Clustering Based Packet Scheduling Adaptive to the Network Load in LTE-Advanced Networks

R. Kausar\textsuperscript{1}, Y. Chen\textsuperscript{2}, K. K. Chai\textsuperscript{3}
\textsuperscript{1,2,3} School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK.
\textsuperscript{1} rehana.kausar@elec.qmul.ac.uk.
\textsuperscript{2} yue.chen@elec.qmul.ac.uk.
\textsuperscript{3} michael.chai@elec.qmul.ac.uk.

Abstract

This paper investigates authors’ previously proposed novel clustering based packet scheduling algorithm for the downlink transmission of LTE-Advanced networks, under variable network conditions. Numerous simulations are run to investigate the performance and validity of this algorithm under different network scenarios such as equal number of real time and non-real time users, real time users more than non-real time users and vice versa. Under each scenario, the total number of users is also varied to validate the algorithm both for different network scenarios and for the variable overall network load. The key performance indicators are average delay, the delay viability and packet drop rate of real time users, minimum throughput of non-real time users, and system throughput and fairness among users. The simulation results show that the algorithm maintains the service level and system level performance under each network scenario and variable network load.

Keywords: Long Term Evolution-Advanced (LTE-A), Packet Scheduling (PS), Quality of Service (QoS), RT, NRT, Packet Drop Rate (PDR).
1. Introduction

The author’s previously proposed algorithm [1] improves the service level performance by enhancing the Quality of Service (QoS) provision for real time (RT) traffic while maintaining the throughput for Non Real Time (NRT) traffic at required level. In addition it improves the system level performance by maintaining a good trade-off between system throughput and fairness among users. In [1] K-mean clustering algorithm is integrated in the packet scheduling architecture to reduce the average delay and the delay viability of real time users. Hebbian Learning process is integrated in the time domain of packet scheduling architecture, to reduce PDR and to improve system level performance.

The rapid development in all Internet Protocol (IP) based next generation of mobile communications networks, such as Long Term Evolution-Advanced (LTE-A), is expected to support the outburst of high-speed packet based applications. These applications have a large variety of QoS requirements such as reduced delay and PDR and high throughput. Radio resource management (RRM) faces challenges when comes across such large variety of conflicting QoS requirements [2]. The reason being limited radio resources, rapidly changing wireless channel conditions and ever increasing number of mobile users. Packet scheduling (PS) being one of the cores of RRM is very crucial to make and effective utilisation of available radio resources [3]. The variable QoS requirements of different traffic types must be analysed and weighted to reach at a balanced solution [2].

The current work related to PS in Orthogonal Frequency Domain Multiple Access (OFDMA) systems takes into account Queue State Information (QSI) as well as Channel State Information (CSI) [4, 10]. The combined consideration of QSI and CSI significantly improves the support of QoS guarantees to Real Time (RT) and Non Real Time (NRT) traffic types. Adaptive TD Scheduling Algorithm (ATDSA) [11] proposed by the author of this paper considers both CSI and QSI in taking scheduling decisions and improves service level and system level performance in mixed traffic environment, in LTE-A downlink transmission.

However to deal with the variable QoS requirements of different traffic types, clustering is fairly good technique to integrate in PS algorithms to rearrange users’ priority according to their QoS requirements. Clustering is a technique to group together a set of items having similar characteristics [12]. In PS domain, where there is always a need to change priorities of different users according to their QoS requirements, clustering is a good technique to make groups of users with similar QoS requirements. It helps in setting priorities of users by sorting these groups in a proper way. Clustering has already been used in many domains especially on Web aiming at improving Web applications [12, 13]. In the Web usage domain, clustering improves market segmentation and the performance of internet search engines [12].

Clustering based scheduling gives variable priorities to different traffic types and to users belonging to same traffic type. For example in [14], a clustering based scheduling algorithm is used to organise nodes of a network into groups based on the number of their requests per channel. The transmission priority then starts from the group with the highest requests. It improves network performance in terms of higher network throughput while keeping mean packet delay at lower levels as compared to the conventional scheduling algorithms.

