Uplink Scheduler for SC-FDMA based Heterogeneous Traffic Networks with QoS Assurance and Guaranteed Resource Utilization

Rukhsana Ruby, *Student Member, IEEE,* Victor C.M. Leung, *Fellow, IEEE,*
and David G. Michelson, *Senior Member, IEEE*

***Abstract*—[1] Ensuring QoS provision for a network with heterogeneous traffic is considered as a difficult task because of the conflicting requirements of different traffic. Emergent high bandwidth 3G/4G technologies, such as LTE (Long Term Evolution), LTE-advanced are off the shelf in order to meet the demands of new applications by providing high data rate while simultaneously maintaining their individual QoS criteria. These systems recommend SC-FDMA (Single Carrier Frequency Division Multiple Access) as the uplink transmission scheme. In this paper, we have formulated the uplink scheduling problem for SC-FDMA based heterogeneous traffic systems considering all standard specific constraints and individual user QoS demand. In order to capture the QoS criteria of different traffic, we have adopted a utility function which is already used for the downlink operation of CDMA based systems. To facilitate the interest of the service providers, we consider an additional concept, opportunity cost function which is constructed based on the granular resource utilization. Dual decomposition method has been used in order to solve the formulated problem. Having noticed the high computational complexity of the optimal solution, we have given a sub optimal algorithm with relatively lower complexity. For evaluating the performance of our proposed uplink scheduling scheme, we assume, the network consists of best effort traffic, traffic with delay bound and traffic with throughput requirement. Finally, extensive simulation has been conducted to justify the efficacy and effectiveness of our scheme comparing with other existing solutions of an exemplary SC-FDMA based system, i.e., LTE system.**

***Index Terms*—Scheduling, Convex Optimization, QoS, Heterogeneous Traffic, SC-FDMA.**

I. Introduction

In order to meet the growing demand of high data rate, OFDMA (Orthogonal Frequency Division Multiple Access) based OFDM (Orthogonal Frequency Division Multiplexing) is the key technique of 3G/4G technologies such as LTE, LTE-advanced because of its immunity to inter-symbol interference and frequency selective fading. Despite numerous advantages of OFDM and OFDMA, their major disadvantage is their waveforms have high peak to average power ratio (PAPR). To reduce PAPR, LTE, LTE-advanced agree on using SC-FDMA technique for the uplink transmission which imposes contiguous subchannel allocation to a user. Typical scheduling of such systems involves the determination of a set of users, assignment of subchannels to these users and settling of transmit power for each subchannel. Uplink scheduling even for the conventional SC-FDMA based systems with best effort traffic is considered as challenging because of individual users' power constraint, discrete nature of subchannel assignment while maintaining contiguity pattern. With the invention of numerous applications with distinct fascinating features, now a days, the cellular systems are more likely to carry heterogeneous traffic with different QoS demands. Scheduling subchannels while satisfying diverse QoS of different new applications brings more challenges to the uplink scheduling problem on the top of its own inherent difficulties. Beside all these system and traffic specific issues, the service providers may want to have domination on the subchannel allocation determined by their in-house policy. This introduces further challenges to the uplink scheduling problem in SC-FDMA based systems conveying heterogeneous traffic.

Resource allocation problems for OFDM based networks especially downlink LTE systems have appeared in some survey papers recently [1], [2]. In homogeneous traffic networks, for two types of traffic, e.g., elastic traffic, traffic with QoS requirements, scheduling problems have extensively been studied. For the elastic traffic, in order to perform the scheduling decision, utility of the users is represented by a concave function [3]. Subject to the objective of maximizing the sum of general concave utility functions, one recent work [4] has proposed some computationally efficient algorithms. Channel aware throughput maximization technique for such systems is known as MT (Maximum Throughput). For fair resource allocation, typical PF (Proportional Fair) metric [5], [6], [7] is the ratio of instantaneous data rate on a subchannel and past average throughput. In [8], [9], PF scheme is formulated as an optimization problem with the objective of maximizing achieved throughput under the typical constraints of the system.

Among QoS aware schedulers in homogeneous traffic networks, [10] ensures guaranteed throughput by separating jobs in the time and frequency domains. In the time domain, they first separate traffic based on their current and pre-specified settled throughput. Then, they apply some PF scheme on the same priority users in order to obtain final subchannel assignment. More works on providing throughput guarantee for real time flows include [11], [12]. [13] calculates the priority indicator using HOL (Head of Line) packet delay and ensures target delay bound for the delay sensitive traffic. M-LWDF (Modified Largest Weighted Delay First) [14] has been

introduced for OFDMA based networks in [15] by putting past throughput in the denominator of the original metric. Two very promising strategies, i.e., LOG and EXP rules have been explained in [16]. Their subchannel selection metric is based on the logarithmic and exponential function of the packet HOL delay respectively. Unlike LOG, EXP scheduler takes the overall network status into account. Similar to these, there is another scheduling rule, i.e., MW (MaxWeight) [17] for delay sensitive traffic. Analytical results in terms of queue distribution for the schedulers with these rules are given in [18], [19], [20]. In [21], the authors adapt EXP rule by taking the characteristics of PF and exponential function of packet HOL delay into account. Two level packet scheduler working in the larger LTE frame and then in the granular frame level for real time traffic has been given in [22]. A similar approach, however working in discrete 3 levels has been presented in [23]. Third level of this scheduler is called cut-in process which discards packets whose delay deadline has been expired. Recent work [24] combines EXP rule, cooperative game and virtual token mechanism in order to ensure bounded delay and guaranteed bit rate for the users.

Scheduling problem in heterogeneous traffic networks is considered as more challenging compared to the homogeneous one because of the conflicting requirements of different traffic. There are three design issues need to be considered while formulating the scheduling problem of an ideal scheduler for heterogeneous traffic networks. Wireless users experience varying channel quality condition time to time due to the stochastic fading effects and hence, their achievable data rate are affected. The scheduler should choose the user with good channel quality condition in order to maximize the system throughput. However, if the user with better channel quality condition is always selected, the users with worse channel condition may starve. This issue is known as fairness and another very important factor to keep into consideration. Third, different applications have different QoS requirements and the utility function of the scheduler should be defined in such a way that it has the ability to capture those metrics simultaneously and effectively.

Similar to homogeneous traffic networks, downlink scheduling has extensively been studied for OFDM based networks carrying heterogeneous traffic. These works mainly adopted three categories of designs. First, some works gave strict priority to the high priority traffic compared to the low priority ones, such as [25], [26]. [25] has proposed a solution for VoIP and data traffic while giving strict priority to VoIP users. A scheduling policy with strict priority across classes is also studied in [26]. Within a class, proposed scheduler does chunk by chunk resource allocation. The work in [27] has given strict priority to SIP traffic over other data traffic. Second, the authors in [28] have addressed mixed traffic (delay sensitive, guaranteed throughput, best effort) by designing the scheduler on the time and frequency domains. The priority sets are populated based on the QCI (QoS Class Indicator) of each data flow and classified as GBR (Guaranteed Bit Rate) and non-GBR. Third, the scheduler is designed by representing the satisfaction of users by the utility function. In [29], [30], different utility functions are used to represent different types of traffic (traffic with data rate requirement, traffic with delay constraint). Although [29] did not consider the channel condition of users in the scheduling decision, [30] takes all required issues into account. Beside these, there are some general mechanisms for handling mixed traffic. For example, the authors in [16] have proposed a low complexity RB (Resource Block) allocation algorithm using LOG/EXP, EDF (Earliest Deadline First) rules and tested their scheme with mixed streaming and live video traffic. The work in [31] studied a scheduling policy that gives equal priority to all packets with different QoS unless their delay reaches close to their deadline.

Unlike downlink operation, not much works have been conducted for the uplink access of such systems. Most works focused on the throughput maximization objective. For example, [32]–[35] have given different heuristics based on different principles upon stating that the problem is NP-hard. Although these works take subchannel contiguity constraint into account, skip individual users' power constraint. However, this constraint is very important factor in scheduling decision and limits overall system performance. On the other hand, there are very few works for the uplink scheduling with QoS assurance. These works are mainly for the homogeneous delay sensitive traffic [36], [37]. One recent work on energy efficient QoS assured scheduler is [38], where QoS is considered as user's instantaneous rate. For heterogeneous traffic environments, recently, [39] has proposed a scheduler which is based on the time and frequency domain scheduler [28]. One drawback of this work is they did not consider traffic class in their design. Moreover, they evaluated the performance of this scheduler in their customized simulation scenario which is not practical.

Having noticed the drawback of [39] and in order to design a utility based scheduler for the heterogeneous traffic environment, in this paper, we have proposed an uplink scheduling scheme for a SC-FDMA based system carrying heterogeneous traffic. This work is based on the utility function [40] has used. Difference is, their solution is suitable for the downlink scheduling and designed for CDMA based systems. The utility function of [40] is general, one single function can handle diverse types of traffic, instantaneous channel condition and so forth. Typically, this utility function is used in economics in order to maintain the social welfare (another name of fairness) of the society, however rarely applied in communications. Moreover, this utility function can distinguish inter or intra class based traffic prioritization. Beside capturing three key design issues of heterogeneous traffic environment by this utility function, in order to assist the service providers, proposed uplink scheduling problem consists of a constraint which imposes some control on subchannel allocation. The distinctive novelty and contributions of this paper are briefly summarized as follows.

- For representing the satisfaction of a user, we have adopted the utility function already used for the downlink scheduling operation of a CDMA based system [40]. Although we have adopted their utility function and same definition of QoS measures for different traffic, we have solved a distinct problem for a different system, i.e., uplink scheduling problem of SC-FDMA based systems.

Furthermore, the way they have solved the downlink scheduling problem is different compared to us. Given pre-defined physical layer data rate of the users and given certain system capacity, their work schedules a set of users instead of CDMA tones. Whereas, our scheduling scheme determines every single subchannel-user mapping, their assigned power which is the true notion of an ideal scheduler given the incoming packets in each scheduling epoch.

- For obeying every detailed specifications of the standard [41], our problem formulation takes individual user's power and subchannel contiguity constraints into account. Individual user's power constraint is required in order to correctly measure overall system performance. On the other hand, in order to avoid high PAPR of the generated waveform, we assign contiguous subchannels to each user.
- Beside standard specific constraints, our formulation consists of another factor, opportunity cost function. Physical meaning of this function is how much utilization of resource the service provider can sacrifice in order to achieve other system or user specific benefits. Cost function can work in both granular and aggregate resource utilization level. For this work, we assume it to be dependent on subchannel utilization as later one can be achieved through the former one, however not vice versa. Furthermore, the service providers may relate opportunity cost function to revenue as desired.
- By setting certain parameter for the opportunity cost function, our scheduling scheme can be transformed to two extreme ends of scheduling mechanisms, i.e., MT and PF schedulers. By providing enough evidence, we have proved this statement analytically.
- Dual decomposition method has been used in order to show the optimal solution structure of our formulated problem. Finally, from the guiding principles of the optimal solution, we have given a low complexity scheduling algorithm.
- Extensive simulation has been conducted assuming the network has best effort traffic, traffic with throughput requirement and traffic with delay bound. While comparing with other scheduling solutions of a sample 3G/4G system, such as LTE system, we have proved the effectiveness and efficacy of our scheme.

Rest of the paper is organized as follows, Section II gives the overview and components of the system while elaborating detailed formulation of the problem. Solution approach and the resultant algorithm are presented in Section III. We investigate some interesting characteristics of our scheduling scheme in Section IV. In order to show the effectiveness of the proposed scheme, we provide simulation results followed by the simulation methodology in Section V. Finally, Section VI concludes the paper with some directions of future research.

