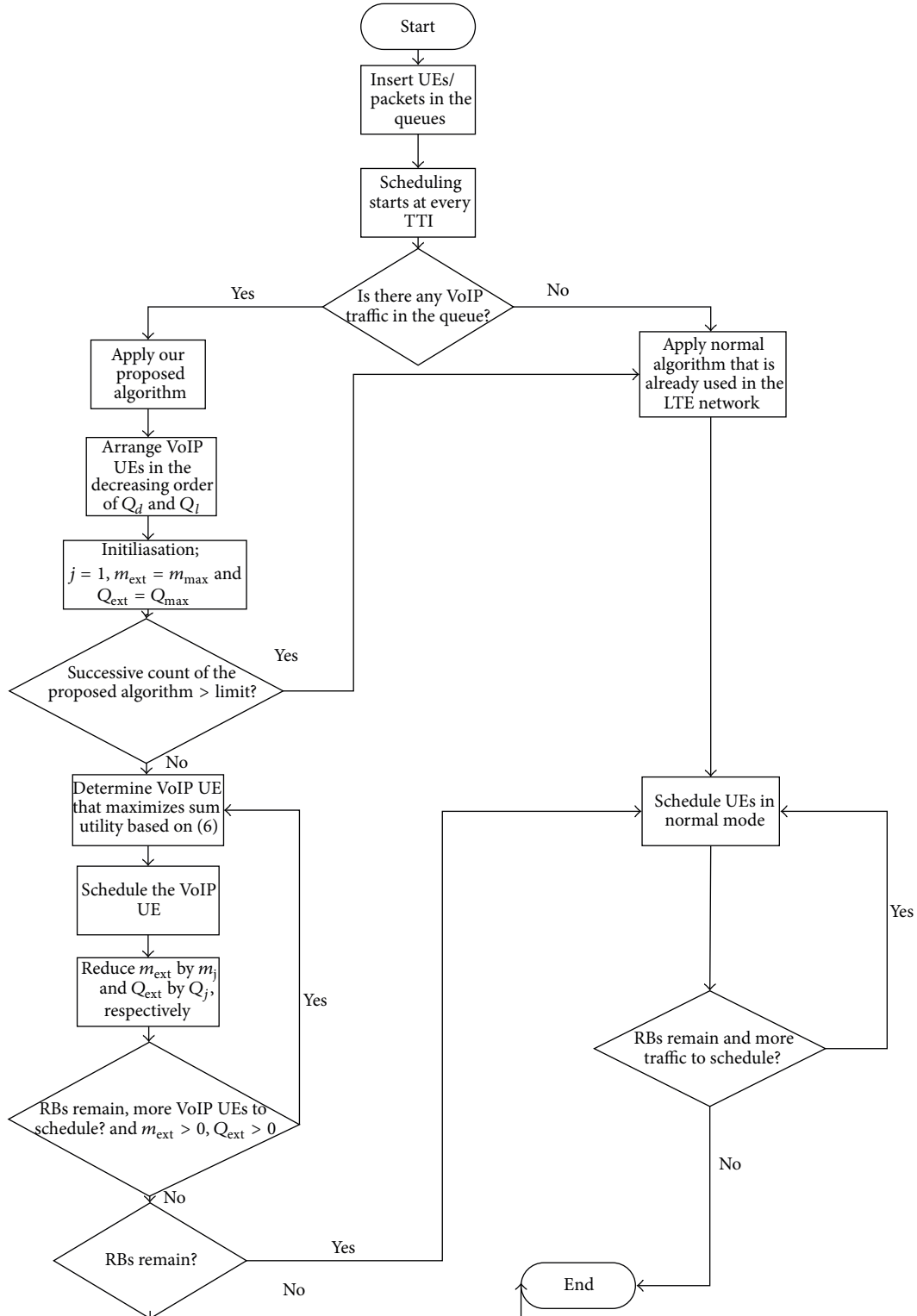
Figure 2: Algorithm flow chart.