The work in this paper is inspired by the fact that different traffic types have variable QoS requirements. All requirements need be taken into account in scheduling decisions to effectively utilise the available radio resources. It is important to arrange the users’ scheduling order according to their QoS requirements. Therefore a clustering approach is integrated in ATDSA, to group RT users based on their PDR. Creating such groups and prioritising them properly could lead to higher network performance without aggravating the scheduling algorithm.

The rest of the paper is organised as follows. Section 2 describes the system model used for these simulations along with presents performance metrics to analyse PS performance. Section 3 and 4 give a brief description of cross layer design PS architecture and Clustering Based scheduling Algorithm [1]. Simulation model and results have been presented in section 5. Finally, section 6 gives the conclusion and future work.
2. System Model

An OFDMA system is considered in which minimum allocation unit is one PRB is containing 12 sub-carriers in each Transmission Time Interval (TTI) of 1 ms 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 (1, 8, 16).

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

Where, $Y_{(k,m)}(t)$ is the data symbol from eNodeB to user $k$ at sub-channel $m$, $X_{(k,m)}(t)$ is the input data symbol, $[H_{(k,m)}(t)]^*$ is the complex channel gain of sub-channel $m$ for user $k$, and $Z_{(k,m)}(t)$ is the complex White Gaussian Noise [8]. It is assumed, as in [7, 3, 8, 16], 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, assuming that the channel state information (CSI) $H_{(k,m)}(t)$ is known by the eNodeB as in [8]. The achievable throughput of a user $k$ on sub-channel $m$ can be calculated by (2) as used in [11, 12].

$$C_{(k,m)}(t) = B \log_2 \left[ 1 + \frac{|H_{(k,m)}(t)|^2}{\sigma^2 \Gamma} P_m(t) \right] \text{(2)}$$

Where, $B$ is the bandwidth of each PRB, $\sigma^2$ is the noise power density and $\Gamma$ is a constant signal-to-noise ratio (SNR) gap and has a simple relationship with the required Bit Error Rate (BER).

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

3. Packet Scheduling Architecture

A cross layer design QoS architecture is shown in Fig. 1 as used in [10], which takes into account of the information from Application, Network, Medium Access Control (MAC) and Physical layer, to take scheduling decisions. It consists of different functionalities such as, queue management, Adaptive Time Domain (TD) Scheduler and Frequency Domain (FD) Scheduler.

![Fig. 1. The cross layer packet scheduling architecture](image)

The input is a mixed traffic which is classified in four traffic types: Control traffic (i.e. control information), RT traffic (i.e. voice), NRT traffic (i.e. streaming video) and Best Effort (BE) traffic (i.e. email, SMS). Service specific queue sorting algorithms and ATDSA are used to sort users of different traffic types and to take decisions on the proportion of available radio resources to RT and NRT traffic types, respectively as used by the author in [11]. To exploit multiuser diversity in the TD and FD, the Channel Quality Information (CQI) reports are fed back to queue management and FD scheduler, respectively. A brief description on queue sorting algorithms and ATDSA is given in Section 3.1 and 3.2 [1].
3.1. Queue Management

Users in the control queue are sorted by Round Robin (RR) as control information is equally important for all users. Control queue is at the top and is allocated resources before all other queues. This is because control information is the most important.

$$P_{k}^{RT}(t) = \left( \frac{T_k^{wating}(t)}{DB^{RT}(t)} \times [H_k^{RT}(t)]^2 \right) + [Q_k(t)]^2$$

(4)

Where $P_{k}^{RT}(t)$ is the priority, $T_k^{wating}(t)$ is waiting time, $H_k^{RT}$ is the channel gain and $Q_k(t)$ is queue length of RT user $k$ at time $t$. The priority metric takes into account waiting time, channel conditions and queue length of RT users to improve fairness, multiuser diversity in the TD and delay viability respectively.