II. System Model & Problem Formulation

We consider uplink scheduling problem of a typical SC-FDMA based cellular network. Transmission bandwidth of the system is orthogonally divided into a number of subchannels. We denote set $\mathbf{N}$ contains subchannels available at each scheduling epoch. The number of users in the system is M. Users are again categorized into C classes, where class c has higher priority than class $c+1$. Let M_c denotes the number of class c users; $M = \sum_{c}^{C} M_c$ and the set holding all users is $\mathbf{M}$. Each class is accompanied with diverse kinds of QoS criterion. At each scheduling instant, all users transmit their channel state information to the base station. In the channel coherent time, based on the traffic type, channel state and their QoS requirements, the base station assigns appropriate subchannels to the corresponding users as well as determines power of each subchannel. Purpose of the scheduler equipped in the centralized controller is to maximize the sum satisfaction of all users.

A. Problem Formulation

Since SC-FDMA based resource allocation allows the service providers to allocate resource in granular subchannel level, for a particular user i of a certain class c, we define the utility function for each and every subchannel $j, \forall j \in \mathbf{N}$. At scheduling time instant t, the satisfaction of user i of class c on subchannel j is perceived by the utility function $U_{cij}\left(\{X^z_{cij}(t)\}_{z=1}^{m_{ci}}\right)$, where $\{X^z_{cij}(t)\}_{z=1}^{m_{ci}} = \{X^1_{cij}(t),\ X^2_{cij}(t),\ \cdots\ X^{m_{ci}}_{cij}(t)\}$. $\{X^1_{cij}(t),\ \cdots X^{m_{ci}}_{cij}(t)\}$ are computed quantitative QoS measures in terms of (c,i)th user's satisfactions in the uplink system during the scheduling decision of subchannel j at time instant t. Typically, QoS measures are the average throughput, current data rate, average delay etc. m_{ci} represents the maximum number of QoS measures for user (c,i). Therefore, we can write the objective function as

$$\max \sum_{c}^{C}\sum_{i}^{M_c}\sum_{j}^{N} U_{cij}(\{X^z_{cij}\}_{z=1}^{m_{ci}}(t)). \tag{1}$$

If x_{cij} denotes the fraction of subchannel j allocated to user (c,i), the total allocation across all users should be no larger than 1, i.e.,

$$\sum_{(c,i)\in\mathbf{M}} x_{cij}(t) \le 1,\ \forall j \in \mathbf{N}. \tag{2}$$

According to the standard, there is a constraint on x_{cij} to be an integer $\{0,1\}$. We will see later, how we have handled this constraint while solving this problem. Beside this, due to the contiguity constraint of SC-FDMA technique, subchannels are allocated to user (c,i) in contiguous manner. It implies

$$\begin{aligned} &\text{if } x_{cin}(t) = 1 \ \&\& \ x_{ci(n+1)}(t) = 0, x_{cij}(t) = 0,\ n+2 \le j \le N \\ &\text{if } x_{cin}(t) = 1 \ \&\& \ x_{ci(n-1)}(t) = 0, x_{cij}(t) = 0,\ 1 \le j \le n-2. \end{aligned} \tag{3}$$

Each user (c,i) has power limitation during a scheduling epoch and its maximum power is denoted by P_{ci}. For transmitting on subchannel j, we denote (c,i)th user's transmission

power is p_{cij} and total power used on all allocated subchannels cannot exceed the maximum power.[2]

$$\sum_{j\in\mathbf{N}} p_{cij}(t) \leq P_{ci},\ \forall (c,i)\in\mathbf{M}. \quad (4)$$

Moreover, there are upper and lower bounds of each QoS measure and these are pre-defined values set by the users. For example, zth QoS measure of user (c,i) has an upper bound $\nu_{ci}^{z,max}$ and a lower bound $\nu_{ci}^{z,min}$ (e.g., maximum and minimum average throughput).

$$\nu_{ci}^{z,min} \leq\ X_{cij}^{z}(t) \leq\ \nu_{ci}^{z,max},\ \ \forall (c,i)\in\mathbf{M},\ \forall z, 1\leq z\leq m_{ci}. \quad (5)$$

Constraints presented in Equations 2, 3, 4 and 9 are either enforced by the standard or set by the users. Another constraint we would like to introduce is for the sake of service providers, i.e., opportunity cost. The concept of opportunity cost can be used to manage the tradeoff between fairness and resource utilization. In order to ensure fairness across the network, the scheduler may be forced to serve low rate generating users resulting in rate loss. To limit this rate loss, we have proposed a cost function called opportunity cost. Similar to other metrics, opportunity cost is also defined in granular user and subchannel level. At scheduling instant t, while allocating subchannel j, we define opportunity cost for user (c,i) is $OC_{cij}(t)$. $OC_{cij}(t)$ is a function of rate for user (c,i) on subchannel j, $R_{cij}(t)$. $R_{cij}(t)$ is given by

$$R_{cij}(t) = x_{cij}(t) log\left(1+\frac{p_{cij}(t)e_{cij}(t)}{x_{cij}(t)}\right), \quad (6)$$

where e_{cij} is the normalized received SNR per unit transmit power of user (c,i) on subchannel j from the base station. In the multi-cell scenario, e_{cij} indicates the SINR instead of SNR and it is a function of power used by the users of neighboring cells on subchannel j. At the beginning of scheduling instant t, the scheduler gathers complete information of this metric for all users and all subchannels. Maximum rate that the scheduler can earn out of subchannel j at time t is given by

$$R_j^{max}(t) = \max_{(c,i)\in\mathbf{M}} R_{cij}(t),\ \forall j\in\mathbf{N}. \quad (7)$$

Using these metrics, $OC_{cij}(t)$ is defined as follow

$$OC_{cij}(t) = R_j^{max}(t) - R_{cij}(t). \quad (8)$$

In other way, the opportunity cost is a measure of how much rate the network operator would forgo if user (c,i) is selected for the transmission on subchannel j at scheduling instant t while there is some other user $(c,i)^*$ that generates the highest rate on this subchannel. Taking the network operators' interest into account, the objective function in Equation 1 can further be constrained by the opportunity cost function

$$OC_{cij}(t) \leq H,\ \forall (c,i)\in\mathbf{M}, \forall j\in\mathbf{N}. \quad (9)$$

The network operator can determine the required level of fairness and resource utilization by choosing appropriate value of H. If $H=\varepsilon R_j^{max}$, rate loss is no more than ε% of maximum achievable rate R_j^{max} on subchannel j. Moreover, $H=0$ implies that the network operator cannot tolerate any rate loss, therefore it always picks the highest rate generating user while allocating any subchannel. In order to ignore the opportunity cost, the network operator can set $H=R_j^{max}$. In this case, all users are treated equally by the scheduler.

[2]According to Figure 3 of [41], different levels of power for the allocated RBs (equivalent to subchannels) of a user is allowed.

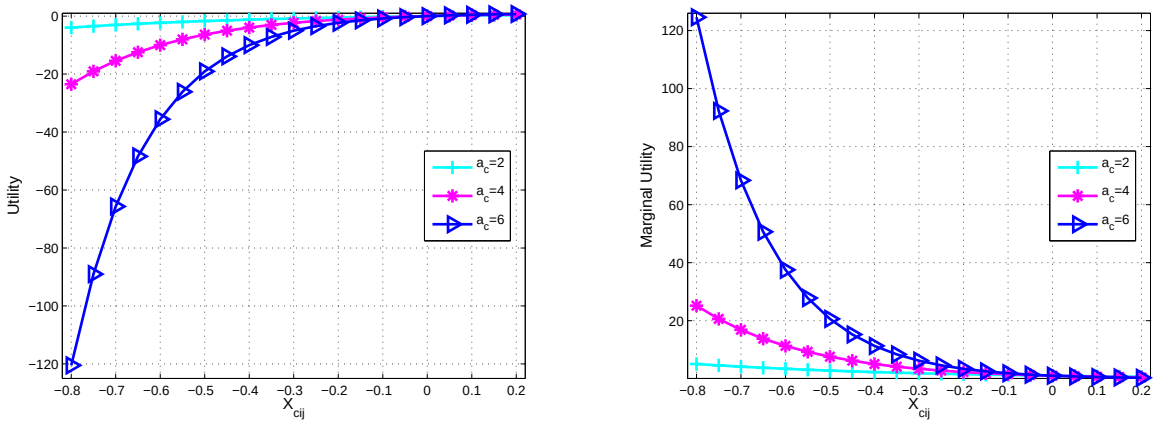
B. The Utility Function

In this section, we introduce a utility function which is able to meet all requirements of an ideal utility function. In order to satisfy all requirements of the problem described in the previous subsection, the feasible utility function should have concavity property. The more the allocated resource, the more satisfied the user is, i.e., utility function should be non-decreasing function of $X_{cij}^{z}, \forall j\in\mathbf{N}, \forall z\in[1,m_{ci}]$. While allocating a subchannel at time instant t, if the values of all QoS measures for a user (c,i) are minimum, the utility value for that user attains its unique minimum value U_{ci}^{min}, whereas when the values reach to their maximum, the utility appears to its unique maximum value U_{ci}^{max}. In rest of the other cases, the utility value remains in between these two quantities, or in other way, $U_{ci}^{min}\leq U_{cij}(\{X_{cij}(t)\}_{z=1}^{m_{ci}})\leq U_{ci}^{max}$. Furthermore, once the utility value for a user reaches to its maximum value, additional allocated resource cannot deviate it from its maximum quantity U_{ci}^{max}.

In addition of having above fundamental properties, the utility function should support inter-class prioritization. We denote a parameter $a_c, \forall c, 1\leq c\leq C$ to distinguish inter-class prioritization. Larger the value of a_c, the higher the priority of class c. Proposed utility function is already used for the scheduling purpose in CDMA based networks. As our problem is different, we interpret this function in a different way. The utility function for user (c,i) while allocating subchannel j at time instant t is given by

$$U_{cij}\left(\{X_{cij}^{z}\}_{z=1}^{m_{ci}}(t)\right) \ =\ 1\ -\ e^{-a_c\sum_{z=1}^{m_{ci}} X_{cij}^{z}(t)}. \quad (10)$$

We plot the utility function with respect to arbitrary QoS measure X_{cij} for 3 users with different a_c in Figure 1(a). We observe, utility function has the diminishing property. When the quantitative value of X_{cij} is low, rate of change of the utility (slope) is larger, it implies, if the scheduler gives priority to the user with low QoS measure, contribution of this user towards the maximization of overall utility is higher compared to others. Hence, the scheduler needs to take this into account while making the scheduling decision of subchannel j. In order to show the decreasing trend of slope with the increasing X_{cij}, we plot Figure 1(b). We define the slope of the utility function as marginal utility. Moreover, larger value of a_c makes the utility function steeper with respect to QoS measures. In other way, slope of the utility function is steeper with higher a_c at low X_{cij}. Therefore, for this particular utility function, the slope of a user's utility plays an important role in the scheduling decision of subchannels. From the perspective of economics, this type of utility function has another definition,



(a) Utility comparison for different a_c and X_{cij}.

(b) Marginal utility comparison for different a_c and X_{cij}.

Fig. 1. Characteristics of the utility function.

if the users with low QoS measure are given higher priority gaining additional resource, this maximizes the social welfare as well as fairness of the system.

In order to fully explain the utility function, we need to specify the QoS measures, i.e., $\{X_{cij}^z\}_{z=1}^{m_{ci}}$. QoS measures further require to define the traffic types of which they belong to. We consider, the network has the similar types of traffic as [40] mentioned in their work. Since traffic types are similar, we adopt similar form of QoS definitions. However, our QoS measure is the quantity when we schedule subchannels instead of users. Hence, we reiterate the QoS metrics of each traffic type with their exact interpretation in the following.