We designed our network with a number of cells and network elements which are the eNodeB, mobility management/ gateway (MME/GW), and user equipments (UEs). We had 57 cells in total which were divided into 3 sectors with each sector having 19 cells [9]. Some of the parameters used in our simulation procedures can be seen in Table 1. The simulation software used here (LTE-Sim) has different kinds of traffic generators and we used one of them to generate VoIP traffic flows and another to generate FTP flows. The FTP traffic flows were generated using TCP flows. The VoIP traffic generator
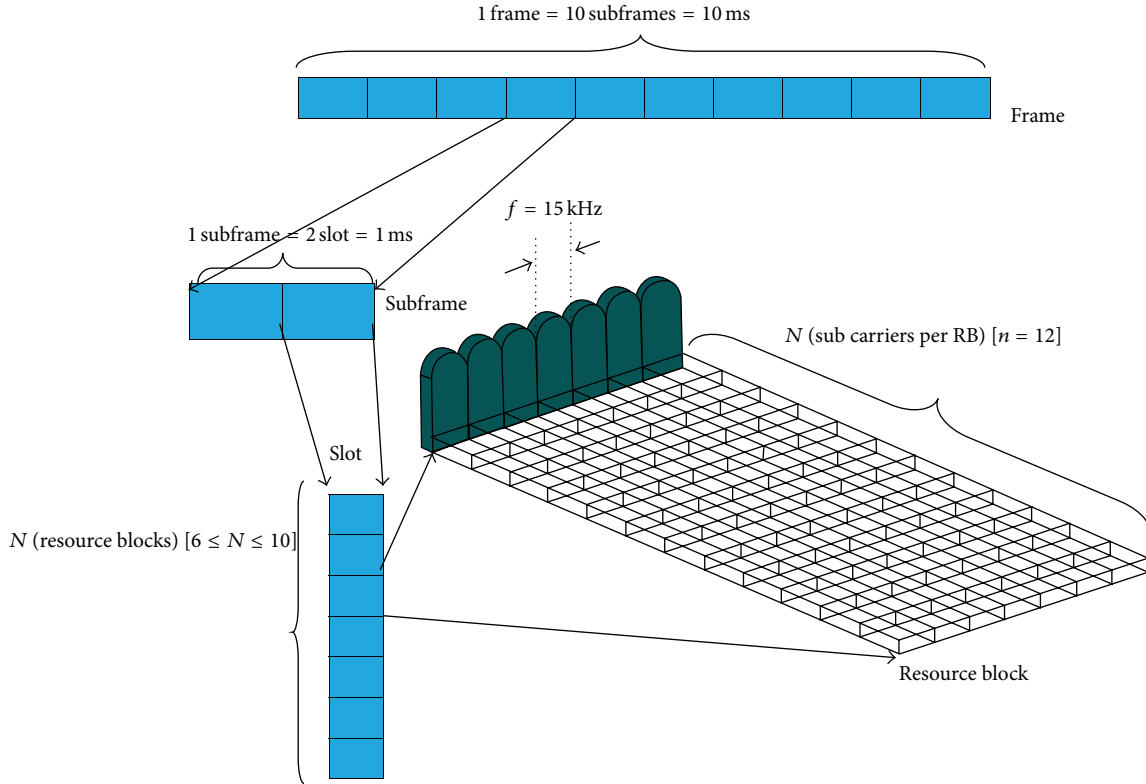
FIGURE 3: Resource grid.

TABLE 1: Simulation parameters.

Simulation parameters	Values
Bandwidth	5 MHz
PRB structure	12 subcarrier and 2 subframes
TTI	1 msec
Number of available PRBs	25
Modulations for AMC	QPSK
Number of sectors	3
Simulation time	1000 TTIs
Number of iterations	5
Cyclic prefix	Normal
Scheduling algorithms	Our proposed algorithm, EXP-PF, and PF
Cell radius	1 Km

used the ON/OFF Markov chain to replicate the voice flows. When the ON period is active, the sending user transmits packets which are of 20 bytes in size. However, when the OFF period is activated, the transmission rate is set to zero since the presence of voice activity detector is assumed. Simulations were run for a number of iterations and in every iteration the seed number was updated. This was done in order to analyze the accuracy and the confidence interval of our simulation results.