The QoS requirement for NRT is defined as $\eta_k(t) > T_k$, where $\eta_k(t)$ is the instantaneous throughput of user $k$ at time $t$ and $T_k$ is throughput requirement of NRT user $k$. Minimum throughput requirements of NRT traffic are fulfilled by sorting NRT queue with (5) as used in [10].

$$P_{k}^{NRT}(t) = \frac{T_k^{wating}}{DB^{NRT}(t)} \times \frac{T_k^{(t)}}{R_k(t)} \times [H_k^{RT}]^2$$

(5)

Where $P_{k}^{NRT}(t)$ is the priority metric, $T_k^{wating}(t)$ is the waiting time, $R_k(t)$ is the average achieved throughput and $H_k^{NRT}(t)$ is the channel condition of user $k$ at time $t$. In (5) $DB^{NRT}$ is the delay upper bound for NRT video packets which is taken equal to RT packet in this paper. Equation (5) takes into account waiting time, minimum throughput requirement and channel conditions to support fairness, throughput guarantee and exploit multiuser diversity in the TD, respectively. Users in BE queue does not have any QoS requirement, however to maintain fairness level among users, proportional Fairness (PF) [13] algorithm is used to sort users in this queue as given in (6).

$$P_{k}^{BE} = \frac{T_k}{R_k}$$

(6)

Where $\eta_k(t)$ is instantaneous and $R_k(t)$ is average achieved throughput of BE user $k$.

3.2. Adaptive TD Scheduling Algorithm (ATDSA)

The ATDSA uses Hebbian learning process to allocate the radio resources dynamically based on the QoS feedback. The Hebbian learning process compares PDR value of RT traffic in the current TTI with the previous values of PDR and changes the weight of RT traffic according to (7) as used in [10].

$$W_{RT}(t) = \begin{cases} W(t-1) + \eta & \text{if } PDR_{RT}(t) > PDR_{RT}(t-1) \\ W(t-1) - \eta & \text{Otherwise} \end{cases}$$

(7)

Where $W_{RT}(t)$ is the weight given to the RT traffic at time $t$, $0 < \eta < 1$ is learning rate. Value of $W_{RT}(t)$ is maintained between 0 and 1 and increases each time by $\eta$, if the PDR of RT traffic increases. It becomes 1 when PDR of RT traffic exceeds the PDR threshold $T$, at which resource allocation to RT traffic is increased. If PDR is lower than the threshold then there is a decrease in RT weight equal to $\eta$. When the weight of RT traffic is decreased, the allocation of radio resources to RT traffic is also decreased so that resources can be allocated to other traffic types.

Let $C$ be the total available radio resources and $\lambda C$ is the proportion of $C$ allocated to RT traffic then $(1 - \lambda)C$ is the proportion of available radio resources allocated to NRT traffic types. The adaptive change in $\lambda$ based on Hebbian learning process is defined by (8).

$$\lambda(t + 1) = \lambda(t) \pm W_{RT}(t) \times \text{Min allocation unit}$$

(8)

Where $\lambda(t)$ is resources allocated to RT traffic at time $t$ and min allocation unit is 1 PRB. The adaptive change in RT capacity based on Hebbian learning process is shown in Fig. 2.
The proportion of available resources allocated to the NRT traffic is further divided in different types of the NRT traffic (NRT streaming video and BE) by prioritizing the streaming queue to guarantee its throughput requirements. As BE queue does not have any QoS requirements and rest of the resources are allocated to the BE queue.

\[
PDR(t) > PDR(t-1) \\
W_{RT}(t) = W_{RT}(t-1) \\
W_{RT}(t) = W_{RT}(t-1) + \eta \\
PDR > \mathcal{T} \\
W_{RT} = 1
\]

Fig. 2. Hebbian learning process

4. Clustering based Scheduling

To reduce average delay of RT traffic and delay viability of individual RT users, clustering is used to group RT users with highly strict QoS requirements and the ones with comparatively far from reaching their upper bound of delay/PDR thresholds. Clustering algorithms are used to group data in such a way that objects with similar characteristics are in one cluster and objects with dissimilar characteristics are in other cluster. In this paper clustering algorithm is integrated in ATDSA [10] (presented by author) to further enhance the QoS provision to RT traffic. It is described as below [1].