1) **Traffic with minimum throughput requirement:** In order to ensure minimum throughput, we need to design QoS measure so that the scheduler gives higher priority to the user with lower throughput. While allocating subchannel j, let denote the average throughput achieved by user (c,i) up to time t is $\zeta_{cij}(t)$. Maximum incoming data rate for this user is ζ_{ci}^{max}, whereas the minimum required one is ζ_{ci}^{min}. As we discussed before, the property of our utility function is such that the scheduler gives priority to the user with lower QoS measure. Therefore, QoS metric of user (c,i) for this type of traffic at time t while allocating subchannel j has been defined as $X_{cij}^1(t) = \left(\mu_{ci}^1 - \frac{\zeta_{ci}^{min}}{\zeta_{cij}(t)}\right)$, where $0 <= \mu_{ci}^1 <= 1$. Smaller value of μ_{ci}^1 gives more weight to this metric. This metric is also termed as fairness measure for this type of traffic. If any user obtains lower throughput while the scheduler tends to allocate one subchannel, because of the lower value of this metric, the scheduler is forced to give provision to this user and thus, total utility of the system is maximized and welfare of the system is maintained.
2) **Traffic with bounded delay constraint:** There are some traffics types (e.g., audio, VoIP) which have some certain delay constraint. In order to design metric for this class of traffic, let denote packet HOL delay of user (c,i) is $D_{cij}(t)$ at time instant t while making the scheduling decision of subchannel j. Maximum delay bound that the user (c,i) can tolerate is D_{ci}^{max}. Similar to the case of traffic with throughput requirement, this metric represents the fairness. We define this metric as $X_{cij}^1(t) = \left(\mu_{ci}^1 - \frac{D_{cij}(t)}{D_{ci}^{max}}\right), 0 <= \mu_{ci}^1 <= 1$. Lower value of μ_{ci}^1 gives more weight to this metric. If any user (c,i) of this traffic type starts to experience higher packet delay, $X_{cij}^1(t)$ appears to get lower value while allocating subchannel j and hence, the scheduler gives more priority to that user.
3) **Best effort traffic:** Best effort traffic has usually the lower priority compared to other traffic types described earlier. If any user receives considerably lower throughput comparing with other users in the system, in order to ensure fair distribution of resource, we need to design a metric for this type of traffic. Denote the measure $X_{cij}^1(t) = \left(\frac{\bar{\zeta}_{cij}(t)}{\max\limits_{(c,i)} \zeta_{cij}(t)} - \mu_{ci}^1\right), 0 <= \mu_{ci}^1 <= 1$. Starvation of users with lower average throughput results in unfairness, thus by serving those users at some point the scheduler maintains social welfare of the system. If the scheduler would serve user with higher average throughput, it might further increase that individual user's throughput, however that contribution is lower towards the system compared to the case when the scheduler would serve user with lower average throughput. The role of μ_{ci}^1 is the weight for this measure and larger value gives additional weight to this metric.
4) **Traffic with minimum throughput and bounded delay requirements:** If the traffic has both minimum throughput requirement and bounded delay constraint, QoS measure for this type of traffic can be defined as $X_{cij}^1(t) = \left(\mu_{ci}^1 - w_t \frac{\zeta_{cij}(t)}{\zeta_{ci}^{min}} - w_d \frac{D_{cij}(t)}{D_{ci}^{max}}\right)$. Similar to other types of traffic discussed above, lower value of μ_{ci}^1 gives higher priority to this QoS factor. Whether we want to give priority to delay bound or minimum throughput of this traffic type is determined by the values of w_d and w_t. w_d and w_t are normalized by 1, i.e., $w_d + w_t = 1$. If we

want to ensure equal priority to both delay and throughput of this type of traffic, we can set $w_d = w_t = 0.5$.

No matter the traffic type, we need another QoS metric which gives priority to the user with better channel condition. We define this metric as $X^2_{cij}(t) = \left(\mu^2_{ci} - \frac{R_{cij}(t)}{(\max_{(c,i)} R_{cij}(t))}\right), 0 <= \mu^2_{ci} <= 1$. We normalize the user's instantaneous rate on subchannel j by the maximum possible rate achieved using this subchannel. Normalized rate has been subtracted from μ^2_{ci}, because we want to make sure that the user with better instantaneous rate obtains lower quantity compared to the one with worse channel condition. μ^2_{ci} works as a penalty of not serving users with good channel condition. Larger value of μ^2_{ci} gives more weight to this QoS measure. Users with better channel condition will have lower quantitative value for metric $X^2_{cij}(t)$ and hence, according to the property of the utility function, the scheduler should give more provision to those users while allocating subchannel j at time t.

So, we have concluded that we have two QoS measures: one for ensuring fairness of specific traffic type, $X^1_{cij}(t)$ and another one is common to all users $X^2_{cij}(t)$ for ensuring the provision when users have better channel condition.

III. Solution Approach & Scheduling Algorithm

In the previous section, we have seen that marginal utility of a user is the performance metric to maximize in order to maintain social welfare of the system. Hence, at each scheduling epoch, we want to maximize the sum marginal utility of all users across all subchannels, i.e., we want to maximize $\sum_{(c,i)\in\mathbf{M}}\sum_{j\in\mathbf{N}} a_c exp\{-a_c\left[X^1_{cij}+X^2_{cij}\right]\}$. It implies, the objective function is to maximize $\sum_{(c,i)\in\mathbf{M}}\sum_{j\in\mathbf{N}} -a_c\left[X^1_{cij}+X^2_{cij}\right]$. We denote R^{max}_j by Γ^{max}_j. Since the objective function and constraint in Equation 8 have Γ^{max}_j, it is required to add an additional constraint in the problem formulation in order to support this assignment. Taking all constraints, resultant formulation yields

$$\max_{(\mathbf{x},\mathbf{P})} \sum_{(c,i)\in\mathbf{M}}\sum_{j\in\mathbf{N}} a_c\left(\frac{R_{cij}}{\Gamma^{max}_j} - \mu^2_{ci} - X^1_{cij}\right) \quad (11)$$
$$\sum_{(c,i)\in\mathbf{M}} x_{cij} \le 1,\ \forall j\in\mathbf{N},\ \sum_{j\in\mathbf{N}} p_{cij} \le P_{ci},\ \forall (c,i)\in\mathbf{M}$$
$$\text{if } x_{cin}=1 \ \&\&\ x_{ci(n+1)}=0, x_{cij}=0, n+2\le j\le N$$
$$\text{if } x_{cin}=1 \ \&\&\ x_{ci(n-1)}=0, x_{cij}=0, 1\le j\le n-2$$
$$\Gamma^{max}_j - R_{cij} \le \varepsilon\Gamma^{max}_j,\ R_{cij}\le\Gamma^{max}_j.$$

The problem in Equation 11 has no duality gap and we can solve it by formulating it as a dual problem with associated dual variables $\boldsymbol{\alpha} = (\alpha_{ci})_{(c,i)\in\mathbf{M}}$ for constraint 2, $\boldsymbol{\beta} = (\beta_j)_{j\in\mathbf{N}}$ for constraint 3 and $\boldsymbol{\gamma} = (\gamma_{cij})_{(c,i)\in\mathbf{M},j\in\mathbf{N}}$ for constraint 9 and $\boldsymbol{\delta} = (\delta_{cij})_{(c,i)\in\mathbf{M},j\in\mathbf{N}}$ for the last supportive constraint. Resultant lagrangian looks like

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\mathbf{x},\mathbf{P}) := \sum_{(c,i)\in\mathbf{M},j\in\mathbf{N}} a_c\left(\frac{R_{cij}}{\Gamma^{max}_j} - \mu^2_{ci} - X^1_{cij}\right) \quad (12)$$
$$+\sum_{(c,i)}\alpha_{ci}\left(P_{ci}-\sum_j p_{cij}\right)+\sum_j\beta_j\left(1-\sum_{(c,i)}x_{cij}\right)$$
$$+\sum_{(c,i)}\sum_j\gamma_{cij}\left(\varepsilon\Gamma^{max}_j-\Gamma^{max}_j+R_{cij}\right)+\sum_{(c,i)}\sum_j\delta_{cij}\left(\Gamma^{max}_j-R_{cij}\right).$$

From duality theory, optimal solution to problem in Equation 12 is given by

$$\min_{(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta})>=0}\ \max_{(\mathbf{x},\mathbf{P})} L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\mathbf{x},\mathbf{P}). \quad (13)$$

In order to solve this problem, first we find the optimal value of $\mathbf{x}$ and $\mathbf{P}$ given fixed value of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. Once we obtain $\mathbf{P}$, we rearrange the lagrangian in such a way that it becomes the function of mutually exclusive per user cost function, denoted by β_{cij}. Optimal value of β_j is the maximum possible value of β_{cij} over all users for subchannel j taking the subchannel contiguity constraint 3 into account. Because of the subchannel contiguity constraint, multi user diversity of OFDMA systems may not be achievable. Finally, optimal value of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ is obtained by the help of subgradient based numerical search method. Taking the derivative of Equation 12 w.r.t. p_{cij} and following the K.K.T condition, we obtain optimal p_{cij}, i.e.,[3]

$$p^*_{cij} = x_{cij}\left[\frac{a_c(1+\gamma_{cij}-\delta_{cij})}{\alpha_{ci}} - \frac{1}{e_{cij}}\right].$$

Substituting $\mathbf{p}^*$ into Equation 12, we obtain

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\mathbf{x}) := \sum_{(c,i)}\sum_j x_{cij}\left(f(\alpha_{ci},\gamma_{cij},\delta_{cij},e_{cij})-\beta_j\right) \quad (14)$$
$$+\sum_{(c,i)}\alpha_{ci}P_{ci}+\sum_j\beta_j+\sum_{(c,i)}\sum_j\gamma_{cij}\Gamma^{max}_j(\varepsilon-1)+\sum_{(c,i)}\sum_j\delta_{cij}\Gamma^{max}_j,$$

where $f(\alpha_{ci},\gamma_{cij},\delta_{cij},e_{cij}) = a_c\left(\frac{h_{cij}}{\Gamma^{max}_j}-\mu^2_{ci}-X^1_{cij}\right) + \gamma_{cij}h_{cij} - \delta_{cij}h_{cij} + \frac{1}{e_{cij}} - \frac{a_c(1+\gamma_{cij}-\delta_{cij})}{\alpha_{ci}}$, $h_{cij} = log\left(\frac{a_c e_{cij}(1+\gamma_{cij}-\delta_{cij})}{\alpha_{ci}}\right)$, $\Gamma^{max}_j = \max_{(c,i)} h_{cij}$. We denote (c,i)th user's cost function on subchannel j $f(\alpha_{ci},\gamma_{cij},\delta_{cij},e_{cij})$ by β_{cij}. Given that $x_{cij}\in[0,1]$, optimal value of $\mathbf{x}$ is obtained by following the procedure below. For the sake of procedure, we copy elements of set $\mathbf{N}$ in another set $\mathbf{N}'$.

1) First, for each subchannel $j\in\mathbf{N}'$, find the best subchannel metric among all users and denote it by $\tilde{\beta}_j = \max_{(c,i)\in\mathbf{M}}\beta_{cij}$. Second, find a subchannel permutation $\{\nu_j\}_{j\in\mathbf{N}}$ such that $\tilde{\beta_{\nu_1}} >= \tilde{\beta_{\nu_2}} >= \cdots >= \tilde{\beta_{\nu_N}}$. Select the subchannel with index ν_1 and its designated user $(c,i)|_{\nu_1}$ (to which it obtains maximum value of cost function), check whether the selected subchannel and its designated

[3]In the multi-cell scenario, if the number of users in all cells is M' and the number of subchannels is N, we will require to solve $M'N$ number of linear equations in order to obtain optimal p_{cij}. This is because, when we take the derivative of Equation 12 w.r.t. p_{cij}, it becomes the linear function of other users' power in the neighboring cells including p_{cij}.

user meets the subchannel contiguity constraint. It means, if the designated user has already some subchannels allocated, selected subchannel should be the contiguous to its allocated subchannel block, otherwise selected subchannel should be belonged to the designated user without any hesitation. If the subchannel with index ν_1 fails to satisfy subchannel contiguity constraint, then the subchannel with index ν_2 is chosen and this operation continues until the selected subchannel satisfies the subchannel contiguity constraint. We denote finally selected subchannel as ν^* and its designated user is $(c,i)|_{\nu^*}$. There might be ties in this assignment which can be resolved by some inefficient search. Therefore, optimal $\mathbf{x}$ for subchannel ν^* is obtained as follows.

$$x^*_{cij} = \begin{cases} 1 & \text{if } j = \nu^*, (c,i) = (c,i)|_{\nu^*} \\ 0 & \text{if } j = \nu^*, \{(c,i) \in \mathbf{M} | (c,i) \neq (c,i)|_{\nu^*}\} \end{cases}.$$

2) Since the lagrangian is a sum of users' cost function β_{cij}, we can minimize $L(.)$ over $\boldsymbol{\beta}$ for the given values of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ by setting $\beta^*_{\nu^*} = \tilde{\beta}_{\nu^*}$. More than one user can obtain the value $\beta^*_{\nu^*}$, however ties can be broken arbitrarily without losing the optimality.
3) Remove subchannel ν^* from set $\mathbf{N}'$.