6.2. Simulation Results. We used LTE-SIM to analyze the performance of our proposed algorithm for heterogenous traffic types in 3G LTE. We compared our proposed algorithm with that of other scheduling algorithms: EXP-PF, PF, and U-delay proposed in the literature. The metrics used in this analysis are throughput, delay, packet loss ratio, traffic sent, and traffic received.

6.3. Real-Time Traffic Type. As described earlier in this paper, the real-time traffic type used is the VoIP application flows. The metrics used are VoIP delay, throughput, and VoIP packet loss ratio.

6.3.1. Throughput Comparison. We analyzed user throughput for all three schedulers while gradually increasing the number of VoIP users. The fact that VoIP packets are very small makes it hard for them to use all the available PRBs efficiently. However, as presented in Figure 4, it is clear that system throughput improved as VoIP users increased, implying that the PRBs utilization had improved. Analyzing Figure 4, we can see that our proposed algorithm has better throughput than that of EXP-PF and PF presented in [9]. These simulation results indicate that our algorithm improves the system's performance. This in return assures the QoS of VoIP users.

6.3.2. VoIP Delay Comparison. We also investigate VoIP delay while gradually increasing the number of VoIP users. Figure 5 shows the VoIP delay. Looking at Figure 5, VoIP delay increased when more VoIP users (i.e., from 6 users to 12 users) were introduced in the network. This is mainly
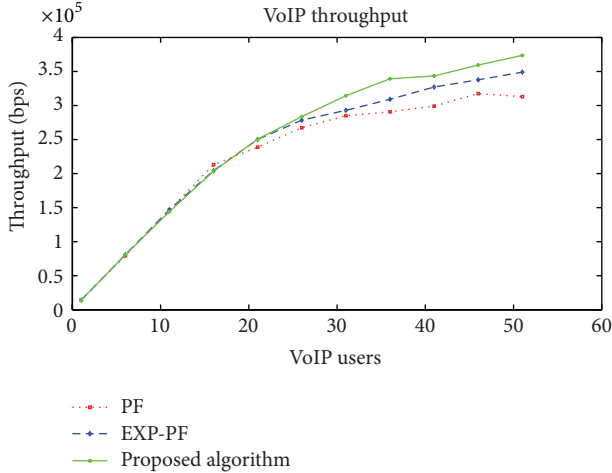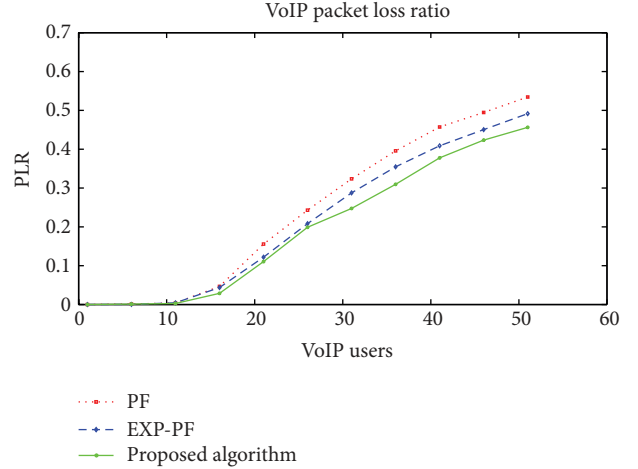
FIGURE 4: Throughput.
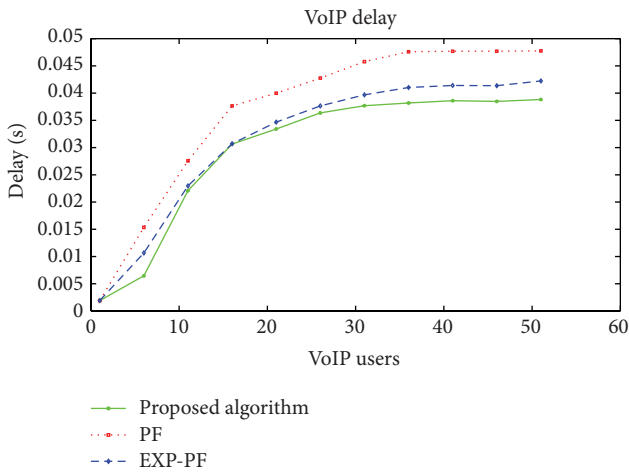

FIGURE 6: VoIP packet loss ratio.

