Packet Scheduling Architecture with Service Specific Queue Sorting and Adaptive Time Domain Scheduling Algorithms for LTE-Advanced Networks

Rehana. Kausar, Yue. Chen, Kok. Keong. Chai, John Schormans
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, UK
rehana.kausar,yue.chen,michael.chai@elec.qmul.ac.uk

***Abstract*— In this paper, a cross layer design packet scheduling architecture is proposed for Long Term Evolution-Advanced downlink transmission, to guarantee the support of quality of service requirements in a mixed traffic environment. The proposed architecture uses service specific queue sorting algorithms for different traffic types and an adaptive time domain scheduling algorithm to adaptively allocate available resources to real time and non real time traffic. Multiuser diversity is exploited both in the time domain and frequency domain by jointly considering the channel state information and queue state information. The aim is to improve the support of QoS guarantees to real time voice and non real time streaming video traffic and to maintain a good trade-off between system throughput and user fairness by optimizing the use of available radio resources. Results show that proposed packet scheduling architecture reduces delay, delay viability and packet drop rate of real time traffic while satisfying minimum throughput requirements of non real time traffic and it maintains the system throughput and fairness among users at good level.**

Keywords-LTE-A; Packet Scheduling (PS); OFDMA; Quality of Service (QoS); mixed traffic.

I. INTRODUCTION

Long Term Evolution Advanced (LTE-A) is an all-IP (Internet Protocol) based future wireless communication network, which is aiming to support a wide variety of applications and services with different Quality of Service (QoS) requirements. It is targeting at superior performance in terms of spectral efficiency, fairness, QoS support and service satisfaction as compared to the existing Third Generation Partnership Project (3GPP) wireless networks.

To achieve the goal, Radio Resource Management (RRM) plays a vital role. Packet Scheduling (PS) being one of the core functionalities in RRM is very crucial to optimise the network performance and it has been under extensive research in recent years. Different PS algorithms have been deployed aiming at utilising the scarce radio resource efficiently. A QoS aware Packet Scheduling Architecture (PSA) is presented in [1], which takes into account different prioritizing stages such as QoS aware queue sorting and adaptive Time Domain (TD) scheduler with built-in congestion control to the existing conventional QoS aware PS algorithms. Delay dependent queue sorting algorithm for Real Time (RT) traffic reduces average delay of RT traffic and built-in congestion control policies reduce Packet Drop Rate (PDR) by adaptively allocating radio resources to RT and Non Real Time (NRT) traffic types based on QoS feedback of RT traffic. And by exploiting multiuser diversity in the TD and Frequency Domain (FD), system overall throughout is improved. By prioritising users with longer delays in RT and NRT streaming video traffic and using conventional Proportional Fairness (PF) algorithm to sort users in RT, NRT and Best Effort (BE) queues respectively, the proposed PSA in [1] maintains a good trade-off between system throughput and user-fairness. However, there is still need of further work on PSA [1] in order to meet the requirements of QoS for RT traffic and throughput requirements of NRT traffic. Service specific queue sorting algorithms are needed for each queue to guarantee the QoS support. In addition, the fix built-in congestion control policies used in [1] need to be replaced with an adaptive scheme to make the PSA capable to adapt to the network conditions, traffic patterns, system load and the QoS requirements of different traffic types.

Thus, the functionalities of queue sorting and adaptive TD scheduler are enhanced by extended research on these algorithms. New queue sorting algorithms for RT and NRT queues have been proposed to further improve the support of the provision of QoS guarantees for both RT and NRT traffics [2]. The results show that the queue sorting algorithms have reduced average delay, delay viability and PDR of RT traffic while satisfying the minimum throughput requirements of NRT streaming video traffic at the cost of minor delays in the BE traffic. It also shows that system overall throughput and user fairness are maintained at good level. To emphasise the significance of new queue sorting algorithms, the adaptive TD scheduler with built-in policies as in [1] was not used instead the users were picked from the queues one-by-one from each queue starting from the top most queue by simple fair scheduling method.

In [1], the λ denotes the proportion of available Physical Resource Blocks (PRBs) assigned to RT users and $(1-\lambda)C$ to NRT users where C is the total number of PRBs available. The initial value of λ is decided based on the trade-off between the average delay of RT and NRT traffic as shown in Fig. 1. Then the value of λ is adaptively adjusted according to the PDR of RT traffic using built-in congestion control policies. In [1] however, only the average delay of RT and NRT traffic is considered to set an initial value of λ which is not very realistic as other performance metrics such as throughput of NRT traffic, system throughput and fairness

among users should also be taken into account while setting this value. An extensive research has been done to adjust the initial value of λ so that a stability region can be found which takes into account various performance measures such as average delay of RT and NRT traffic, minimum throughput of NRT traffic and overall system throughput and fairness among users instead of only considering trade-off between RT and NRT average delay. After setting the initio value of λ, a new adaptive TD scheduling algorithm is used to make adaptive TD scheduler capable of controlling PDR at all traffic patterns instead of using a fix traffic pattern with only a number of built-in policies as in [1]. The results in [1] only consider a traffic pattern in which RT and NRT users are equal which is should be analysed by considering variable number of RT and NRT users as the number of RT and NRT users may vary with time. That is why the behaviour of the proposed PSA is analysed under different traffic patterns with varying number of RT and NRT users.

In this paper, the proposed PSA with new queue sorting algorithms [2] and novel adaptive TD scheduler algorithm is presented to enhance PS performance both at service level and network level. At the service level, the QoS of RT and NRT streaming video traffic are significantly improved as compared to the existing PS algorithms. At the system level, overall system throughput performance and fairness among all users are improved in the new PSA. As described above, this work is based on [1] that was presented in UBICOMM 2010.

The remainder of this paper is organized as follows. In Section II, the related work on PS algorithms is discussed. System model is presented in Section III and the proposed PSA with new queue sorting and adaptive TD algorithm is described in Section IV. In Section V, the proposed packet scheduling algorithm and performance metrics to analyse the proposed packet scheduling algorithm, are presented. The results and discussion section (Section VI) presents analytical results from different perspectives to show the performance of the proposed PSA; the first part of Section 5 presents a set of results to compare the performance of PSA with existing QoS aware PS algorithm and the second part evaluates the performance of PSA under different traffic patterns. Finally, conclusion and future work are presented in Section VII.

II. RELATED WORK

The classic packet scheduling algorithms exploiting multiuser diversity are the MAX C/I and Proportional Fairness (PF) algorithms. MAX C/I algorithm allocates a physical resource block (PRB) to a user with the highest channel gain on that PRB, and can maximize the system throughput [1] [3-4]. PF algorithm takes fairness among users into consideration and allocates resources to users based on the ratio of their instantaneous throughput and its acquired time averaged throughput [1] [5]. However, these algorithms aim only at improving resource utilization based on channel conditions of users; QoS requirements, for example delay requirements of real time (RT) traffic or minimum throughput requirements of non-real time (NRT) traffic, are not considered at all. In the next generation of mobile communication networks, apart from system throughput and user fairness, the crucial point is to fulfil users' QoS requirements in a multi-service, multi-user mixed traffic environment. This is because different traffic types are competing for radio resources to fulfil their QoS requirements. To allocate radio resources efficiently and intelligently in such complex environments is challenging. Various methods have been proposed aiming to use radio resources efficiently to fulfil QoS requirement of different traffic types [6-8].