Substituting optimal $\mathbf{x}$ and $\boldsymbol{\beta}$ into the lagrangian, resultant lagrangian yields

$$L(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\delta}) = \sum_{(c,i)}\sum_{j}\left[\beta_{cij} - \beta^*_j\right]^+ + \sum_{(c,i)} \alpha_{ci} P_{ci} + \sum_j \beta^*_j + \sum_{(c,i)}\sum_j \gamma_{cij}\Gamma^{max}_j(\varepsilon - 1) + \sum_{(c,i)}\sum_j \delta_{cij}\Gamma^{max}_j. \quad (15)$$

Notice that, the lagrangian in Equations 15 is the function of $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. Now, the optimal $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ can be obtained by minimizing $L(.)$ and this is the optimal solution of Equation 15. We have adopted the subgradient based search approach in order to obtain optimal $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$. Subgradient search requires the following updates in each iteration

$$\alpha_{ci}(t+1) = \alpha_{ci}(t) - \kappa(t)\left(P_{ci} - \sum_j p^*_{cij}(t)\right) \quad (16)$$

$$\gamma_{cij}(t+1) = \gamma_{cij}(t) - \kappa(t)\left(\varepsilon\Gamma^{max}_j(t) - \Gamma^{max}_j(t) + h_{cij}(t)\right) \quad (17)$$

$$\delta_{cij}(t+1) = \delta_{cij}(t) - \kappa(t)\left(\Gamma^{max}_j(t) - h_{cij}(t)\right). \quad (18)$$

Step size $\kappa(t)$ in iteration $t+1$ is given by

$$\kappa(t) = \frac{\tilde{L} - L\left(\boldsymbol{\alpha}(t),\boldsymbol{\gamma}(t),\boldsymbol{\delta}(t)\right)}{\left|\frac{dL(.)}{d\boldsymbol{\alpha}(t)}\right|^2 + \left|\frac{dL(.)}{d\boldsymbol{\gamma}(t)}\right|^2 + \left|\frac{dL(.)}{d\boldsymbol{\delta}(t)}\right|^2}, \quad (19)$$

where $\left|\frac{dL(.)}{d\boldsymbol{\alpha}(t)}\right| = \sqrt{\sum_{(c,i)\in\mathbf{M}}\left(P_{ci} - \sum_j p^*_{cij}(t)\right)^2}$, $\left|\frac{dL(.)}{d\boldsymbol{\gamma}(t)}\right| = \sqrt{\sum_{(c,i)\in\mathbf{M}}\sum_{j\in\mathbf{N}}\left(\varepsilon\Gamma^{max}_j(t) - \Gamma^{max}_j(t) + h_{cij}(t)\right)^2}$ and $\left|\frac{dL(.)}{d\boldsymbol{\delta}(t)}\right| = \sqrt{\sum_{(c,i)\in\mathbf{M}}\sum_{j\in\mathbf{N}}\left(\Gamma^{max}_j(t) - h_{cij}(t)\right)^2}$. $\tilde{L}$ is the estimate of the lagrangian determined from the previous iterations. Given that $P = \max_{(c,i)\in\mathbf{M}} P_{ci}$ and $e^{max} = \max_{(c,i)\in\mathbf{M}}\max_{j\in\mathbf{N}} e_{cij}$, for this problem, each element of $\frac{dL(.)}{d\boldsymbol{\alpha}}$ is bounded within the range $[0,P]$ and that of vectors $\frac{dL(.)}{d\boldsymbol{\gamma}}$, $\frac{dL(.)}{d\boldsymbol{\delta}}$ is within $[0,(1+loge^{max}P)]$. Under this statement, optimal value of the lagrangian is achievable within the finite number of iterations. Detailed procedure of achieving convergence is given in Exercise 6.3.2 of [42]. This problem has $M+2MN$ number of dual variables and it requires several thousands of iterations to achieve convergence. Moreover, in each iteration, there is an issue of resolving ties which requires some inefficient search. Given these disadvantages, this procedure as a scheduler may not be efficient to implement in fast time scale. Hence, we have proposed a sub optimal algorithm presented in *Algorithm 1*.[4]

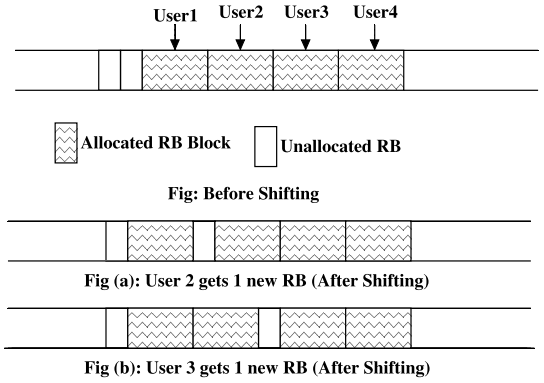


Fig. 2. Sample subframe structure for stage 2 of Algorithm 1.

In *Algorithm 1*, $g_{ci}(j)$ is given by

$$g_{ci}(j) = \sum_{n\in\Omega'_{ci}(j-1)} log(1+p^*_{cin}e_{cin}) - \sum_{n\in\Omega_{ci}(j-1)} log(1+p^*_{cin}e_{cin}), \quad (20)$$

where $\Omega'_{ci}(j) = \Omega_{ci}(j-1)\cup l_{ci}(j)$. $l_{ci}(j)$ is the best unallocated subchannel for user (c,i). p^*_{cij} is the power on subchannel j after doing the power control on the set $\Omega_{ci}(j)$ or $\Omega'_{ci}(j)$ for user (c,i). Power control of user (c,i) on the set $\Omega_{ci}(j)$ is equivalent to solving the following optimization problem

$$\underset{\sum_{n\in\Omega_{ci}(j)} p_{cin} = P_{ci}}{\arg\max} \sum_{n\in\Omega_{ci}(j)} log\left(1+p_{cin}e_{cin}\right). \quad (21)$$

Problem in Equation 21 can be solved by dual formulation which we have given in Section 5.4 of our recent paper [44]. Complexity of this operation is found to be $|\Omega_{ci}(j)|$. Furthermore, in the 1st stage of the algorithm, $\Gamma_{ci}(j) = g_{ci}(j)$. In the 2nd stage, we define $U_s(j)$ as $\sum_{s\in L_{sj}}\sum_{(c,i)\in L_j} -a_c\left(X^1_{ci} + X^2_{ci}\right)$ and the parameters inside the sum term are determined presuming the network status at the beginning of current scheduling epoch.

[4]In the optimal solution of multi-cell scenario, the power allocated to the subchannels of a user can be less than its maximum allowable power. Hence, it is not trivial to design a sub optimal algorithm for the multi-cell scenario. However, [43] is a useful reference to design an algorithm for our problem.

Algorithm 1 Sub Optimal Subchannel Allocation Algorithm

1: ε is tolerable rate loss percentage and set by the service provider.
2: Set $j := 0$, $\Omega_{ci}(j) := \emptyset$ for each user (c,i).
3: **while** $j < N$ **do**
4: Set $j := j+1$.
5: Get the best subchannel index $l_{ci}(j)$ for each user (c,i).
6: If user (c,i) had already subchannels allocated, $l_{ci}(j)$ is selected from at most 2 contiguous subchannels, otherwise $l_{ci}(j)$ is arbitrarily chosen.
7: Calculate metric $g_{ci}(j)$ and $\Gamma_{ci}(j)$ for each user (c,i).
8: Calculate $\Gamma^{max}(j) := \max_{(c,i)} \Gamma_{ci}(j)$.
9: Calculate $H := \varepsilon\Gamma^{max}(j)$.
10: Determine users $(c,i) \in L_j$ so that $(\Gamma^{max}(j) - \Gamma_{ci}(j)) <= H$.
11: Find $(c,i)^* := \arg\max_{(c,i)\in L_j} a_c \left(\frac{g_{ci}(j)}{\Gamma^{max}(j)} - \mu^2_{ci} - X^1_{cij} \right)$.
12: Assign the jth subchannel to user $(c,i)^*$

$$\Omega_{ci}(j) := \begin{cases} \Omega_{ci}(j-1) \cup l_{ci}(j) & \text{if } (c,i) = (c,i)^* \\ \Omega_{ci}(j-1) & \text{Otherwise} \end{cases}.$$

13: **end while**
14: **repeat**
15: Take one unallocated subchannel j.
16: Let the users set on the right or left of subchannel j is L_j (presented in Figure 2).
17: Obtain the base cumulative rate $\chi(j) = \sum_{(c,i)\in L_j} R_{ci}$.
18: Obtain the base cumulative utility $U(j) = \sum_{(c,i)\in L_j} U_{ci}$.
19: We index the shifting operation by s and the set holding all indexes by L_{sj}.
20: Provide the edge users in set L_j (e.g., user 1 in Figure 2) available subchannels if necessary, $\forall s \in L_{sj}$.
21: Determine the cumulative rate $\chi_s(j)$ and the cumulative utility $U_s(j)$, $\forall s \in L_{sj}$.
22: Determine the shifting set L'_{sj} so that $(\chi^{max}(j) - \chi_s(j)) <= H$ and $U_s(j) >= U(j)$.
23: Find optimal s, $s^* := \arg\max_{s\in L'_{sj}} U_s(j)$.
24: **until** No Improvement is possible

Worst case computational complexity of step 5 is $O(MNlogN)$. Other steps, such as steps 8, 10 have the worst case complexity of $O(M)$ which is dominated by step 5. Hence, total computational complexity of the algorithm before step 14 (or stage 2) is $O(MN^2logN)$. For the 2nd stage of the algorithm, we discuss its worst case complexity as follows. In the worst case, from the 1st stage, 1 user gets 1 subchannel and M subchannels of all users are adjacent. Therefore, remaining number of unallocated subchannels is $N-M$ and the loop of the 2nd stage runs $N-M$ times. Inside the loop, we may need to shift M times and each shifting requires power control and other primitive steps which are of $O(1)$ complexity. Hence, the worst case complexity of the 2nd stage is $(N-M)M$ which is again dominated by the complexity of the 1st stage.

IV. Characteristics of Proposed Scheduling Scheme

In this section, first we prove our scheduling scheme achieves much fairer resource distribution compared to [40]. Second, we discuss the impact of opportunity cost function on scheduling outcome and show that by certain value of H, the proposed scheduling scheme can bridge between MT (which has the best global performance) and PF (which is well-known for proportional fairness) schemes. In the last paragraph, we have briefly discussed the practicality of our scheduling scheme w.r.t. other schedulers implemented in practice.

Lemma 1: Our scheduling scheme is more efficient in terms of achieving fairness compared to [40].