FIGURE 5: VoIP delay.

due to the fact that more users in the network utilize more resources and hence more packets are being dropped. Even though delay was increasing, it did not affect much the QoS of voice quality, because when we look at [22], the approved delay is 250 ms. The improvements in delay are mainly due to the improvements made on this network that we described earlier. Even though there was an increase in VoIP delay when more VoIP users were introduced in the network for all the three schedulers, there are some differences in the three schedulers. When PF and EXP-PF are employed, the VoIP delay increases more than when our proposed algorithm was used. The reason for this, is that our proposed algorithm employs a simple method of allocating resource blocks and scheduling VoIP users which is affected to a lesser degree by the high load factor as compared to the other two scheduling algorithms that employ packet deadline expiration procedure which is highly affected by the high load factor.

6.3.3. VoIP Packet Loss Ratio Comparison. The packet loss ratio is investigated and is shown in **Figure 6**. VoIP PLR

increased when more VoIP users were introduced in the network. Once again, even though the PLR was increasing as the number of VoIP users increased, it did not affect the QoS of VoIP in the network due to the improvements on the transmission rate vector of the network. These techniques improved the overall network performance. Even though there was an increase in VoIP PLR when more VoIP users were introduced in the network for all the three schedulers, we note some differences in the three schedulers. When PF and EXP-PF are employed, the VoIP PLR increases more than when our proposed algorithm is used; this is due to the fact that, with PF and EXP-PF schedulers, when there are many simultaneous real-time packets, the possibility of ejecting the real-time packets that have passed their deadline increases [15]. However, with our proposed algorithm, we do not calculate the deadline expiration factor for VoIP packets; it employs a simple method of scheduling users based on simple metrics, as well as the availability of the resource blocks.

6.4. Best-Effort Traffic Type. The best-effort traffic types used in this paper are FTP traffic. Throughput, FTP packets dropped ratio, FTP traffic sent and received are the metrics used to investigate how the three algorithms handle best-effort traffic.

All the scheduling algorithms sent almost the same amount of FTP traffic as it can be seen in **Figure 7**. All algorithms had the same amount of FTP users hence the same amount of traffic sent.

However, the amount of FTP traffic received is different for all the three scheduling algorithms. This is due to the fact that as FTP traffic is transmitted over an LTE network, some packets will be dropped and based on the type of the scheduler used, different levels of packets are dropped. As it can be seen in **Figure 8**, when using PF scheduling algorithm more FTP traffic is received than our proposed algorithm and EXP-PF scheduling algorithms. This shows that PF is more suitable for best-effort traffic than our proposed algorithm and EXP-PF.
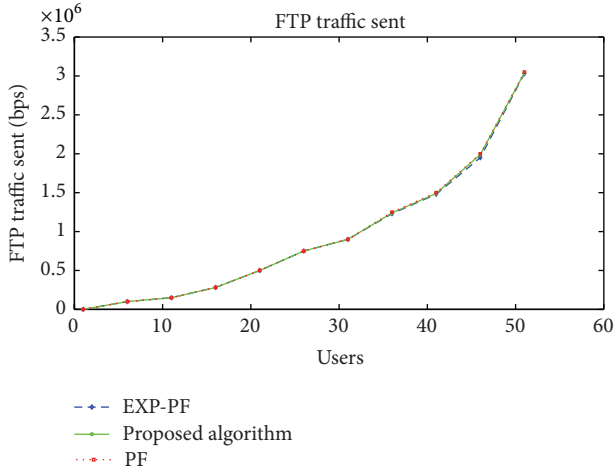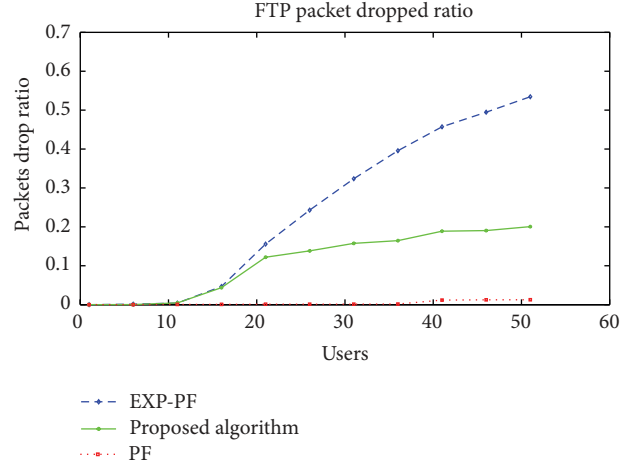
FIGURE 7: FTP traffic sent.
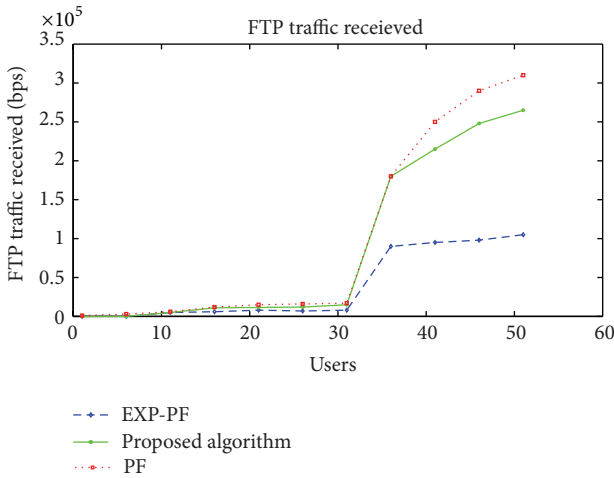