For a clustering process \( Clus \), \( num \) denotes the number of clusters to be created having one central point (Centroid), \( X \) denotes the number of RT users \( X = \{x_1, \ldots, x_n\} \) that is to be clustered and \( \mu_1, \mu_2, \ldots, \mu_{num} \) represent each of the \( num \) clusters consisting of \( |C_1|, \ldots, |C_{num}| \) members that is number of RT users. The clustering process \( Clus \) is defined as an assignment of RT users to the group of users i.e., clusters.

\[
CLUS: \{1, \ldots, n\} \rightarrow \{1, \ldots, num\} \tag{9}
\]

Users belonging to the same cluster have similar requirements regarding their PDR threshold and dissimilar to the users belonging to other clusters. Similarity in clustering is fundamental to make appropriate groups of users. The dissimilarity between two users is evaluated by distance measure. Each point of given set \( X \) is assigned to its nearest Centroid based on Euclidean distance as given in (10).

\[
C_i = \min_{i=1,2,\ldots,n} ||x_i - \mu_k||^2 \tag{10}
\]

Where \( C_i \) is the Centroid to which \( x_i \) is assigned and \( \mu_k \) is the \( k^{th} \) Centroid. The squared Euclidean distance uses the same equation as the Euclidean distance but does not take the square root while evaluating dissimilarity. The mean of all points assigned to a Centroid is calculated and the position of each Centroid is updated by the mean of points assigned to it. This process is repeated until no Centroid is shifted in the next iteration, resulting in \( num \) clusters. Considering all clusters, the clustering process is guided by the cost function \( J \), which is the sum of distances between each user and the Centroid to which it is assigned as given below in (11).

\[
J = \sum_{i=1}^{num} \sum_{x_i \in C_j} dist(x_i, C_j) \tag{11}
\]
The optimization objective is to minimize $J$ so that the dissimilarity between the users of same cluster becomes minimum or null. For the clustering process, K-mean algorithm is used which is widely used clustering algorithm [14]. It minimizes the cost function $J$ given in (11). The process is shown in Fig.3.

The integration of K-mean clustering in ATDSA is abbreviated by ATDSAK. The core idea of integrating clustering in ATDSA is to rearrange RT users’ priority list in the TD based on the PDR threshold of RT traffic, before allocating PRBs in the FD. Once the RT users are grouped into clusters, these clusters are sorted to give priority to the cluster with users having PDR near to the PDR threshold. The users in the sorted clusters are then allocated resources in the FD [1].

5. Simulation Model and Results

5.1. Simulation Model

A single cell with one eNodeB, total system bandwidth of 10 MHz and PRB size of 180 kHz is considered. Total system bandwidth is divided into 55 PRBs. The wireless environment is typical Urban Non Line of Sight (NLOS) and the LTE system works with a carrier frequency of 2GHz. The most suitable path loss model in this case is the COST 231Walfisch-Ikegami (WI) [18] as used in many other papers on LTE. Users are assumed to have a random distribution and the total number of RT users is assumed to be equal to total number of NRT users as in [8, 11]. The delay budget for RT traffic is 40ms in OFDMA-based networks [8, 19] and the required throughput by NRT traffic is taken 240kbps as in [8, 11]. Total eNodeB transmission power is 46dBm (40w) and maximum BER requirement is $10^{-4}$ for all users. The simulation parameters used for system level simulation are based on [19] and these are typical values used in many papers. These parameters are listed in Table 1.