A low complexity QoS aware PF multicarrier algorithm is presented for OFDM system in [9]. The objective is to achieve proportional fairness in the system while improving QoS performance. A greedy method based multi carrier PF criterion is proposed with the consideration that traditional single carrier PF is not suitable for OFDM systems. A subcarrier reassignment procedure is used to further improve QoS performance. This paper proposes PS algorithm specifically for the multimedia traffic and improves QoS, throughput and fairness in the system. However, there is a need to analyze the behaviour of the proposed algorithm when the system has to deal with different traffic types such as interactive, background traffic, etc. In [6], a service classification scheme is used which classifies mixed traffic into different service specific queues and grants different scheduling priorities to them. QoS of RT traffic is improved at the cost of system spectral efficiency, when the RT queue is granted the highest priority. And fairness is significantly improved when fair scheduling is used in the TD to pick users from the queues instead of strictly prioritizing RT traffic queue. Fair scheduling picks users one-by-one from each queue and strict priority empties queues one after other giving priority to RT queue. Conventional PF and MAX C/I are used to prioritise users in the queues. The QoS of RT and NRT traffic can be improved by using service specific queue sorting algorithms to prioritise users. In [10], an urgency factor is used to boost the priority of a particular traffic type. When any packet from a queue is about to exceed its upper bound of delay requirement, its priority is increased by adding an urgency factor. Although most of the packets are sent when they are nearly ready to expire, a lower packet loss rate is achieved thus improving the performance of system by guaranteeing QoS requirements to different traffic types.

In mix traffic scenarios, queue state information (QSI) becomes very important in addition to channel state information (CSI) [11-12]. It can make scheduling decision even more efficient; especially in QoS aware scheduling algorithms it is very crucial. Typically this implies to minimize the amount of resources needed per user and thus allows for as many users as possible in the system, while still satisfying whatever quality of service requirements that may exist [13]. A time domain multiplexing (TDM) system based Modified Largest Waited Delay First (M-LWDF) is presented in [11] which takes into account both QSI and CSI. This algorithm serves a user with the maximum product of Head of Line (HOL) packet delay, channel condition and an arbitrary positive constant. This constant is used to control packet delay distribution for different users.

It updates the queue state after each TTI rather than updating after each sub carrier allocation. M-LWDF significantly improves the support of QoS guarantees to the RT and NRT traffic for TDM systems. In [14], an exponential (EXP) rule is proposed for scheduling multiple flows that share a time-varying channel. The EXP rule is applied in M-LWDF as one of the parameters that equalizes the delays of different RT packets to reduce the PDR of RT traffic due to time-out. M-LWDF algorithm is applied in a frequency domain multiplexing (FDM) system in [15] to optimize sub-carrier allocation in Orthogonal Frequency Division Multiple Access (OFDMA) based networks. It shows improved performance in terms of QoS but like M-LWDF updates the queues state each TTI rather than after each sub-carrier allocation. In [16], M-LWDF for OFDMA systems is modified by updating the queue status after every sub-carrier allocation. It takes into account RT and NRT traffic types and provides better QoS for both services. The results show that the support of provision of QoS guarantees in terms of delay and PDR for RT and minimum throughput for NRT traffic is improved. However this idea can be extended to more effective scheduling framework by adding more traffic types and making resource allocation more adaptive based on the QoS. In [17] an adaptive algorithm with connection admission control (CAC) design is proposed. Due to large number of users and limited PRBs, CAC restricts the ongoing connections to provide required QoS and makes decisions whether to reject or accept new connections. It improves the QoS of RT traffic by prioritizing RT users and delaying users of other traffic types. In [18], a prioritizing function is used for packet data scheduling in OFDMA systems to satisfy QoS requirements of RT and NRT traffics. Priority is associated to different traffic types by setting different values of the prioritizing function. This algorithm allocates resources in a static way by setting the value of priority function for different traffic types and cannot cope with the highly dynamic variation of wireless channel conditions. In [19], a server allocation scheme to parallel queues with randomly varying connectivity is presented. The allocation decision is based on the connectivity and on the lengths of the connected queues only. The main aim of the work presented in [19] is to stabilize different queues. However this allocation policy can minimize the delay and maximize throughput for the special case of symmetric queues i.e., queues with equal arrival, service, and connectivity. However the work proposed by the author aims at considering system level and service level PS performance jointly. That is why various parameters are considered instead of only taking into account the stability of user queues, as in [19]. It takes scheduling decisions based on channel conditions to increase system spectral efficiency, average PDR to reduce PDR and delay viability and queue length to reduce packet delay and make the user queues stable.

As described in [11-18] and certain of the references therein, the PS algorithms improve scheduling performance in different domains separately such as system throughput, user fairness, QoS of RT and NRT traffic types. The Combined consideration of service level (QoS) and system level performance (system throughput and user fairness) improvement has got very little or no attention despite the fact that it is very crucial. Scheduling performance in different domains needs to be united in an efficient PS architecture so that the system can be made cost effective and radio resources may be utilised at the best. PS performance in different areas can be improved jointly by an intelligent PSA which is capable to make scheduling decisions adaptive to the environment and to the achieved performance in terms of QoS of different traffic types. The detailed traffic types can be considered in PSA to make the PS algorithms more realistic.

III. SYSTEM MODEL

An OFDMA system is considered in which minimum allocation unit is one Physical Resource Block (PRB) containing 12 sub-carriers in each Transmission Time Interval (TTI) of 1ms duration. There are K mobile users and M PRBs. The downlink channel is a fading channel within each scheduling drop. The received symbol $Y_{k,m}(t)$ at the mobile user k on sub channel m is the sum of the additive white Gaussian noise (AWGN) and the product of actual data and channel gain, as given in (5) [10-11].

$$Y_{k,m}(t) = H_{k,m}(t)X_{k,m}(t) + Z_{k,m}(t) \qquad (1)$$

where, $Y_{k,m}(t)$ is data symbol from eNodeB to user k at sub channel m, $X_{k,m}(t)$ is the input, $\left[H_{k,m}(t)\right]^2$ is the complex channel gain of sub channel m for user k, and $Z_{k,m}(t)$ is the complex White Gaussian Noise [11]. It is assumed, as in [11] [14-17], that the power allocation is uniform, $P_m(t) = P/M$ on all sub channels where, P is the total transmit power of eNodeB, $P_m(t)$is the power allocated at channel m and M is total number of sub channels. At the start of each scheduling drop, the channel state information (CSI) $H_{k,m}(t)$ is known by the eNodeB.

The achievable throughput of a user k on sub channel m can be calculated by (6) as used in [11] and [12].

$$C_{k,m}(t) = Blog_2\left[1 + \frac{|H_{k,m}(t)|^2}{\sigma^2\Gamma}P_m(t)\right] \qquad (2)$$

where, B is the bandwidth of each PRB, σ^2 is the noise power density i.e., noise power per unit bandwidth and Γ is a constant signal-to-noise ratio (SNR) gap and has a simple relationship with the required Bit Error Rate (BER).

$$\Gamma = \frac{-ln(5BER)}{1.5} \qquad (3)$$

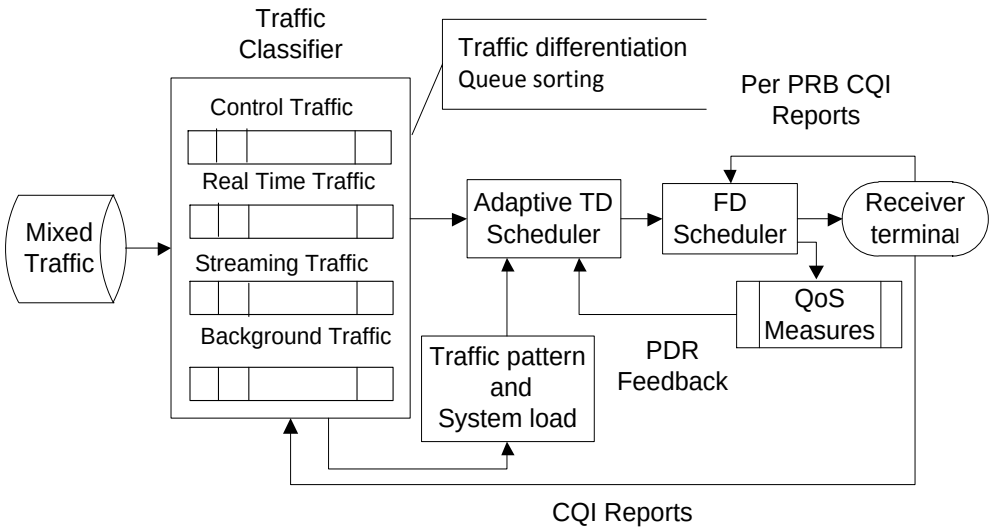

Figure 1. The cross layer packet scheduling architecture

IV. PACKET SCHEDULING ARCHITECTURE (PSA)

A schematic diagram of proposed PSA is shown in Fig. 1. It consists of a traffic classifier, adaptive TD scheduler and FD scheduler. Mixed traffic is classified into service specific queues at classifier stage. Users in these queues are prioritized according to QoS requirements. Adaptive TD scheduler adaptively allocates available radio resources to RT and NRT traffic types based on traffic pattern and system load information from traffic classifier and PDR information from QoS measure unit. QoS measure unit calculates PDR of RT traffic and minimum throughput of NRT traffic in each TTI to analyze the support of QoS provision to RT and NRT streaming video traffic. FD scheduler actually maps these resources to the selected users.