FIGURE 9: FTP packet loss ratio.
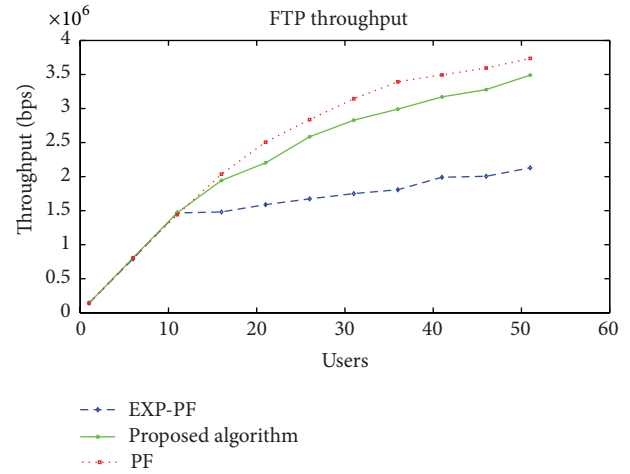

FIGURE 8: FTP traffic received.

FIGURE 10: FTP throughput.

This can be emphasized by the packet loss ratio for all three schedulers shown in Figure 9, PF has lower packet loss ratio than our proposed algorithm and EXP-PF. Our proposed algorithm and EXP-PF are more suitable for real-time traffic as they increase the priority levels for real-time traffic as opposed to best-effort traffic. However, our proposed algorithm uses the adaptive method to control the time that is applied to real-time traffic with in an LTE network. This helps in not starving best-effort traffic.

To analyze the efficiency of all three scheduling algorithms when applied to best-effort traffic, we investigated their throughput. Throughput determines the number of successfully delivered packets over a specific link or channel. As seen in Figure 10, PF again has better throughput than our proposed algorithm and EXP-PF. So in short we can say that PF scheduling algorithm is the best for nonreal-time traffic.

6.5. Comparing Our Proposed Algorithm to U-Delay Algorithm in [3]. In order to analyze the improvement made by our proposed algorithm compared to the algorithm presented

in [3], we compared VoIP throughput of our proposed algorithm to that of U-delay in [3]. Having incremented the utility function from [3], it was necessary to analyze if there was any improvement made by our algorithm from that in [3]. Different metrics could have been used for this analysis but in this paper, VoIP throughput metric was used due to the fact that this metric can show the overall performance of the two algorithms. As presented in Figure 11, our proposed algorithm had better VoIP throughput than that of U-delay in [3]. VoIP throughput of our proposed algorithm increased as the VoIP users increased; this is due to the fact that having incremented the utility function in [3], our proposed algorithm improved PRBs utilization hence improving the system performance of the scheduling algorithm.