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

The performance of author’s novel packet scheduling scheme ATDSAK [1] under varying network load is investigated in this paper. This is to validate the adaptability of ATDSAK in varying network conditions. Simulations are performed by changing number of active RT and NRT users and the results are compared with ATDSAK [1]. The performance validation of ATDSAK was presented in [1] where it was compared against QoS aware mixed traffic packet scheduling algorithm (abbreviated as MIX in all figures) [4] and SWBS [8]. MIX and SWBS packet scheduling algorithms in [4 and 8] respectively are used as reference algorithms for comparison. MIX [4] classifies mixed traffic in different queues and sort users in these queues with queue specific algorithms at classifier, picks users from queues by fair scheduling in the TD and allocates resources to these users in the FD. In Fair TD scheduling, users are picked from the queues one-by-one thus allocating a fair share of radio resources to all queues. SWBS [8] uses a sum waiting time based scheduling algorithm in which sum waiting time of packets is taken into account while prioritizing RT and NRT traffic types. For RT traffic it takes real arrival time and for NRT traffic, it takes virtual sum arrival time of packets which is related to the minimum throughput requirement of NRT traffic.

As in [1], in this paper, the packet arrival process for RT, NRT streaming video and BE traffic is Poisson distribution with 0.35 ON time. The total number of active users is varied from 50 to 100. In [1] the traffic load of RT and NRT traffic is equal and in this paper the traffic load scenarios are as follows:

**RT=NRT**

In this scenario, RT users are 50% of the total active users and NRT users are also 50%. In the simulation results it is represented by ATDSAK.

**RT>NRT**

In this scenario, RT users are 70% of the total active users and NRT users are 30%. In the simulation results it is represented by ATDSAK1.

**RT<NRT**

In this scenario, RT users are 30% of total active users and NRT users are 70%. In the simulation results it is represented by ATDSAK2.

Fig. 4 shows average delay of RT traffic verses total number of active users. Average delay increases with the number of users for all algorithms as shown. ATDSAK shows average delay results of [1]. ATDSAK1 and ATDSAK2 show the average delay of scenario RT>NRT and RT<NRT. As can be seen, there is a minor difference between ATDSAK and other scenarios which validates the performance of the proposed algorithm [1].

![Fig. 4. Average delay of RT traffic](image)

Fig. 5 shows delay viability of RT users verses total number of active users. The delay viability of RT users is even decreased under variable network conditions as shown. This indicates that [1] maintains its performance and is adaptable to the varying network load.
The average PDR of RT traffic as shown in Fig. 6 increases with the number of users for all scenarios. ATDSAK1 shows average PDR values very near to [1] which means the performance of [1] is maintained. However ATDSAK2 shows a little higher PDR as compared to [1] which is because most of the resources are allocated to the massive NRT traffic.

Minimum throughput of NRT streaming video traffic is calculated by (9) and is shown in Fig. 7. ATDSAK1 and ATDSAK2 both fulfill minimum throughput requirements of NRT streaming video traffic as shown. In addition by varying load, the performance has become even better.

System overall throughput is shown in Fig. 8. System throughput increases with number of users. Under all network loads including RT>NRT, RT=NRT and RT<NRT, system throughput remains nearly similar. This validates that author’s proposed algorithm [1].
Fig. 8. System throughput

Fairness among users at RT=NRT is highest because real time traffic is equal to non-real time traffic. The fairness level falls a little with RT>NRT and RT<NRT because of the inequality of RT and NRT users, which is normal.

Fig. 9. Fairness among users

6. Conclusion

In this paper, author’s clustering based packet scheduling algorithm is investigated under variable network conditions to validate the adaptability of this algorithm. Three network scenarios are chosen for this investigation including RT=NRT, RT>NRT and RT<NRT. The total number of active users are varied from 50 to 100 for each network scenario. Simulation results show that the algorithm maintains its performance under various network condition. On service level, average packet delay is reduced and remains nearly same in each network scenario and under variable traffic load, delay viability is even reduced when number of RT and NRT users is varied. The PDR is increased by a minor value under scenario RT<NRT, however the minimum achievable throughput for of NRT traffic under RT<NRT is higher than the rest scenarios. On system level, system throughput is maintained at good level by all scenarios with a slight decrease in fairness among users under network scenario RT=NRT.

References