The detailed description of functionality, algorithms and policies of each proposed PSA stage are described as below.

A. Classifier

The need for traffic differentiation arises when there is a question to deal with mixed traffic demanding different QoS guarantees. In such an environment, it becomes very important to classify traffic into different service queues to enable queue specific prioritizing schemes effectively. Service differentiation is the first step towards optimising the utilization of available radio resources where the available radio resources are allocated according to the well-defined demands of traffic types [1].

In the proposed PSA mixed traffic is classified into four queues; Control, Real Time, Non Real Time and background traffic queue, as in [1]. These queues are represented by control, RT, NRT and BE queue hereafter. The queues at the traffic classifier stage are prioritized in the sequence as discussed above. These classes cover most of the common traffic types such as control information, low latency RT conversational, high throughput NRT streaming video and low priority background data. Control information is the signaling information exchanged between the User Equipment (UE) and eNodeB and it is separated from other data queues and is served before any other data queues. The control queue is always allocated enough radio resources to transmit signaling information to users. Background traffic represents the best effort (BE) class of traffic and does not have any QoS requirements. The service specific queue sorting algorithms used to prioritise users in these queues are as follows.

Control queue

In the proposed classifier, the control information is equally important information between users and therefore it is transmitted in Round Robin (RR) manner for all scheduled users.

RT queue

In RT queue, the delay requirement for each RT user is defined as $d_k < DB_k$ where d_k is the delay of user k, DB_{RT} the delay budget which is the upper delay bound of RT traffic. A delay dependent priority metric is used to sort users in the RT queue. The priority metric is shown in (4). It is the product of normalised waiting time of each RT user and its channel state information, and the product is added with the square of the user's queue length. In this priority metric, users with longer waiting time (normalised by DB), good channel conditions and longer queues, are prioritised in the front of the queue. By prioritising users with longer delays (normalised with DB), the priority metric reduces

average delay of RT traffic significantly. In addition users are given equal opportunity to be scheduled thus improving fairness among users. By giving priority to users with longer queues, this priority metric reduces PDR of RT traffic due to time out. This is because in a user's queues, packet with the longest delay (provided it is not timed out) is transmitted first provided a PRB is allocated to this user. The overall system throughput is improved by exploiting multiuser diversity when users with good channel conditions are prioritised over the users with bad channel conditions.

The priority of an RT user k at time t is given by (4) below [2].

$$P_k^{RT}(t) = \left(\frac{T_k^{waiting}}{DB^{RT}}(t) \times [H_k^{RT}(t)]^2\right) + [Q_k(t)]^2 \qquad (4)$$

where, , $P_k^{RT}(t)$ is the priority of RT user k at time t, $T_k^{waiting}$ is the waiting time of RT user k, DB^{RT} is the delay budget of RT traffic, $H_{k\epsilon K}^{RT}$ is the channel state information of RT user k and $Q_k(t)$ is queue length of user k at time t.

NRT queue

The QoS requirement for NRT streaming video traffic is defined as $r_k(t) \geq T_k$, where $r_k(t)$ is the instantaneous throughput of user k at time t and T_k is throughput requirement of NRT user k. A QoS aware priority metric is used to sort users in NRT queue. The priority metric for NRT queue is shown in (5) It is the product of normalised waiting time, a ratio of minimum required throughput and average achieved throughput, and channel state information, of each NRT user. The priority metric reduces delay of NRT queue users by prioritising users with longer delays and improves fairness among users by allocating them fair share of time, to be scheduled. This is because when users with longer delays are put in the front of queue, then at the end users' total number of scheduling intervals become almost equal. the ratio of minimum required throughput and average achieved throughput increases the priority of users achieving low throughput and tries to allocate to each user equal or more than the minimum throughput required by NRT queue users. Multiplication of channel state information helps improving the overall system throughput by prioritising users with good channel conditions as in (4).

The priority of a NRT user k at time t is given in (5) below [2].

$$P_k^{NRT}(t) = \frac{T_k^{waiting}}{DB^{NRT}} \times \frac{T_k(t)}{R_k(t)} \times [H_k^{NRT}(t)]^2 \qquad (5)$$

where $P_k^{NRT}(t)$ is the priority of NRT user k at time t, $H_k^{NRT}(t)$is the channel state information of NRT user k at time t and DB^{NRT} is the delay budget of NRT streaming video traffic.

The time average throughput of user k, $R_k(t)$ is updated by the following formula as used in [1] [9],

$$R_k(t+1) = \left[1 - \frac{1}{t_c}\right] R_k(t) + \frac{1}{t_c} \sum_{m=1}^{M} r_{k,m}^{'}(t) \qquad (6)$$

where t_c is the length of time window to calculate the average data rate; $1/t_c$ is called attenuation co-efficient with classic value 0.001, $r_{k,m}^{'}(t)$ is the acquired data rate of user k at PRB m if m is allocated to k, else it is zero and r_k is instantaneous and $R_k$is average throughput of user k.

BE queue

BE traffic has no QoS requirements so priority is given to users based only on channel conditions. However to maintain some amount of fairness between users, classic PF algorithm is used as the queue sorting algorithm for BE. The priority metric for BE users is given below in (10),

$$P_k^{BE}(t) = \frac{r_k}{R_k} \qquad (10)$$

where $P_k^{BE}(t)$ is the priority of BE user k at time t.

B. Adaptive TD scheduler

After prioritising users in the queues, adaptive TD scheduler selects the most suitable users from the queues based on the priority of traffic types and the available PRB in the FD.

Packet scheduling algorithm is mainly focused on PRB allocation based on users' channel state information, traffic queue information and QoS requirements. However because of too many users and limited PRBs, it is infeasible to guarantee all ongoing users' QoS in each TTI. In this case, a TD scheduling algorithm is needed to make decisions adaptively whether to admit or reject scheduling request of a user. A novel Adaptive TD scheduling algorithm is proposed in this paper, where it chooses a pool of users from the queues of traffic classifier based on current network conditions and PDR feedback of RT traffic. This algorithm consists of two main steps. In the first step, it allocates the radio resources to RT and NRT traffics based on current traffic pattern, system load and service, and system level performance metrics. In the second step, it adjusts the RT resource allocation at the cost of minor delay in NRT traffic. This algorithm lowers the PDR of RT traffic and at the same time ensures that the minimum throughput requirement of NRT traffic is met. It is achieved by decreasing resource allocation to RT queue and allocating resources to NRT traffic queues when the PDR is lower than the threshold.

The adaptive TD scheduling algorithm works as follows.