7. Conclusion and Future Work

We analyzed the issue of scheduling and PRB allocation for heterogeneous traffic types in 3G LTE, where we projected the packet scheduling and PRB allocation issue as a constrained

VoIP throughput

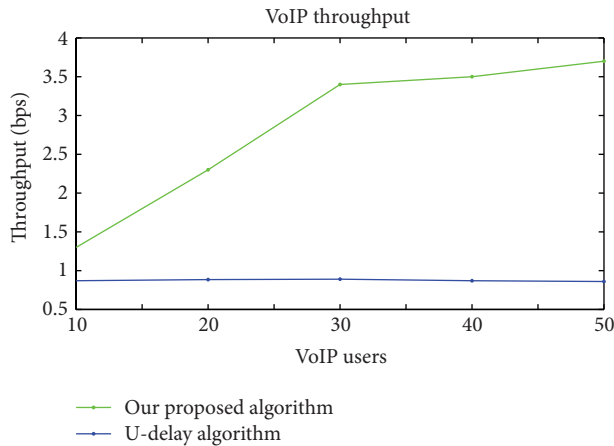——— Our proposed algorithm
——— U-delay algorithm

Figure 11: VoIP throughput of our proposed algorithm VS U-delay in [3].

optimization problem. We provided the optimization objective with the aim of maximizing the expected total utility function under different constraints. Finally, we provided the algorithmic implementation of the proposed algorithm and also studied the performance of the proposed algorithm under different conditions and compared it with other algorithms in the literature.

Our proposed algorithm uses the metric maximization procedure to assign resource blocks to different users with different types of traffic flows. The main metrics used being queue length and transmission rate, this procedure makes our proposed algorithm less complex and it is executed in a short time.

Regarding performance analysis, we analyzed the effect of our proposed algorithm, EXP-PF, and PF scheduling algorithms to the best-effort and real-time traffic. In the real-time traffic (VoIP) scenario, our proposed scheduling algorithm improved the performance of real-time traffic by approximately 10–20 percent in terms of less VoIP packet loss ratio as the number of users increased and less VoIP delay as the number of users increased. The fact that our proposed algorithm employs a simple method of allocating resource blocks and scheduling VoIP user which is less affected by high load factor, improves the performance compared to the other two scheduling algorithms that employ packet deadline expiration procedure which is highly affected by high load factor. However, in the best-effort traffic (FTP) scenario, PF improved the performance of best-effort traffic compared to our proposed algorithm and EXP-PF. This is mainly because it applies fairness to all of its traffic regardless of QoS requirements.

We also analyzed the improvement made by our proposed algorithm compared to U-delay algorithm in [3] which was our benchmark algorithm as far as utility function is concerned. VoIP throughput metric was compared for both algorithms and our proposed algorithm provided better performance than that of U-delay in [3].

In future work, we will improve our proposed algorithm so that it can handle the best-effort traffic. We will also try to employ different tests such as real life scenarios in order to analyze the practicability of our results and to make them more reliable. Our proposed algorithm will also be extended to other real-time applications, i.e, video, as well as extending it to latest LTE advanced technology standards.
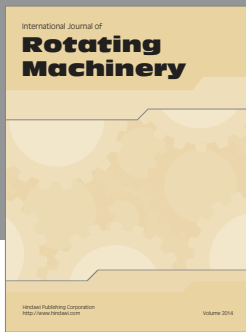
Conflict of Interests

The authors declare no financial or personal conflict of interests which may interfere with the study outcome.

References

[1] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G Evolution: HSPA and LTE for Mobile Broadband*, Elsevier, New York, NY, USA, 1st edition, 2007.

[2] S. Choi, K. Jun, Y. Shin, S. Kang, and B. Choi, "MAC scheduling scheme for VoIP traffic service in 3G LTE," in *Proceedings of the IEEE 66th Vehicular Technology Conference (VTC '07)*, pp. 1441–1445, October 2007.