Proof: Consider an epoch where there is 3 subchannels and 3 remaining users available to schedule. User 1 has the best SNR condition for all subchannels, i.e., $e_{1j} >> e_{2j}, e_{1j} >> e_{3j}, \forall j \in [1,3]$. User 2 has almost similar SNR condition compared to user 3, however slightly better. Packet HOL delay for user 1 is much smaller compared to user 2, i.e., $D_1 << D_2$ (or $\zeta_1 >> \zeta_2$) and $D_2 \approx D_3$ (or $\zeta_2 \approx \zeta_3$). If we would apply the scheduling technique presented in [40], it is very likely that rate obtained by user 1 is much higher compared to that of user 2 or 3 due to the favorable physical layer condition. In our simulation, we observe, if any user has favorable channel condition, it is more likely, that user obtains more subchannels comparing with others, hence, in our scenario, there is a chance that user 1 gets 2 subchannels and user 2 gets 1. However, the result is not fair given the packet HOL delay or throughput. With our scheme, possible subchannel allocations are: **1)** user 1 may get first subchannel due to its very good SNR condition on all subchannels, then subchannel 2 is assigned to user 2 due to its better SNR condition on remaining subchannels and almost similar packet HOL delay or throughput compared to user 3, finally user 3 gets the remaining last unallocated subchannel. **2)** user 1 may obtain 1st subchannel due to the similar reason described above, user 2 gets remaining 2 subchannels because of its SNR condition and worst QoS performance compared to user 3.These 2 scheduling decisions are considered as fair compared to that by [40]. In order to show that there is no scenario that the scheduling technique [40] outperforms ours, we introduce a counter example of this example. Similar to this example, consider user 1 has the best channel condition on all subchannels compared to other two users. Unlike before, average packet HOL delay or throughput of this user is much worse compared to other two users. The channel condition of user 2 is little better compared to user 3 on all subchannels, however much worse compared to user 1. Furthermore, packet HOL delay or throughput for these two users are almost similar, however better than user 1. For the similar reason explained for the previous example, by the scheduler [40], user 1 obtains first 2 subchannels and user 2 obtains the 3rd subchannel. On the other hand, by our scheduling scheme, two possible subchannel allocations are: **1)** user 1 gets first 2 subchannels because of its best channel condition and worse packet HOL delay or throughput. Rest 1 subchannel will be allocated to user 2 because of its slightly

better channel condition compared to user 3. **2)** user 1 obtains all 3 subchannels. The scheduling decisions by both [40] and our scheduling scheme are considered as fair. Hence, there is no scenario for which the scheduler [40] can perform better compared to our scheduler. Our scheduling scheme always achieves better or as good as performance in terms of fairness comparing with [40].

Lemma 2: Under special case ($H = 0$), presented algorithm converges to the MT scheme.

Proof: In the 1st stage of *Algorithm 1*, MT scheme assigns subchannel $l_{ci}(j)$ to user (c,i) when it achieves the maximum quantity for $g_{ci}(j)$ compared to other users; whereas for the 2nd stage, it selects shifting operation which has utility greater than the base cumulative utility and have the maximum value comparing with other shifting operations. Set $H = 0$, at the 1st stage, our scheduler selects only one user for set L_j whose $g_{ci}(j)$ has the maximum quantity in order to satisfy the condition $\Gamma^{max}(j) - \Gamma_{ci}(j) <= 0$. Afterward, since L_j has only one user, eventually this user will be picked up for subchannel $l_{ci}(j)$. For the 2nd stage, similar situation happens, L'_{sj} contains only the shifting operation which incurs the highest utility and has larger value than the base cumulative utility. Hence, converging trend of our algorithm towards MT scheme has been proved.

Lemma 3: For special value (maximum achievable rate, $Rate_{Max}$) of H, presented algorithm transforms to the PF scheme.

Proof: At scheduling epoch t, while scheduling subchannel j, PF algorithm assigns subchannel $l_{ci}(j)$ to user (c,i) if its metric $log_2(1+\frac{g_{ci}(j)}{\zeta_{cij}})$ obtains larger quantity compared to other users. Note that, PF metric is an increasing function of $g_{ci}(j)$ and decreasing function of its long term throughput ζ_{cij}. Consider about our scheme with $H = R_j^{max}$, set L_j contains all potential users in the system. And then, the scheduler picks the user whose marginal utility, i.e., slope obtains the highest quantity. Slope of a user is increasing function of $g_{ci}(j)$ and has a decreasing trend w.r.t. to its long term throughput ζ_{cij}. Mathematically, we can equate both functions in order to find instantaneous value for the constants. Hence,

$$log_2(1+\frac{g_{cij}}{\zeta_{cij}}) \;=\; -a_c\left(\mu_{ci}^1 - \frac{\zeta_{ci}^{min}}{\zeta_{cij}} + \mu_{ci}^2 - \frac{g_{cij}}{\Gamma_j^{max}}\right).$$

Simplifying this, we obtain

$$\mu_{ci}^2 + \mu_{ci}^1 \;=\; -\frac{1}{a_c}\log_2(1+\frac{g_{cij}}{\zeta_{cij}}) + \frac{\zeta_{ci}^{min}}{\zeta_{cij}} + \frac{g_{cij}}{\Gamma_j^{max}}.$$

Therefore, we can conclude, at scheduling instant t while allocating subchannel j, if the value of $\mu_{ci}^2 + \mu_{ci}^1$ is equivalent to the right side of above equation (for user (c,i)) and if we ignore our opportunity cost function, over the infinite time our scheduling algorithm converges to PF scheme. In the similar manner, for the 2nd stage of *Algorithm 1*, we can prove that at scheduling epoch t, certain value of constant $\sum_{(c,i)\in L_j}\mu_{ci}^2 + \mu_{ci}^1$ results in asymptotic convergence towards PF scheme.

Proofs 2 and 3 remind us that the presented scheduling algorithm in this paper is generalized. In general, this scheduler works in between these two extreme cases which has been proved in the next section. Since the optimal solution is an iterative procedure, there is no closed form proof for Lemma 2 and 3 for this. However, for the same condition of these two Lemmas, optimal solution converges to the MT and PF schemes respectively. Adjusting the parameter of the opportunity cost function, we can convert this scheduler to two conventional extreme cases of scheduling scheme.

Furthermore, we have noticed that Huawei [45] has implemented an enhanced PF scheduler for the GBR traffic of an exemplary 3G/4G system, i.e., LTE system. Packet delay budget of different GBR traffic and aggregate RB (Minimal resource unit of LTE systems) quality are considered while designing this scheduler. For the uplink scheduling, this scheduler first calculates users' priority metric which is a function of average packet delay of that user and approximate average channel quality of all RBs and then based on the priority, it allocates RBs among those users. Priority metric is calculated using MW formula and M-LWDF is very similar version of MW rule. Whereas, our scheduler does one by one subchannel (equivalent to RB) allocation based on the instantaneous subchannel's channel quality and average packet delay which is apparently more dynamic. Moreover, our scheduler can deal the traffic with throughput requirement and design of the scheduler is very flexible for diverse QoS (e.g., packet loss ratio, packet jitter) oriented traffic. Beside, the scheduler designed by Ericsson [46] applies conventional traffic policing and shaping concept while allocating RBs among different QoS based traffic. Policing ensures that the users do not get the agreed upon configured rate, whereas traffic shaping makes sure that the users get at least minimum settled QoS specified in their agreement with the vendor. Based on the traffic policing and shaping mechanisms, our scheduler is dynamic and can achieve the similar level of performance as the scheduler equipped with these policies.

V. Performance Evaluation

In this section, we will evaluate *Algorithm 1* for four classes of traffic of a typical 3G/4G system such as LTE system. First, we outline the simulation methodology we have adopted. Then the simulation results splitted in two parts for Homogeneous and Heterogeneous traffic.

A. Simulation Methodology

For setting up the network, we put the base station at the center of a cell, user nodes at different distance surrounding the base station. Cell radius is assumed as 1 km. We run the simulation over 5000k TTIs (Transmission Time Intervals). One TTI is equivalent to 1ms and it consists of 25 RBs. Each RB is analogous to 12 subcarriers [47]. These total 25×12 resource elements are spread over 5 MHZ bandwidth. The theoretical limit [48] of the channel capacity is given by $\beta = \frac{-1.5}{ln(5P_b)}$, where P_b denotes the BER (Bit Error Rate). BER for the channel is configured as 10^{-6}. Each user's maximum power is set as 220 mW. In order to calculate log-normal

shadowing effect of the channel, we assume the reference distance is 1 km and the SNR for this reference distance is 28 dB. Reference shadowing effect is the log normal distribution with variance 3.76. With this variance, log-normal shadowing power is determined as 10.6 dB according to [49]. Rayleigh fading effect is captured with a parameter a such that $E[a^2] = 1$. Channel gain for a particular user (c,i) over RB j is computed by Equation 22.

$$G_{cij,dB} = (-\kappa - \lambda log_{10} d_{ci}) - \xi_{cij} + 10 log_{10} F_{cij}. \quad (22)$$

In Equation 22, the first factor κ captures propagation loss, the value of which is 128.1 dB. d_{ci} is the distance in km from user (c,i) to the base station and λ is the path loss exponent which is set to a value 3.76. The second factor ξ_{cij} captures log-normal shadowing effect for the reference distance. Whereas the last factor F_{cij} corresponds to Rayleigh fading effect. Feedback duration due to the exchange of CSI (Channel State Information), scheduling decisions between the users and base station is considered as negligible. Perfect CSI estimation is assumed at the base station.

In order to demonstrate the ability of our uplink scheduling scheme, we need to justify that users with different QoS measures obtain expected service which has already been specified theoretically. In the previous section, we have designed the utility function with 3 different QoS measures. Now, we want to define 4 different classes for the users where each class is accompanied with one QoS measure. Four different classes are: VoIP (class 1), audio streaming (class 2), video streaming (class 3) and FTP (class 4). Class 1 has the highest priority and class 4 is the type of lowest priority. Class 1 and 2 traffics have delay bound constraint. Video streaming has minimum throughput constraint[5], whereas class 4 is the type of best effort traffic. In order to distinguish priority of different classes, we set the parameters for a_c and μ_{ci} which are shown in Table I.

For VoIP traffic, we have taken AMR (Adaptive Multi rate) codec [50] method. According to this model, packets are generated using a negative exponentially distributed ON-OFF pattern to replicate the talk and silent duration of a VoIP call. The mean duration of ON and OFF periods are 3 s. During the ON period, in every 20 ms interval, a voice packet of 244 bits is generated. Including the compressed IP/UDP/RTP header, the data rate for each VoIP flow becomes 13.6 Kbps. According to [51], the maximum acceptable delay for voice is 250 ms. Considering delays induced by the core network and the delay for RLC and MAC buffering, the tolerable delay at the radio interface should be at most 100 ms [47] which represents a very strict requirement. For modeling audio streaming traffic, we have also used AMR codec. Size of the packets generated for each audio streaming user is uniformly distributed between 244 and 488 bits and hence the data rate varies from 12 Kbps to 64 Kbps. The maximum delay threshold has been set for each user 150 ms. Video streaming is modeled with a minimum data rate of 64 Kbps and a maximum data rate of 384 Kbps. Size of the packet in each video flow is uniformly distributed between 1200 and 2400 bits. Hence, the resultant each video flow generates the number of packets which is uniformly distributed between 1 and 3 at the interval of 19 ms.[6] The data rate of each user with FTP traffic is assumed as maximum 128 Kbps with the size of a packet 1200 bits. Packet generation interval of each FTP flow is 10 ms. In the simulation, we assume each user is equipped with one class of traffic and it keeps holding that flow until the simulation is finished. All the results described in the following subsection are average of 20 simulation runs. As the simulation tool, we have modified the matlab based LTE simulator [52] with all functionalities of our scheduling scheme.

[5]There are two types of video traffic, e.g., interactive video with stringent real-time requirements; video streaming and video download with some minimum bandwidth requirement. In our simulation, each video flow is of second type.

TABLE I
SIMULATION PARAMETERS

Traffic Type	a_c	μ^1_{ci}	μ^2_{ci}
VoIP	4	0.4	0.3
Audio	3.5	0.4	0.3
Video	3.0	0.4	0.3
FTP	2.5	0.4	0.6

B. Simulation Results

First we show the simulation results for homogeneous traffic, i.e., VoIP and traffic with throughput requirement (video streaming). With regard to this experimentation, we have compared the results obtained by our scheduling algorithm with the existing work. For VoIP traffic, we have compared our results with M-LWDF [37], EXP [16] rules and [36]. EXP and M-LWDF rules are specifically developed for the downlink scheduling of delay sensitive multimedia traffic in OFDM based systems, whereas [36] is for the uplink scheduling scheme. Since these schemes have limitations and are not directly comparable to ours, we have extracted scheduling rules from those works and substitute in our algorithm in order to have valid comparison. The scheduler under M-LWDF rule assigns RB $l_{ci}(j)$ to the user $(c,i)^*$ abiding by the formula

$$(c,i)^* = \arg\max_{(c,i)\in\mathbf{N}} \frac{g_{ci}(j)}{\zeta_{ci}} D_{cij}.$$

And, the scheduler with EXP rule obeys the following rule for RB $l_{ci}(j)$

$$(c,i)^* = \arg\max_{(c,i)\in\mathbf{N}} b_{ci} exp\left(\frac{a_{ci} D_{cij}}{1+\sqrt{(1/N)\sum_{(c,i)} D_{cij}}}\right) g_{ci}(j),$$

where $b_{ci} = 1/E[g_{ci}]$ and $a_{ci} = \frac{6}{D^{max}_{ci}}$.