Let the total number of available PRBs are denoted by C. Let λC be the proportion of PRBs assigned to RT traffic users and $(1-\lambda)C$ is assigned to NRT traffic users. At the first step the default value of λ is set from the built-in

policies based on different parameters then the value of λ is adaptively adjusted according to PDR of RT traffic. A built-in policy defines the resources reserved for RT and NRT traffics e.g., policy (60%, 40%) means that 60% of the available PRBs are reserved for RT traffic and 40% are reserved for NRT traffic. At the start of transmission, TD adaptive scheduling algorithm uses a default policy to distribute PRBs between RT and NRT traffic types. A default policy is set at a point where the PS algorithm performs well in terms of all performance metrics thus improve the PS performance at service level and system level. This is defined as a stability region at which the PS algorithm produces balanced performance regarding all performance metrics. The default value of λ is adjusting with the current network condition. This is because the network conditions are changing rapidly in wireless environments. In this way the main challenanges in setting the default value of λ are; i) finding a stability region and ii) updating the value of λ based on changing wireless conditions. To find a stability region is subject the practical user distribution and total number of active users. It means that there may be different traffic patterns such as RT users are equal to NRT user or RT users may be lesser or more than NRT users. Similarly, the number of active users can vary with time. The default value of λ can have different values under different traffic patterns and varying system load. To set a stable default value of λ under different traffic patterns and with variable system load, a series of experiments have been done as described below.

Results Analysis of Built-in Policy

In this section an analysis is presented based on a series of simulation results which is done to make the PS algorithm work effectively under different traffic patterns and with variable system load. For this analysis, the QoS of RT traffic (delay, PDR), QoS of NRT traffic (minimum throughput), system throughput, user throughput fairness and a trade-off between system throughput and user fairness are analysed at the system load varying from 40 to 100 active users in a single cell scenario. These simulations have been conducted in the following traffic patterns.

- RT users $=$ $NRT\ users$
- RT users $>$ $NRT\ users$
- RT users $<$ NRT users

For each traffic pattern, simulations are run for network loads varying from 50 to 100. The reason of running these simulations is that PS performance behaves differently under different traffic patterns and different system load, and there is a need of finding out a stability region where PSA can produce balanced performance in terms of all PS performance metrics used in this paper. Fig.2 shows an example how these simulations are run. In Fig.2 the average delay of 80 users is calculated by using different built-in policies. This is to find a policy where the average delays of RT and NRT traffic are balanced. As shown, both RT and NRT traffic have a balanced delay at policy (70%, 30%). Trade-off between RT and NRT delay shows an insight how the simulations are run for analysis purpose.

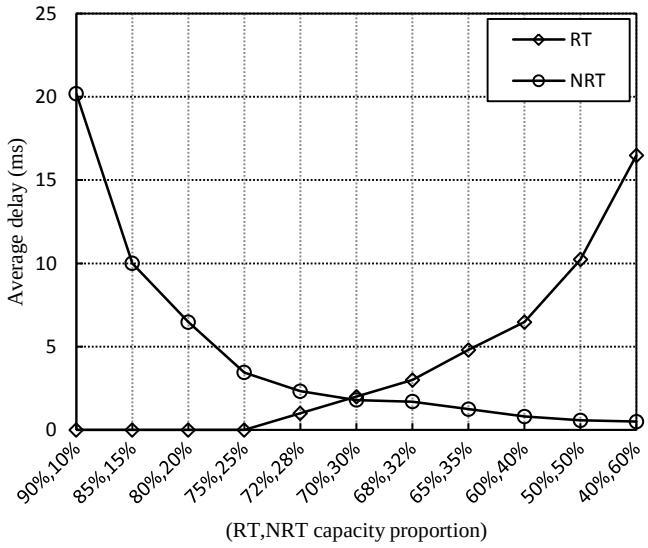

Figure 2. A trade-off between delay of RT and NRT traffic.

This trade-off value is different under different network loads for the same traffic pattern. And this trade-off value is different under different traffic patterns with the same system load. This difference appears for other performance measures such as in PDR, minimum throughput, and fairness etc. For all performance metrics, a balanced point is traced out by considering QoS requirements of RT and NRT traffic types. Based on all information, a stability region is analysed to set the default value of λ. The conclusion of the series of all experiments is as follows.

For the first traffic pattern (RT users > NRT users), if the total number of active users in the cell are more than 60, the built-in policy (70%, 30%) works well in terms of QoS of RT and NRT traffics and system throughput and user fairness. This is shown by analysis results that the system can work well at this policy for all traffic loads greater than 60. If the number of active users is lesser than 60, then policy (60%, 40%) works well and comes up with required performance guarantees. For second network condition (RT users = NRT users), policy (50%, 50%) works well for all network loads. For third network scenario (RT users < NRT users), if the number of users is greater than 60, policy (30%, 70%) works well and for a load lesser than 60, policy (20%, 80%) works well. In this way this algorithm reserves radio resources for RT and NRT traffic types based on a stability region where the proposed algorithm shows a balanced performance under variable system load and specific traffic pattern, in terms of all aimed performance metrics. The next step is to further improve QoS of RT and NRT streaming video traffic by making adaptive changed in the value of λ based on PDR of RT traffic. If the PDR of RT

traffic is increased above a certain threshold, RT resource allocation is increased at the cost of minor delay in NRT traffic. And if PDR of RT traffic is lower than the threshold, RT resource allocation is decreased by diverting resources to NRT traffic types.

The adaptive change in the value of λ for the second step of adaptive TD scheduler follows the following rule (11).

$$\lambda(t+1) = \begin{cases} \lambda(t) & if\ PDR_{RT}(t) = \varphi \\ \lambda(t) + \eta & if\ PDR_{RT}(t) > \varphi \\ \lambda(t) - \eta & if\ PDR_{RT}(t) < \varphi \end{cases} \quad (11)$$

where η is the increment/decrement of the resources reserved for RT traffic and φ is the PDR threshold set for RT traffic. Packets of RT users are dropped when they exceed upper bound of delay. PDR is calculated by QoS measure unit of the proposed PSA in each TTI and is fed back to the adaptive TD scheduler. And based on PDR value adaptive TD scheduler adaptively increment or decrement resources reserved for RT and NRT traffic.

If PDR of RT traffic is equal to φ, value of λ will not change. However if PDR is higher than φ, there will be an increment equal to η in the resource allocation to RT traffic, and if PDR is lower than φ, the resource allocation to RT traffic will be decremented by the same amount η.

This algorithm lowers the delay, delay viability and PDR of RT traffic and considers the minimum throughput requirements of NRT traffic to be satisfied at the same time. This is achieved by increasing NRT resource allocation when PDR is under the threshold.

C. FD scheduler

In the frequency domain, PRBs are mapped to the users. Multiuser diversity is exploited by using channel dependent frequency domain proportional fairness (PF-FD) algorithm. Per PRB CQI reports of each user are fed back to this stage and for each scheduling unit, the best PRB is selected and allocated to it.

V. THE PROPOSED PACKET SCHEDULING ALGORITHM AND PERMANCE METRICS

In this section the packet scheduling algorithm and the performance metrics for its performance evaluation are given.

A. Packet Scheduling Algorithm

A list of prioritized users is generated after applying queue sorting and adaptive TD scheduling algorithms at the classifier and adaptive TD scheduler stage of the PSA, respectively. The proposed PSA flow and the PRB allocation method is formalized in the following algorithm. At a given time *t*, PRBs are allocated to the prioritized users by this algorithm.

Algorithm: The packet scheduling algorithm.

1: Initialization: Let $k = 1,2,3,\ldots..K$, Q_{RT} = RT users, Q_{NRT} = NRT users, Q_{BE} = BE users, subject to $\{Q_{RT}, Q_{NRT}, Q_{BE}\} \subseteq K$,
Let $m = 1,2,3,\ldots..M$, and $Q_k(t) = total\ packets\ of\ user\ k$;
2: Calculate $r_{k,m}(t)$ for all $k \in K$ according to (2);
3: For every user $k \in Q_{RT}$, for very user $k \in Q_{NRT}$ and for every user $k \in Q_{BE}$,
Calculate priority according to (4), (5) and (6) respectively;
4: Sort users in each queue in descending order of the priority;
5: Set the default value of λ from built-in policies based on the stability region; RT capacity = λ;
6: Allocate $(1-\lambda)$ capacity to Q_{NRT} and Q_{BE} queue;
7: Select a set of users from the queues based on capacity allocation;
8: Assign a user k with the highest priority with a PRB m,
Subject to $m = arg\ \max_m (r_{k,m})$;
9: Update $r_k = r_k + r_{k,m}(t)$;
10: Update $M = M - \{m\}$ and $K = K - \{k\}$;
11: Update $Q_k(t) = Q_k(t) - \{Transmitted + dropped\}\ packets$;
12: If $Q_k(t) \leq 0$ then
Remove this user from user list K and allocate m to next user in the user set;
13: Go to step 8 if the PRB list is not empty else go to next TTI;
14: Update average achieved throughput $\overline{R_k(t+1)}$ for all users.