[3] S. Ali and M. Zeeshan, "A utility based resource allocation scheme with delay scheduler for LTE service-class support," in *Proceeding of the IEEE Wireless Communications and Networking Conference (WCNC '12)*, pp. 1450–1455, Shanghai, China, April 2012.

[4] Rohde and Schwarz, "UMTS Long Term Evolution (LTE) technology introduction," Tech. Rep., Rohde and Schwarz, 2007.

[5] 3GPP, "Physical layer aspects for evolved UTRA," 3GPP Technical Report 25.814, version 7.1.0, 2006.

[6] J. Huang, V. G. Subramanian, R. Agrawal, and R. A. Berry, "Downlink scheduling and resource allocation for OFDM systems," *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 288–296, 2009.

[7] R. Musabe, H. Larijani, B. G. Stewart, and T. Boutaleb, "A new scheduling scheme for voice awareness in 3G LTE," in *Proceedings of the 6th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA '11)*, pp. 300–307, October 2011.

[8] R. Musabe and H. Larijani, "Evaluation of a new scheduling scheme for VoIP with mobility in 3G LTE," in *Proceedings of the 5th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ '12)*, pp. 8–13, 2012.

[9] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE cellular systems: an open-source framework," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 2, pp. 498–513, 2011.

[10] 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 *Proceedings of the IEEE 9th Malaysia International Conference on Communications with a Special Workshop on Digital TV Contents (MICC '09)*, pp. 815–820, December 2009.

[11] G. J. Choi and S. Bahk, "Cell-throughput analysis of the proportional fair scheduler in the single-cell environment," *IEEE Transaction Vehicular Technology*, vol. 56, pp. 766–778, 2007.

[12] R. Basukala, H. A. M. Ramli, and K. Sandrasegaran, "Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system," in *Proceedings of the 1st South Central Asian Himalayas Regional IEEE/IFIP International Conference on Internet (AH-ICI '09)*, November 2009.

[13] 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 Communications Surveys and Tutorials*, vol. 15, no. 2, pp. 678–700, 2013.

[14] S. S. Gokhale and S. K. Tripathi, "Routing metrics for best-effort traffic," in *Proceedings of the 11th International Conference on Computer Communications and Networks*, pp. 595–598, October 2002.

[15] B. Goode, "Voice over Internet Protocol (VoIP)," *Proceedings of the IEEE*, vol. 90, no. 9, pp. 1495–1517, 2002.

[16] K. Gonia, "Latency and QoS for Voice over IP," SANS Institute-As part of GIAC Practical Repository, 2004.

[17] G. Zhang and M. Hillenbrand, "Implementing SIP and H.323 signalling as Web Services," in *Proceedings of the 30th EUROMICRO Conference*, pp. 265–271, September 2004.

[18] S. Saha and R. Quazi, "Priority-coupling-a semi-persistent MAC scheduling scheme for VoIP traffic on 3G LTE," in *Proceedings of the 10th International Conference on Telecommunications (ConTEL '09)*, pp. 325–329, August 2009.

[19] Y. Kim, "An efficient scheduling scheme to enhance the capacity of VoIP services in evolved UTRA uplink," *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, Article ID 732418, 2008.

[20] IETF, "Attribute list extension for the service location protocol," IETF RFC 3059, 2001.

[21] T. Henttonen, K. Aschan, J. Puttonen et al., "Performance of VoIP with mobility in UTRA long term evolution," in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 2492–2496, Singapore, May 2008.

[22] ITU-T, "One way transmission time," ITU-T Recommendation G.114, 2003.

[23] TSG-RAN1, "LTE physical layer framework for performance verification," Tech. Rep. 3GPP TSG-RAN1 R1-070674, 2007.

[24] X. Qiu, "On the performance of adaptive modulation in cellular systems," *IEEE Transactions on Communications*, vol. 47, no. 6, pp. 884–895, 1999.

[25] H. Holma and A. Toskala, *WCD MA for UMTS: HSPA Evolution and LTE*, John Wiley & Sons, 4th edition, 2007.