[6]Each video flow is designed following the trace file "sony" taken from "http://trace.eas.asu.edu/h264/". Statistics of this trace is: **a.** Inter-arrival time between two bursts is constant (33 ms).**b.** Burstiness of the video is determined by the size of the burst. Maximum frame burst of this video is 326,905 bytes and the minimum one is 20,209 bytes. Burstiness of our each video flow is determined by the number of packets and packet size. According to our statistics, maximum frame size of each video flow is 7200 bits and the minimum one is 1200 bits.

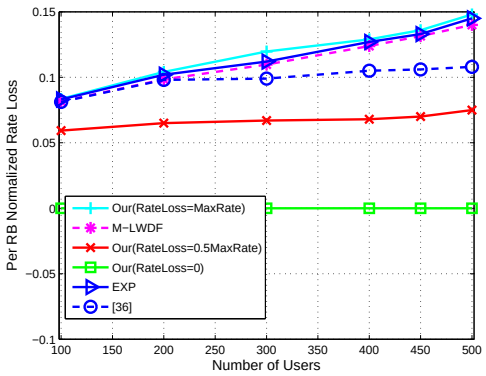


Fig. 4. Per RB normalized rate loss comparison between our scheme and other (VoIP).

For ensuring guaranteed bitrate, we did not find much work in the literature except [10]. Therefore for the homogeneous video streaming setup, we have done performance comparison with [10] as well as with the MT and PF schemes.

Although [16] is the solution for mixed traffic, they did not consider the traffic with guaranteed bit rate in evaluating performance. Most recent work dealing all types of traffic is [28] and we have taken it as performance benchmark while presenting the results of mixed traffic obtained by our scheme.

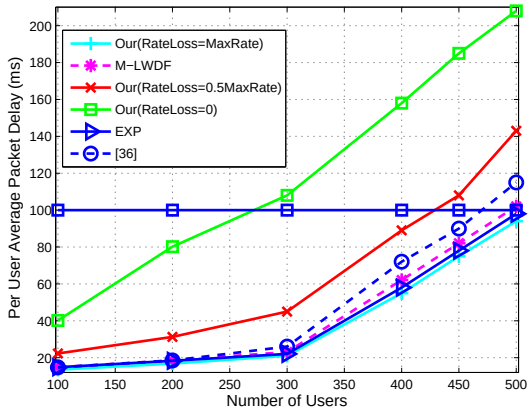
For the performance metrics, we consider average packet delay in the system, average normalized throughput, proportion of per RB rate loss on behalf of service provider, per RB normalized utilization, total system's effective rate, fairness. For measuring fairness over all users in the system, we have used well-known fairness indicator named as Jain Fairness Index (JFI) [53]. Proportion of per RB rate loss is the indication of how much proportion of rate is sacrificed from the maximum one (that could be achieved) while taking the scheduling decision of each RB. On the other hand, per RB utilization is the measure of RB utilization over the entire simulation interval. While computing RB utilization, 0 is counted when any RB stays vacant and contributes to the average per RB utilization. Furthermore, we consider users are spread uniformly between 0.5 km to 0.8 km distance from the base station.

In the second part of simulation results, we have studied the performance of multiplexed traffic. The simulation scenario is designed in such a way that we see the strength of our scheduling scheme which can prioritize different classes of traffic and can reach to an elegant scheduling decision.

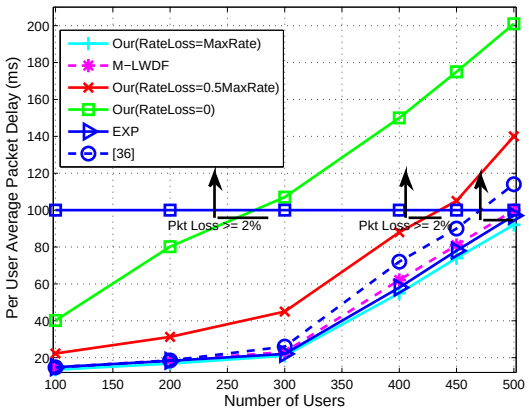
1) Homogeneous Traffic (VoIP): Figure 3(a) depicts the average packet delay of VoIP traffic with respect to total number of users in the system. First observation from this figure is, with the increased number of users, the average packet delay is increased for all cases. We have shown the results of our algorithm for different maximum tolerable rate loss, i.e., $Rate_{Max}$, $0.5Rate_{Max}$ and 0. As we mentioned before that, if maximum tolerable rate loss is $Rate_{Max}$, it treats all users as if the network operator does not have problem with any rate loss, this setting treats all users equally. Figure indicates, for the parameters given in Table I, our scheme (with maximum tolerable rate loss $Rate_{Max}$) has better performance compared to the scheduler with EXP and M-LWDF rules. The M-LWDF and EXP rules are specifically designed for the delay sensitive multimedia network. The formula of the M-LWDF scheduler is based on the product of the packet HOL delay and PF factor. Hence, resultant scheduling policy ensures fairness among the users. The scheduler with EXP rule is more robust than the scheduler with M-LWDF rule. This is because, the exponential function grows much faster with its argument. Furthermore, EXP rule also takes the overall network status into account, because the delay of considered user is somehow normalized over the sum of the experienced delay of all users. Better performance of our scheme w.r.t. M-LWDF and EXP rules justifies its efficacy for the usage of delay sensitive VoIP traffic.[7] At low load, [36] performs almost in the similar manner compared to ours with negligible deviation. However, at high load, the users with moderately better channel condition get more priority while the users with worst channel condition almost starve, because scheduling rule depends on the time difference between current and the recent burst. Therefore, within a fixed time period, the users with good channel condition get scheduled while keeping the users with worst channel starved. While serving these users, another burst of traffic arrives for all users and therefore delay based metric is replaced by the same value for all users including the users with worst channel. Therefore, in the second round, for the same value of delay based metric, same set of users with good channel condition get served which results in starvation for the users with worst channel. If the delay based metric of [36] would consider the time of first burst instead of recent, it would perform better at high load. When the maximum tolerable rate loss is 0, our scheme performs poorly. This is because, the scheduler cannot tolerate any rate loss and it only serves high rate generating users which causes higher packet delay for other users. Due to higher packet delay of low rate generating users, resultant average packet delay of the system gets higher. Performance of our scheduler with the maximum tolerable rate loss $0.5Rate_{Max}$ has in-between performance of other two for the similar reason. Instead of using infinite buffer, if we use finite buffer at each terminal with limited size, we observe packet loss. We define the outage of a system with VoIP users when its packet loss exceeds 2%. This is because, if a user suffers at least 2% packet loss, it is likely that the packets of that user cannot be decoded. Therefore, due to the limited buffer even though the average packet delay of the system is below the noted limit, QoS is not met for the users in the system because of more than 2% packet loss. Hence, we see in Figure 3(b) that the coverage of the system with finite buffer is at earlier point compared to that with infinite buffer. This justification applies for all schedulers.

Since our scheduler with the maximum tolerable rate loss $Rate_{Max}$ experiences lower average delay, it implies, the scheduler gives more priority to the users with worse channel

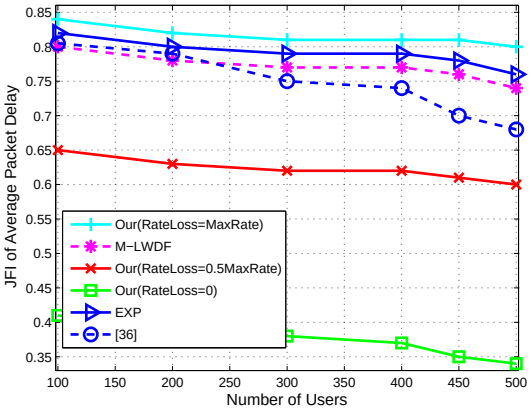
[7]By adjusting the parameters μ_{ci}^1, μ_{ci}^2 in our scheduler and the parameters a_{ci} and b_{ci} in EXP scheduler, better or comparable performance (w.r.t. our scheduler) can be achievable by the scheduler with EXP or M-LWDF rule.



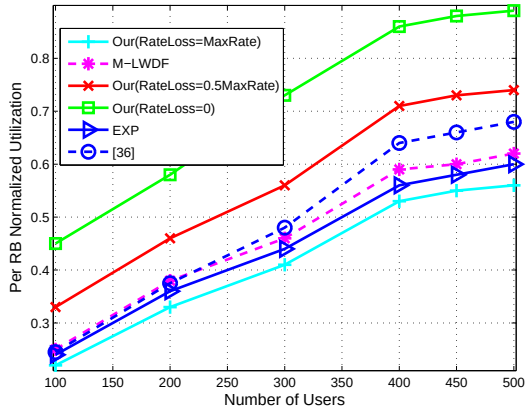
(a) Per user average packet delay comparison between our scheme and others (infinite buffer).



(b) Per user average packet delay comparison between our scheme and others (finite buffer).



(c) JFI of average packet delay comparison between our scheme and others.



(d) Per RB normalized utilization comparison between our scheme and others.

Fig. 3. VoIP.

condition compared to the scheduler with M-LWDF or EXP rule. Therefore, in terms of fairness, our scheduler with the maximum tolerable rate loss $Rate_{Max}$ outperforms the rest others as depicted in Figure 3(c). The utility function combined with its parameters of our scheduler are specifically designed to preserve fairness across the system while exploiting users' varying channel quality condition. With the decrementing maximum tolerable rate loss, we see decrementing JFI. This behavior is expected, because the scheduler picks up selectively high rate generating users and hence the scheme with the maximum tolerable rate loss 0 has the worst fairness. Metrics such as percentage of RB utilization and rate loss are the best for this case and have been illustrated in Figures 3(d) and 4 respectively. With the decrementing tolerable rate loss, the scheduler gradually ignores users with better channel condition and tries to serve users whose average delay tends to deviate from the prescribed bound and hence, we see decrementing RB utilization and incrementing rate loss.

2) Homogeneous Traffic (Video Streaming): In order to present results for this case, similar simulation setup as in the previous subsection has been undertaken. Figure 5(a) shows average normalized throughput with the increasing total

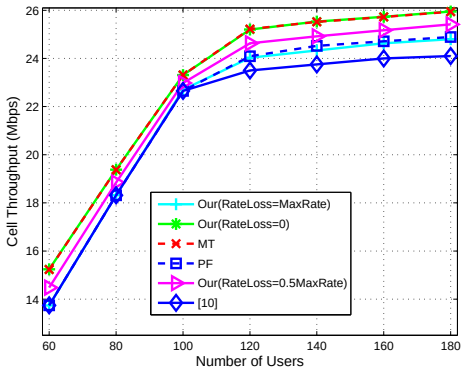
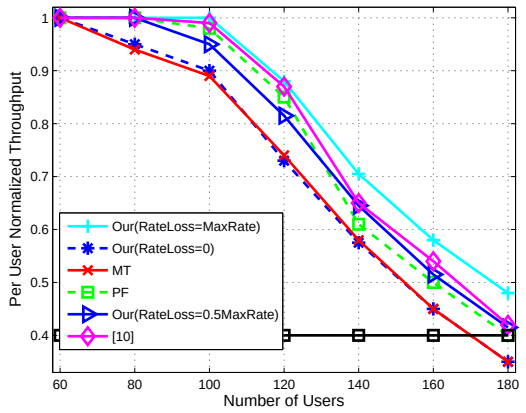
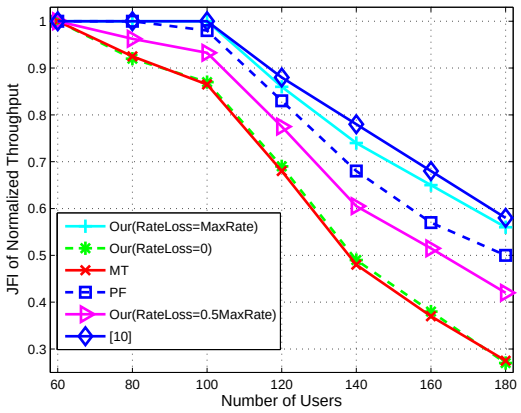


Fig. 6. Cell throughput comparison between our scheme and others (Video).