Figure 3. The proposed packet scheduling algorithm.

Resource allocation is completed when all PRBs are allocated.

B. Performance Metrics

We analyze the propose packet scheduling framework under performance metrics of system throughput, fairness among users and QoS of RT and NRT traffic types.

The system throughput is the sum of average throughput across all the users [20]. Individual user throughput helps calculating minimum throughput requirements of NRT users and system overall throughput is used to analyze network level PS performance in terms of system spectral efficiency.

To measure the fairness among users Raj Jain fairness index is adopted which is defined as below [20-21].

$$Fairness = \frac{\left[\sum_{k=1}^{K} \tilde{R}_k\right]^2}{K\sum_{k=1}^{K}\left(\tilde{R}_k\right)^2} \quad (12)$$

The value of fairness index is 1 for the highest fairness when all users have same throughput such as at lower system loads. In (12), K represents total number of users and $\tilde{R}_k$ is the time average throughput of user k.

User delay is equal to the number of TTIs in which the user is not scheduled and average delay of RT traffic is the total delay experienced by all RT users divided by the total

number of users. The PDR is calculated by the ratio of number of packets dropped (due to time out) to the total number of RT packets as given in (14) as used in [2] [16].

$$p_k^{PDR} = \frac{n_k^{dropped}}{n_k^{total}}. \quad (13)$$

Where p_k^{PDR} is the PDR, $n_k^{dropped}$ is total number of dropped packets by RT user k and n_k^{total} is total number of packets generated by RT user$k$. Overall PDR for RT traffic is calculated by taking the ratio of total packets dropped by all RT users to the total number of packets of RT users. And the delay violation probability is taken as the PDR of a user k with the maximum value of PDR out of all RT users as given by (15), as used in [2] [16].

$$Delay\ viability = \max_{k \in RT}(p_k^{PDR}) \quad (14)$$

Where p_k^{PDR} is the PDR of RT user k. The long-term minimum throughput is taken as the minimal throughput among all NRT streaming video traffic users and is given by (15) [2] [16],

$$r_{min} = \min_{k \in NRT} r_k. \quad (15)$$

Where r_{min} is the minimal throughput of all NRT streaming video traffic users and r_K is the throughput achieved by NRT streaming video traffic user k.

VI. SIMULATION RESULTS AND DISCUSSION

Simulation model used in all simulations is presented in this section. The results obtained are discussed in detail in this section.

A. Simulation model

The proposed PSA for LTE-A networks is simulated using a single cell OFDMA system with total system bandwidth of 10 MHz which is divided into 55 PRBs and PRB size is 180 kHz. The total system bandwidth is divided into 55 PRBs.

The wireless environment is typical Urban Non Line of Sight (NLOS) and the LTE-A system works with carrier frequency of 2GHz. The most suitable path loss model in this simulation is COST 231Walfisch-Ikegami (WI) [3] as used by many other literatures on LTE. In the simulation we assume all users are random distributed.

In the simulations, we take full buffer traffic model and packet is fixed to 180 bits/s. The first simulation is to compare the performance of the proposed PSA against the existing QoS aware PS algorithm. In these simulations the total number of RT users is equal to the total number of NRT users. The simulation results are shown in Figs. 4 to 9.

In the second set of simulation, the performances of the proposed PSA are analysed under different traffic patterns as mentioned in Section III. The second set of the simulation results compare the PS performances of the proposed PSA at different traffic patterns and variable system loads. The results are shown in Figs. 10 to 12.

The simulation parameters for the system level simulation are based on [22] and these values are used typically in most of the literatures. The simulation parameters and configurations are shown in Table 1.

TABLE 1 SIMULATION PARAMETERS

Parameter	Value/comment
Cell topology	Single cell
Cell Radius	1 km
UE distribution	Random
Smallest distance from UE to eNodeB/m	35 m
Path Loss model	COST 231 Walfisch-Ikegami (WI) model
Shadow fading standard deviation	8 dB
System bandwidth	10 MHz
PRB bandwidth	180 kHz
Carrier frequency	2 GHz
BS transmission power	46dBm(40w)
Traffic model	Full buffer

In each of the simulation, the delay upper bound for RT traffic is set to 40 ms [16] [23] which is equivalent to 40 time slots. The minimum throughput required by NRT streaming video traffic users is to 240 kbps as used in [2] and [16]. The total eNodeB transmission power and Bit error rate (BER) for all users are set to 46dBm (40w) and 10^{-4} respectively.

In [1], each user is assumed to have one service type and one scheduling unit (SU) carries the information about the user, service type and buffer status. However in this paper three separate traffic models are used for RT, NRT streaming video and BE traffic. For RT traffic, an "ON and OFF" traffic model is used with 35% "ON" time, and the packets are generated by using Poisson distribution. Poisson distribution is also employed for NRT streaming video and BE packet generation. The BER without buffering for RT traffic, NRT streaming and BE traffics are 0.1579, 0.8596 and 0.7448 respectively. For the BER with buffering of RT, NRT and BE, the values are 9.864e-007, 9.9219e-007 and 9.881e-007 respectively. Using the above simulation model, all simulations are run in Matlab R2009a on Windows 7 with 2.4 GHz CPU and 4-GB RAM.

B. Simulation results

The performance of the proposed PSA is evaluated against the standalone PF and QoS aware SWBS algorithm [15]. In the simulation result figures, PPSA represents the proposed PSA, PF represents the proportional fairness

algorithm and SWBS represents the QoS aware packet scheduling algorithm.

First, we present results for the QoS support to RT and RT streaming video traffic. The conventional PF algorithm does not take into account QoS support to RT and NRT streaming video traffic and is not considered in these results. Fig.4 shows the average delay of RT traffic by the proposed PSA and SWBS algorithm. Both PPSA and SWBS show almost same average delay for a system load $K < 70$ as the total number of active users is small and users are frequently scheduled. At higher system loads, the average delay shown by both algorithms increases because of resource competition. However the performance of SWBS is poorer than the proposed PPSA. As can be seen for $K = 100$, average delay of PPSA is 0.56 ms which is significantly lower than the average delay by SWBS. This is because the proposed PSA is designed in such a way that it reduces user delay by giving high priority to the users with longer delays. At lower system load, both algorithms show almost the same performance because the available resources are sufficient enough to meet the requirements of all users.

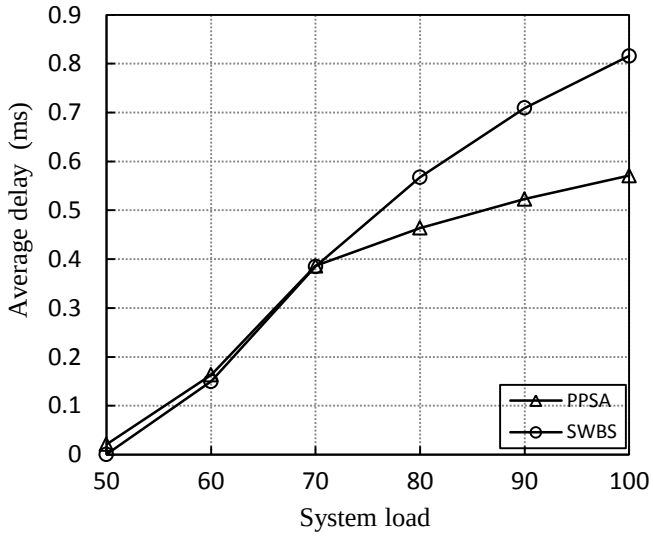

Figure 4. Average delay of RT traffic.

The PDR performance is analysed as the average PDR of RT traffic and the delay viability of RT users. The average PDR of RT traffic is calculated by (13) and is shown in Fig. 5. The performance shown by PPSA and SWBS is same for lower system loads when $K < 80$. When $K > 80$, the average PDR increases significantly with the user number. However PPSA can still maintains the best PDR performance. At $K = 100$, the PDR performance of PPSA is 30 % better than SWBS as shown. This is due to the particular design of adaptive TD scheduler in the proposed PSA as described in Section IV

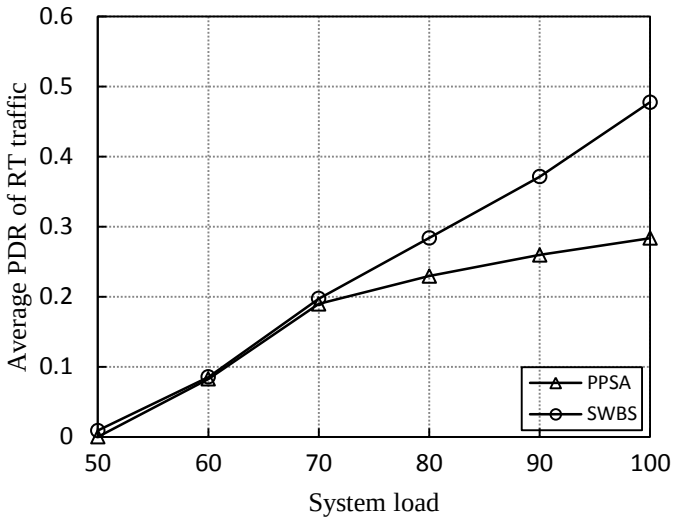

Figure 5. Average PDR of RT traffic.

The delay viability is a measure of difference of PDR among RT users and shows the highest PDR by an RT user. It is calculated by (14) and is shown in Fig.6. As can be seen, delay viability for both algorithms increases with the number of users. However PPSA shows higher performance as compared to SWBS particularly at higher system loads. The proposed PSA at its classifier stage, prioritises users with longer queues and transmits packets with the highest delay (provided it is not timed out), once PRB is allocated to the user. In this way it significantly reduces the number of dropped packets due to time out. Delay viability is further reduced by PSA when adaptive TD scheduling algorithm adaptively adjusts the radio resource allocation to RT traffic based on PDR threshold. That is why it has capability to keep the PDR of each user lower than the PDR shown by other algorithm.

The QoS support for NRT traffic is analysed by the minimum throughput of streaming video traffic as shown in Fig. 7. It is calculated by (5) for the proposed PSA, SWBS and conventional PF algorithm. The results for PF algorithm are included hereafter because it is designed to improve system throughput and fairness among users. While showing results on throughput of users and system throughput and fairness among users, PF shows comparable results.

The proposed PSA and SWBS can support minimum throughput guarantee of streaming video traffic and achieve more than required throughput($R_k = 240kbps$). However the conventional PF algorithm can only support minimum throughput guarantee at lower system load, $K = 50$. When $K > 50$, minimum throughput achieved by PF decreased and becomes 135 kbps at $K = 100$ as shown.

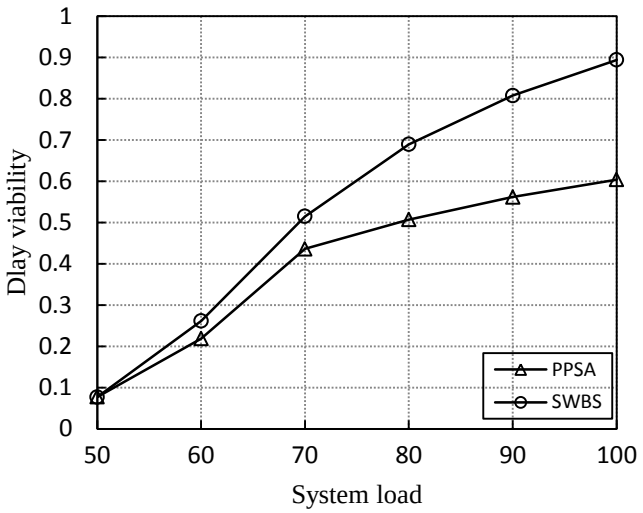

Figure 6. Delay delay viability of RT users.

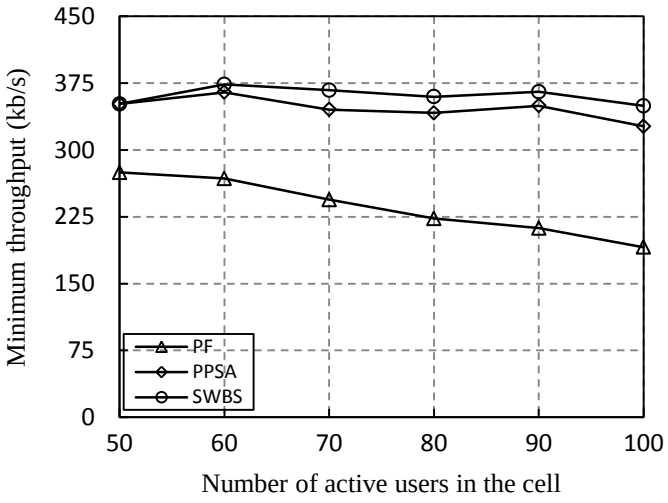

Figure 7. Minimum throughput of streaming video traffic.

The fairness performance is analysed by (12) for PF, proposed PPSA and SWBS algorithm and is shown in Fig.8. The proposed PSA significantly improves The PF algorithm shows the highest fairness among all algorithms at all system loads because it has a fairness control in its design. Fairness achieved by PPSA is almost similar to that achieved by PF up to a system load of 70 active users. However it is slightly lower than PF at system load higher than 70 active users. This is because at higher system loads, the resource competition among users increases significantly and PPSA is a QoS aware algorithm. It is designed to balance the PS performance in terms of all performance metrics. That is why at higher system load its fairness performance decreases slightly as shown. The SWBS algorithm however shows the lowest performance because in its design there it lacks fairness control.

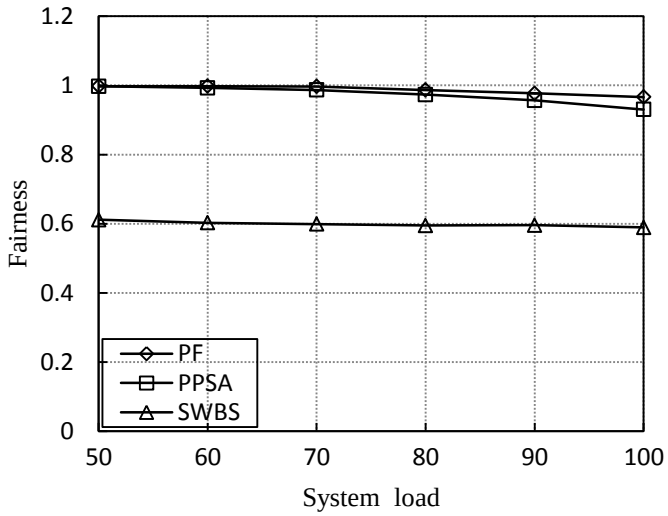

Figure 8. Fairness among users.

We define the average system throughput as the average transmitted bits per second in the system [9]. Fig.9 shows system throughput achieved by the proposed PSA, SWBS and conventional PF algorithm. As can be seen, the proposed PSA achieves the highest throughput among all the algorithms. PF algorithm also achieves a high throughput because it is designed to make a good trade-off between system throughput and user fairness thus maintain system throughput at good level. However its performance results are poorer than that for PSA at all system loads. For example at a system load of 70 active users, system throughput achieved by PPSA is 24Mbps which is 6Mbps higher than PF algorithm and 9 Mbps higher than SWBS algorithm. This is because the proposed PSA exploits multiuser diversity both in the TD and FD and always gives priority to users with good channel condition. The SWBS algorithm achieves the lowest system throughput because it is only designed to improve QoS of RT and NRT traffic types and does not improve overall system throughput.