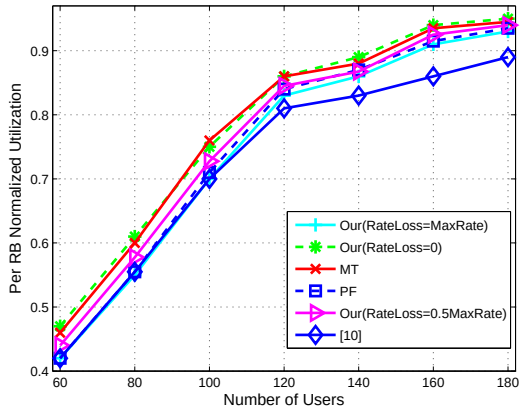
number of users. It is expected that increased number of users deteriorates per user normalized throughput due to the limited number of RBs available at each scheduling epoch. In the figure, we have shown the performance of our scheme with three maximum tolerable rate loss, i.e., $Rate_{Max}$, $0.5Rate_{Max}$ and 0. Our scheme with maximum tolerable rate loss $Rate_{Max}$



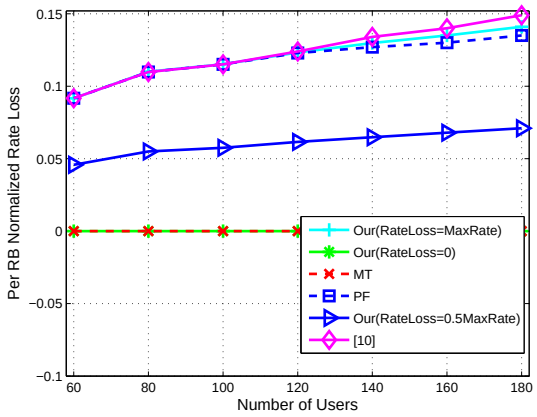
(a) Per user normalized throughput comparison between our scheme and others.



(b) JFI of normalized throughput comparison between our scheme and others.



(c) JFI of average packet delay comparison between our scheme and others.



(d) Per RB normalized rate loss comparison between our scheme and others.

Fig. 5. Video streaming.

has the best performance. Because of the fairness measure in terms of per user average throughput inside the exponential utility function, our scheme ensures fairness across the users. In [10], the scheduler works in both time and frequency domains. Time domain task partitions the users based on their throughput being less and greater than the noted limit. Former list has absolutely higher priority than the later one. Moreover, priority metric for the first list is determined by blind equal throughput, whereas the metric for the second one is ratio of instantaneous wide band channel quality and past average throughput. From the sorted priority list, N number of users are passed to the frequency domain scheduler which applies PF scheme in order to finalize the scheduled users. PF scheme uses the product of instantaneous RB quality and inverse throughput in order to ensure fairness. This scheduler pre-processes the users while giving priority to the ones with lower throughput before applying PF technique on them. Therefore, at low load, PF technique and [10] perform almost similarly as the higher priority list selected by the time domain scheduler is empty and hence, there is no basic performance difference between them. However, at high load, higher priority list [10] gets bigger and bigger and consequently it performs better compared to the PF one. It is unusual to see that MT scheme has lower per user normalized throughput, however there is a intuitive reason behind it. This scheduler gives higher priority to the users with better channel condition even though those contribute less towards overall system throughput and hence, several users with worse channel remain under-provisioned. Same thing happens to our scheduler with the maximum tolerable rate loss 0, better rate generating users are given priority in the scheduling decision at the expense of low rate generating users and hence, the resultant normalized average throughput of the system is lower. If we use limited buffer at each user terminal, we notice even lower coverage for all cases because of the buffer overflow resulting in packet loss.

Unlike the results discussed in the previous paragraph, we observe better performance for MT scheduler or our scheme with the maximum allowable rate loss 0 in terms of overall cell throughput or percentage of RB utilization. Results with this respect are given in Figures 6 and 5(c) respectively. MT or our scheme with the maximum tolerable rate loss 0 selects only the users with better channel condition or higher rate generating users. In either case, such behavior of the scheduler increases cell throughput or enhances per RB utilization, however at

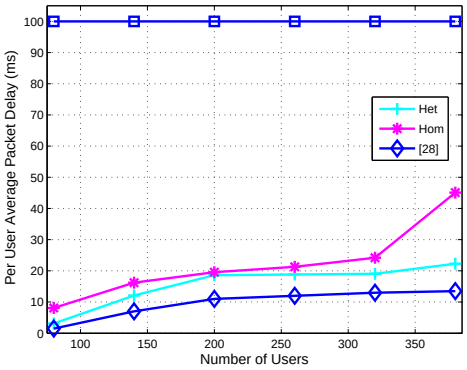


Fig. 7. Per user average packet delay comparison between homogeneous and heterogeneous setup (VoIP).

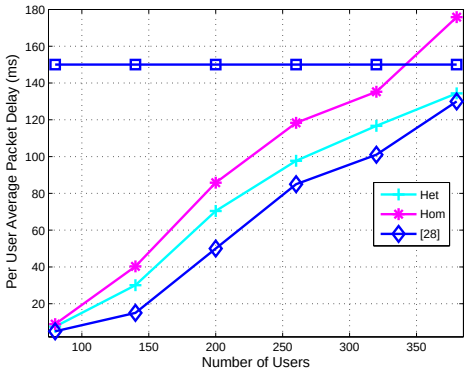


Fig. 8. Per user average packet delay comparison between homogeneous and heterogeneous setup (Audio).

the expense of throughput of the users with worse channel condition or low rate generating users. Since [10] keeps busy in serving low rate generating users at high load, per RB utilization is the worst for this scheduler. Because of giving privilege to the best users at each scheduling epoch, percentage of rate loss is 0 for MT scheme or our scheme with maximum tolerable rate loss 0 as depicted in Figure 5(d). Whereas, the scheduling decision of other scheme [10] not necessarily depends on user channel condition or rate, they check per user's instant throughput and relevant constraint. And, hence, the scheduler of this policy suffers higher rate loss. Fairness in terms of JFI comparison among all schemes are given in Figure 5(b). Because of the exponential nature of the utility function if any user goes under the minimum throughput, our scheduler (with maximum tolerable rate loss $Rate_{Max}$) is forced to serve that user and hence, ensures fairness across the users while exploiting their instantaneous channel condition. Because of the nature of the utility function, PF scheme ensures fairness across the system, however not as good as our scheme. At low load, the scheduler [10] behaves like PF scheme and so thus fairness. However, at higher load, its time domain scheduler is almost ignorant of the spectral efficiency, gives priority to the users with lower throughput and hence, achieves the highest fairness compared to all. Our scheme does not deviate much from [10] in terms of fairness. With decreasing maximum tolerable rate loss, our scheme also suffers from fairness measure because of giving privilege to the higher rate generating users.

3) Heterogeneous Traffic: In order to prove that our approach can handle multiplexed traffic efficiently, we have deployed N users splitted equally and uniformly into 4 classes. Moreover, maximum tolerable rate loss is assumed as $Rate_{Max}$ and buffer size in each terminal is considered as infinite. Figures 7 and 8 compare the average packet delay of VoIP and audio streaming users respectively with that of [28]. In the same figure, we have shown the results from the homogeneous setup obtained by our scheme. As VoIP has higher priority than the audio streaming, in the heterogeneous setup, VoIP users will always incur lower packet delay comparing with the audio streaming users. Moreover, heterogeneous VoIP

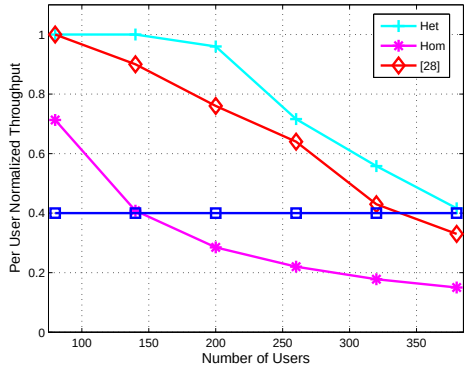


Fig. 9. Per user normalized throughput comparison between homogeneous and heterogeneous setup (Video).

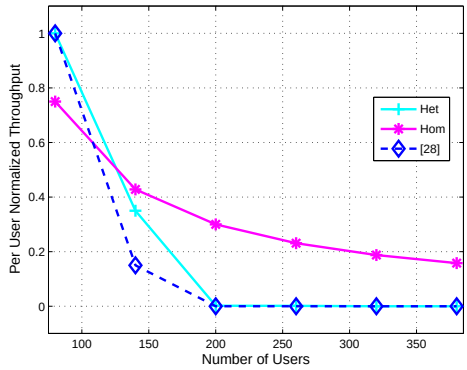


Fig. 10. Per user normalized throughput comparison between homogeneous and heterogeneous setup (FTP).

users always have lower average packet delay than the case with homogeneous setup. This is because, total users in the multiplexed case is equally divided among rest 3 lower priority classes, and hence, there are fewer VoIP users in this case comparing with the single type one. [28] always gives the highest priority to VoIP users whenever they have data in the queue no matter the accumulated delay of the users is much lower than the bounded limit. On the other hand, audio streaming users under multiplexed setup not necessarily always have lower packet delay comparing with that of homogeneous case although there are some lower priority users along with them. Since the data rate of each audio streaming users is higher than that with VoIP users, at low load, when the network is unsaturated, multiplexed less number of audio streaming users have enough resource to get their packets transmitted. However, as we increase the load, the number of VoIP users increases, they may occupy entire resource of the network, in that case, we will be able to see higher packet delay for the multiplexed case than the single class case. Audio streaming is the second highest priority class and data rate of the users with this class is much larger than that with VoIP, the scheduler in [28] apparently serves the users of this type whenever they have packets in the queue no matter their packet HOL delay is lower than the noted limit. This reason results in lower average packet delay compared to our scheme. However, when the network becomes saturated with the VoIP users, audio streaming users will achieve closer or worse performance compared to ours.

Figures 9 and 10 depict average normalized throughput for the video streaming and FTP users. Similar to the previous figures, here, we have shown the results for the homogeneous setup achieved by our scheduler. Since video streaming is of higher priority, users of this type incurs higher throughput than that of FTP. Similar to the audio streaming users, at low load, under multiplexed case, video streaming users achieve better performance comparing with the homogeneous users. In addition, each video streaming user has much higher data rate comparing with the combined VoIP and audio streaming users. If we increase the number of users in the network and it goes close to saturation with the VoIP and audio streaming users, video streaming users start to starve and in that point, we will see poor performance of heterogeneous video streaming users. Same reasoning applies to FTP users, at low load, we will see better performance for the heterogeneous case. However, when all the resource of the network is consumed by all higher priority traffic, performance of heterogeneous FTP users starts to deteriorate. Because of the blind provision towards VoIP and audio streaming users given by the scheduler [28], video streaming users perform poorly compared to ours. For the similar reason, the throughput of FTP users is even worse comparing with our scheme.

Figure 11 compares admission region of VoIP and video streaming users in the system under multiplexed setup with [28]. [28] redundantly gives more provision to VoIP users, average per user packet delay designed by them is lower than that by our scheme as depicted in Figure 7. It starves video streaming users although the delay of VoIP users is lower than the required limit. Therefore, from the figure, we observe, although the number of VoIP users with QoS assurance is equal to our scheme, number of video streaming users with guaranteed throughput is lower at increased load.

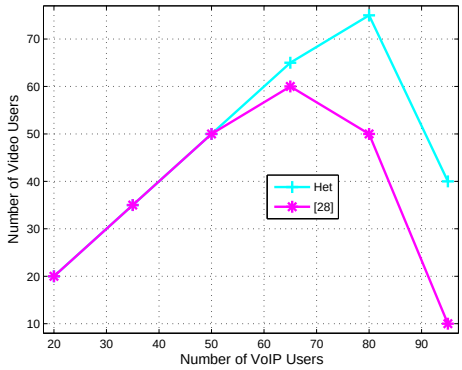


Fig. 11. Admission region with QoS guarantee comparison between our scheme and other.

VI. Conclusion & Future Work

3G/4G technologies such as LTE or LTE-advanced are specifically envisioned to meet the growing demand of high data rate for different applications and SC-FDMA is the recommended uplink multiple access scheme of such systems. Varying emergent applications prompt the development of heterogeneous traffic networks with diverse QoS requirements. Beside the demands of end users, resource utilization is an important matter to consider in order to assist the network operators. In this paper, we have presented an uplink scheduling technique for SC-FDMA based heterogeneous traffic networks which is able to maintain varying QoS provision of end users while keeping the resource utilization (in subchannel level) of the service providers in the prescribed range. In order to solve this problem, we have first formulated the problem considering all standard specific constraints while capturing QoS factors of the users in a smart utility function. In addition to these, the formulation consists of a constraint which allows the service provider to keep granular subchannel utilization level in some certain range. Having considered the discrete nature of subchannel allocation, we have shown the optimal solution structure of the problem which has high computational complexity. From the guiding principles of the optimal solution structure, we have proposed a sub optimal algorithm with polynomial time complexity. Furthermore, we have evaluated our scheduler in a network which has three different types of users, i.e., traffic with delay constraint, traffic with minimum data rate requirement and best effort traffic. Results obtained from the extensive simulation show that proposed scheme exhibits the tradeoff between the fairness of end users and the resource utilization of the service providers. By showing the simulation results for both homogeneous and heterogeneous traffic networks in terms of several performance metrics, we have justified the efficacy and effectiveness of our scheduling scheme envisioned to be deployed in future 3G/4G like LTE systems. Investigating the long term capacity of our

proposed scheduler analytically is our next step to pursue. We may need to resort to the simplified channel and user traffic model in order to achieve this objective.

REFERENCES

[1] R. Kwan and C. Leung, "A survey of scheduling and interference mitigation in lte," *JECE*, vol. 2010, pp. 1:1–1:10, Jan. 2010. [Online]. Available: http://dx.doi.org/10.1155/2010/273486

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

[3] F. P. Kelly, A. K. Maulloo, , and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237–252, 1998.

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

[5] P. Kela, J. Puttonen, N. Kolehmainen, T. Ristaniemi, T. Henttonen, and M. Moisio, "Dynamic packet scheduling performance in utra long term evolution downlink," in *3rd International Symposium on Wireless Pervasive Computing (ISWPC)*, 2008, pp. 308–313.

[6] D. Tse and P. Viswanath, "Fundamentals of wireless communication," in *Cambridge University Press; Edition 1*, 2005.

[7] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for ofdm systems," *IEEE Trans. Wirel. Commun.*, vol. 8, no. 1, pp. 288–296, 2009.

[8] R. Kwan, C. Leung, and J. Zhang, "Proportional fair multiuser scheduling in lte," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 461–464, 2009.

[9] F. Calabrese, C. Rosa, K. Pedersen, and P. Mogensen, "Performance of proportional fair frequency and time domain scheduling in lte uplink," in *European Wireless Conference (EW)*, 2009, pp. 271–275.

[10] G. Monghal, K. Pedersen, I. Kovacs, and P. Mogensen, "Qos oriented time and frequency domain packet schedulers for the utran long term evolution," in *IEEE VTC-Spring*, 2008, pp. 2532–2536.

[11] N. Chen and S. Jordan, "Throughput in processor-sharing queues," *IEEE Trans. Automatic Control*, vol. 52, no. 2, pp. 299–305, 2007.

[12] ——, "Downlink scheduling with probabilistic guarantees on short-term average throughputs," in *IEEE WCNC*, 2008, pp. 1865–1870.

[13] D. Skoutas and A. Rouskas, "Scheduling with qos provisioning in mobile broadband wireless systems," in *European Wireless Conference (EW)*, 2010, pp. 422–428.

[14] M. Andrews, "Cdma data qos scheduling on the forward link with variable channel conditions," *Bell Laboratories*, Apr 2000.

[15] H. Ramli, R. Basukala, K. Sandrasegaran, and R. Patachaianand, "Performance of well known packet scheduling algorithms in the downlink 3gpp lte system," in *IEEE 9th Malaysia International Conference on Communications (MICC)*, 2009, pp. 815–820.

[16] B. Sadiq, R. Madan, and A. Sampath, "Downlink scheduling for multiclass traffic in lte," *EURASIP J. Wirel. Commun. Netw.*, vol. 2009, pp. 14:9–14:9, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/510617

[17] G. Song and G. Li, "Cross-layer optimization for OFDM wireless networks part i: Theoretical framework," *IEEE Trans. Wirel. Commun.*, vol. 4, no. 2, March 2005.

[18] V. J. Venkataramanan and X. Lin, "On wireless scheduling algorithms for minimizing the queue-overflow probability," *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 788–801, 2010.

[19] A. L. Stolyar, "Large deviations of queues sharing a randomly time-varying server," *Queueing Syst. Theory Appl.*, vol. 59, no. 1, pp. 1–35, may 2008. [Online]. Available: http://dx.doi.org/10.1007/s11134-008-9072-y

[20] B. Sadiq and G. de Veciana, "Large deviation sum-queue optimality of a radial sum-rate monotone opportunistic scheduler," *CoRR*, vol. abs/0906.4597, 2009.

[21] R. Basukala, H. Mohd Ramli, and K. Sandrasegaran, "Performance analysis of exp/pf and m-lwdf in downlink 3gpp lte system," in *First Asian Himalayas International Conference on Internet*, 2009, pp. 1–5.

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

[23] W. K. Lai and C.-L. Tang, "Qos-aware downlink packet scheduling for LTE networks," *Computer Networks*, no. 0, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128613000352

[24] M. Iturralde, A. Wei, T. Ali Yahiya, and A. L. Beylot, "Resource allocation for real time services using cooperative game theory and a virtual token mechanism in lte networks," in *IEEE CCNC*, 2012, pp. 879–883.

[25] M. Lerida, "Adaptive radio resource management for voip and data traffic in 3gpp lte networks," in *KTH Royal Institute of Technology, Stockholm*, 2008.

[26] M. Gidlund and J.-C. Laneri, "Scheduling algorithms for 3gpp long-term evolution systems: From a quality of service perspective," in *IEEE ISSSTA*, 2008, pp. 114–117.

[27] M. Wemersson, S. Wanstedt, and P. Synnergren, "Effects of qos scheduling strategies on performance of mixed services over lte," in *IEEE PIMRC*, 2007, pp. 1–5.

[28] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, "Multi-qos-aware fair scheduling for lte," in *IEEE VTC-Spring*, 2011, pp. 1–5.

[29] K. hao Liu, L. Cai, X. Shen, and S. Member, "Multi-class utility-based scheduling for uwb networks," *IEEE Trans. Veh. Technol.*, 2008.

[30] X. Wang, G. Giannakis, and A. Marques, "A unified approach to qos-guaranteed scheduling for. channel-adaptive wireless network," *IEEE Proc.*, vol. 95, pp. 2410–2431, 2007.

[31] H. Lei, M. Yu, A. Zhao, Y. Chang, and D. Yang, "Adaptive connection admission control algorithm for lte systems," in *IEEE VTC-Spring*, 2008, pp. 2336–2340.

[32] H.-L. Chao, C.-K. Chang, and C.-L. Liu, "A novel channel-aware frequency-domain scheduling in lte uplink," in *IEEE WCNC*, 2013, pp. 917–922.

[33] I. Wong, O. Oteri, and W. Mccoy, "Optimal resource allocation in uplink sc-fdma systems," *IEEE Trans. Wirel. Commun.*, vol. 8, no. 5, pp. 2161–2165, 2009.

[34] F. Liu, X. She, L. Chen, and H. Otsuka, "Improved recursive maximum expansion scheduling algorithms for uplink single carrier fdma system," in *IEEE VTC-Spring*, 2010, pp. 1–5.

[35] X. Wang and S. Konishi, "Optimization formulation of packet scheduling problem in lte uplink," in *IEEE VTC-Spring*, 2010, pp. 1–5.

[36] H. Safa, W. El-Hajj, and K. Tohme, "A qos-aware uplink scheduling paradigm for lte networks," in *IEEE 27th International Conference on Advanced Information Networking and Applications*, ser. AINA '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 1097–1104. [Online]. Available: http://dx.doi.org/10.1109/AINA.2013.38

[37] S. Kwon and N.-H. Lee, "Uplink qos scheduling for lte system," in *IEEE VTC-Spring*, 2011, pp. 1–5.

[38] H. Ye, G. Lim, L. Cimini, and Z. Tan, "Energy-efficient resource allocation in uplink OFDMA systems under QoS constraints," in *IEEE MILCOM*, Nov 2013, pp. 424–428.

[39] S. N. K. Marwat, Y. Zaki, C. Görg, T. Weerawardane, and A. Timm-Giel, "Design and performance analysis of bandwidth and QoS aware LTE uplink scheduler in heterogeneous traffic environment." in *IEEE IWCMC*. IEEE, 2012, pp. 499–504.

[40] B. Al-Manthari, H. Hassanein, N. A. Ali, and N. Nasser, "Fair class-based downlink scheduling with revenue considerations in next generation broadband wireless access systems," *IEEE Trans. Mobile Computing*, vol. 8, pp. 721–734, 2009.

[41] T. S. 3GPP, "Group services and system aspects - policy and charging control architecture," p. 23, Nov 2009.

[42] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 2003.

[43] E. Hossain, L. B. Le, and D. Niyato, *Radio Resource Management in Multi-Tier Cellular Wireless Networks*. Amazon Digital Services, Inc.: Wiley; 1 edition, 2013.

[44] R. R and V. Leung, "Uplink scheduling solution for enhancing throughput and fairness in relayed LTE networks," *Accepted for IET Commun.*, 2013.

[45] L. HUAWEI TECHNOLOGIES CO., "eran scheduling feature parameter description," 2013. [Online]. Available: http://www.huawei.com

[46] M. Mojtahed and S. Xirasagar, "Quality of service over lte networks," 2013. [Online]. Available: www.lsi.com

[47] G. T. . v0.1.1, "Further advancements for e-utra, physical layer aspects," p. 23, Nov 2008.

[48] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Trans. Commun.*, vol. 47, no. 6, pp. 884–895, 1999.

[49] V. Garg, "An fdd multihop cellular network for 3GPP-LTE," June 2007.

[50] K. YONG-SEOK, "Capacity of voip over hsdpa with frame bundling," *IEICE Trans. Commun.*, vol. E89-B, pp. 3450–3453, 2006.

[51] J. Puttonen, T. Henttonen, N. Kolehmainen, K. Aschan, M. Moisio, and P. Kela, "Voice-over-ip performance in utra long term evolution downlink," in *IEEE VTC-Spring*, 2008, pp. 2502–2506.

[52] J. Blumenstein, J. C. Ikuno, J. Prokopec, and M. Rupp, "Simulating the long term evolution uplink physical layer," in *53rd International Symposium ELMAR*, Zadar, Croatia, September 2011.
[53] J. R., C. D.M., and H. W, "A quantitative measure of fairness and discrimination for resource allocation in shared systems," *DEC Research Report TR-301*, 1984.