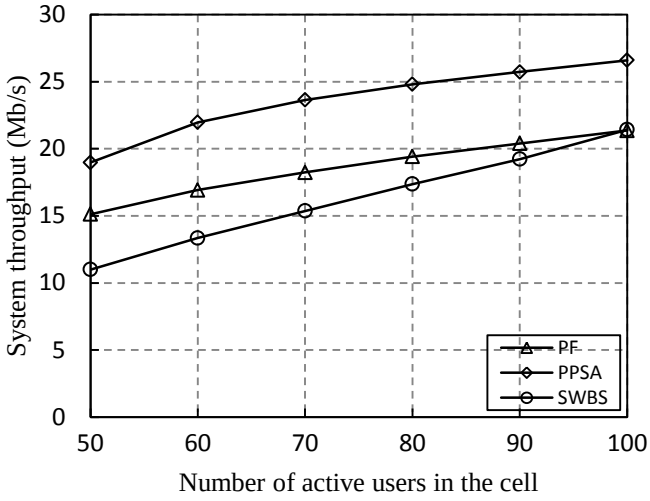

Figure 9. System throughput.

C. Performance results for different traffic patters

In this paper simulations have also been conducted to analyse the performance of the proposed PSA with three traffic patterns with number of users varying from 50 to 100. This is to prove the validity of the proposed PSA in varyig network condtions in terms of traffic patterns and system loads. In the first set of traffic pattern, the total number of active users are equally divided between RT and NRT traffic. The second set of traffic pattern consists of 70% RT active users and 30% of NRT active users. In the third set of traffic pattern, there are 30% of RT users and 70% of NRT users.

In these results the default value of λ is set according to the current traffic pattern and system load. This value is then adaptively adjusted based on PDR of RT traffic as discussed in Section IV.

In the previous set of results, the proposed PSAperformance is analysed on the service level by average delay, delay viability and PDR of RT traffic and minimum throughput of NRT traffic and on the system level by system throughput and fairness among users. In this set of simulation results, the performance results of the proposed PSA for varying network conditions are given. Thses results are shown in Figs. 10 to 12. As can be seen, there is not any huge different in the PS performance at the service and system level. And the proposed PSA is capable to maintain good performance under all varying network conditions . Average delay for RT traffic at $RT = NRT$ and $RT > NRT$ is almost equal, however its value at $RT < NRT$ is slightly lower. Delay viability at $RT = NRT$ and at $RT < NRT$ is almost same and its value for $RT > NRT$ shows very small difference at system load when $K < 60$. PDR with all traffic patterns is almost equal except at a system load when $K = 100$, where it shows verty slight difference. Thesupport for minimum throughput gaurantee to NRT streaming video traffic is well satisfied at all traffic patterns and all users achieve a throughput higher than the requirement $R_k > 240kbps$, as shown.System overall throughput value is almost same at all conditions and fairness at $RT < NRT$ is slightly lower than the other two network conditions

VII. Conclusion and Future Work

In this paper, we have presented a QoS aware packet scheduling architecture which is composed of three main units for the resource allocation in the downlink transmission of OFDMA-based LTE-A networks. The queue sorting algorithms at the classifier stage segregate mix traffic into service specific queues and prioritize users in these queues according to their QoS requirements. The novel adaptive TD scheduling algorithm sets a default value of radio resources for RT and NRT traffic based on traffic pattern and system load at first step. The default value is the changed adaptively based on PDR of RT traffic. In this way it helps maintaining good performance of the proposed PSA with variable conditions of traffic patterns, system load and PDR of RT traffic. In the FD the prioritized list of users is allocated PRBs in such a way that those users get the best PRB available. It helps improving the system spectral efficiency significantly. In this way the proposed PSA provides better QoS to different traffic types. It is able to improve system spectral efficiency by optimizing the use of given radio resources and maintains certain degree of fairness among users at the same time, by adaptively providing just enough resources to RT traffic and distributing extra resources efficiently to NRT services. The results show an improved QoS of RT traffic and a better trade-off between user fairness and system overall throughput. The performance comparison under different traffic patterns and with variable system loads also show that good performance of proposed PSA is maintained with variable conditions.

This work mainly focus on user-level PS performance by evaluating average delay and average PDR of RT traffic, delay viability of RT users and minimum throughput guarantees to NRT users. However packet-level PS performance may be evaluated by calculating jitter which is an importance performance metric at packet-level.

References

[1] Rehana. K., Yue. Chen, Kok. K. C., Laurie C., and John S., ”QoS aware mixed traffic packet scheduling in OFDMA based LTE-Advanced networks”, UBICOMM 2010, Copyright © IARIA 2010, pp. 53-58.

[2] Rehana. K., Yue. C., and Kok. K. C., ” Service Specific Queue Sorting and Scheduling Algorithm for OFDMA-Based LTE-Advanced Networks” Sixth International Conference on Broadband and Wireless Computing, Communication and Applications 2011, Barcelona, Spain [accepted].

[3] Harri H., and Antti T., “LTE for UMTS OFDMA and SC-FDMA Based Radio Access”, John Wiley and sons ltd 2009, pp 181-190.

[4] Martin S. (2008,April 23).WirelessMoves, 3GPP Moves on: LTE-Advaced. Last viewed 23 Jan. 2012 at 18:35 GMT. Website: hhttp://mobilesociety.typepad.com/mobile_life/2008/04/3gpp-moves-on-l.html

[5] Stefania S., Issam T., and Matthew B., "The UMTS Long Term Evolution Forum Theory to Practice", 2009 John Willey & Sons Ltd.ISBN: 978-0-470-69716-0.

[6] Jani P., Niko K.,. H., Martti M. and Mika R., "Mixed Traffic Packet Scheduling in UTRAN Long Term Evaluation Downlink" IEEE 2008, pp 978-982.

[7] Won-Hyoung P., Sunghyun C. and Saewoong B., "Scheduling design for multiple traffic classes in OFDMA networks", IEEE 2006, pp,790-795.

[8] Bilal S., Ritesh M., and Ashwin S., "Downlink Schedulinh for Multiclass Traffic in LTE", EURASIP Journal on Wireless Communications and Networking, Vol. 2009, Article ID 510617, 18 pages.

[9] Zhen K., Yu-Kwong, and Jianzhou W., "A low complexity QoS aware proportinal fair multicarrier scheduling algorithm for OFDM systems" vehiculer transaction on IEEE technology, June 2009, volume 58.

[10] Gutierrez I., Bader F., Pijoan J. L., "Prioritization function for packet scheduling in OFDMA systems", Wireless internet conference 08, Nov. 2008, Maui, USA.

[11] Andrews P., Kumaran K., Ramanan K., Stolyar A., Whiting P., Vijayakumar R., "Providing quality of service over a shared wireless link", Communication magazine, IEEE, vol.39, 2001, pp.150-154.

[12] Suleiman Y. Y., and Khalid A. B., "Dynamic buffer management for multimedia QoS beyond 3G wireless networks", IAENG International Journal of computer science, 36:4, IJCS_36_4_14, Nov. 2009.

[13] Toskala A., and Tiirola E., "UTRAN Long Term Evaluation in 3GPP," Proceedings of IEEE Personal Indoor and Mobile Radio Communications Conference (PIMRC'06), September 2006.

[14] Sanjay. S., and Alexander L. S, "Scheduling for Multiple Flows Sharing a Time-Varying Channel: The Exponential Rule" Bell Labs, Lucent Technologies, NJ 07974.

[15] Parimal P., Srikrshna B., and Aravind R., "A subcarrier allocation algorithm for OFDMA using buffer and channel state information", Vehicular Technology Conference, 2005. VTC-2005-Fall.2005 IEEE 62nd, pp.622-625.

[16] Jun S., Na Y., An L., and Haige X., "Opportunistic scheduling for heterogeneous services in downlink OFDMA system," School of EECS, Peking University, Beijing, P.R.China, IEEE computer Society 2009, pp.260-264.

[17] Haipeng L. E. I., Mingliang, A. Z., Yongyu C., and Dacheng Y., "Adaptive Connection Admission Control Algorithm for LTE Systems", IEEE 2008, pp. 2336-2340.

[18] Kian C.B., Simon A., Angela D., "Joint Time-Frequency Domain Proportional Fair Scheduler with HARQ for 3GPP LTE Systems", IEEE 2008.

[19] Leandros T., and Anthony E., "Dynamic Server Allocation to Parallel Queues with Randomly Varying Connectivity" IEEE Transaction on Information Theory, Vol. 39, N0. 2, March 1993, pp. 466-478.

[20] Lin X, and Laurie C. "Improving fairness in relay-based access networks," in ACM MSWIM, Nov.2008, pp. 18-22.

[21] Chisung B., and Dong-Ho C., "Fainess-Aware Adaptive Resource Allocation Scheme in Multihop OFDMA Systems," Communications Letters, IEEE, vol.11, pp. 134-136, Feb. 2007.

[22] 3GPP TSG-RAN, "TR25.814: Physical Layer Aspects for Evolved Utra". Version 7.0.0,June 2006.

[23] Ekstrom H., Furuskar A., Karlsson J., Meyer M., parkvall S., Torsner J., and Wahlqvist M., "Technical Solution for 3G LTE," IEEE Communications Magazine", vol. 44, March 2006, pp.38-45.

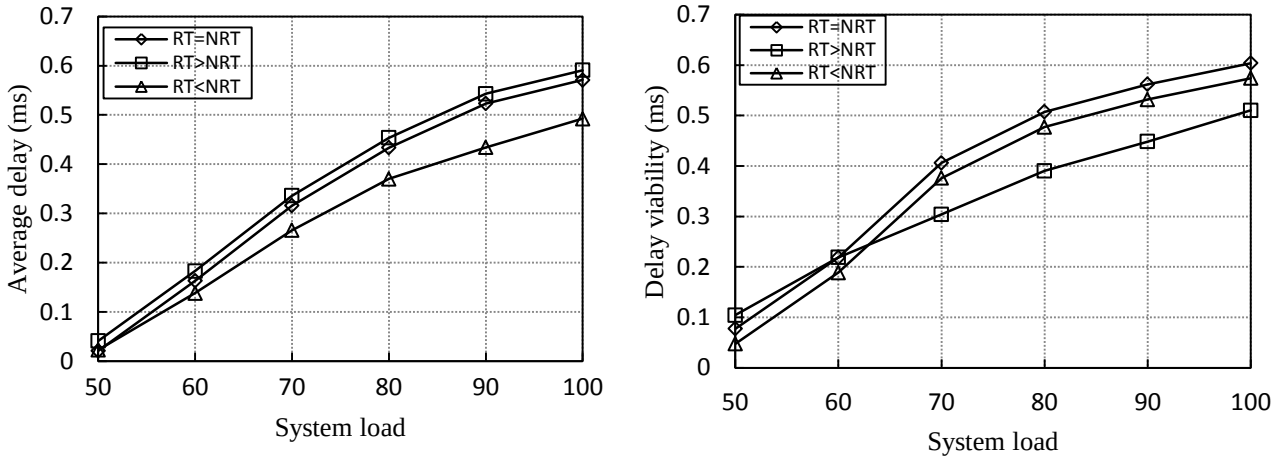

Figure 10. Average delay and delay viability of RT traffic.

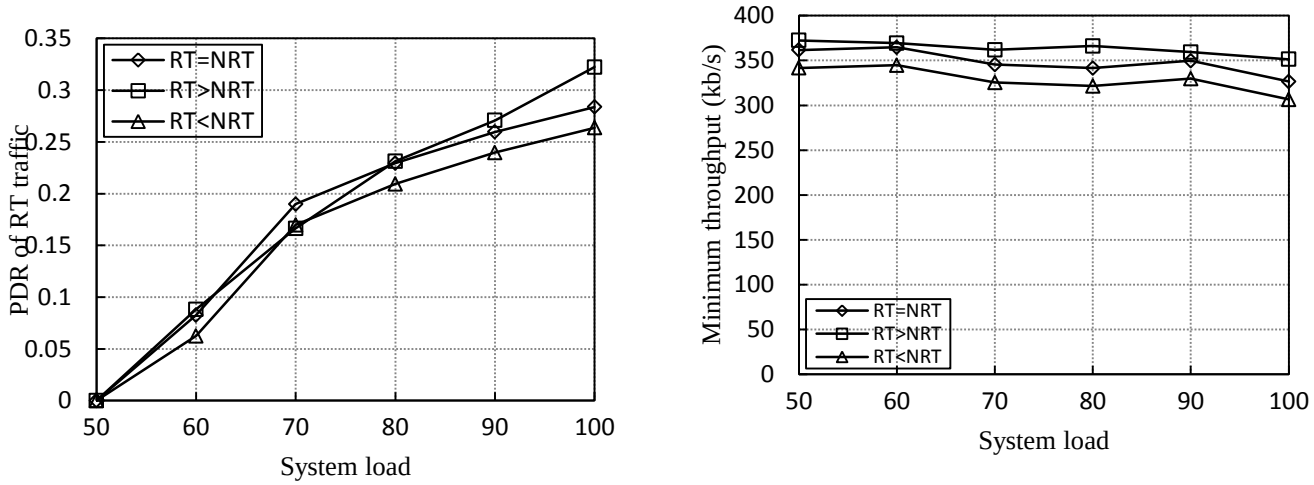

Figure 11. PDR of RT traffic and minimum throughput of NRT traffic.

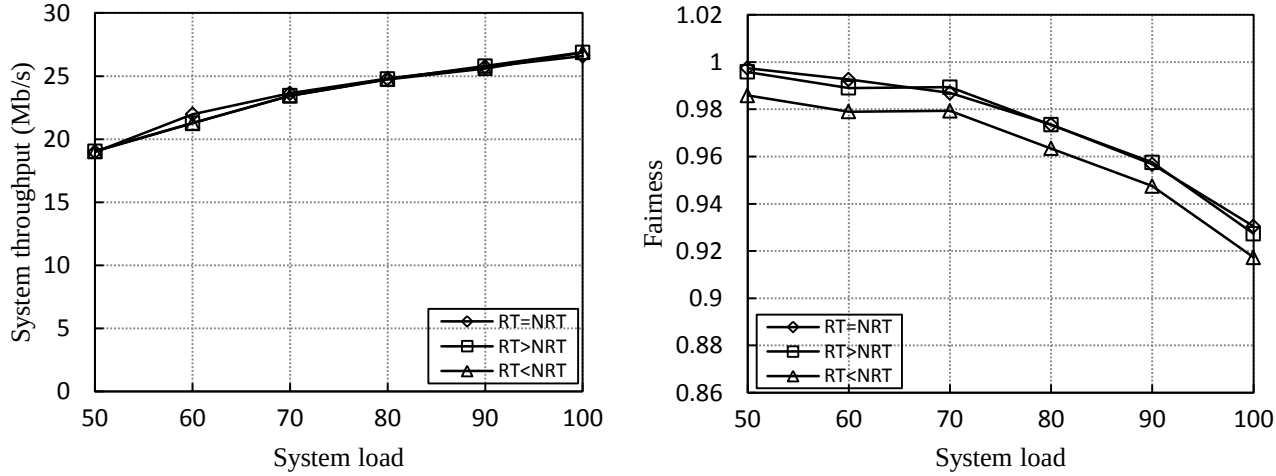

Figure 12. System throughput and fairness among users.