Enhanced PF Scheduling Algorithm for LTE Networks

Renê Pomilio de Oliveira, Lourival Aparecido de Góis and Augusto Foronda
Federal University of Technology – Paraná, Ponta Grossa, Brazil

Abstract: Scheduling Algorithm is a key Radio Resource Management (RRM) mechanism to achieve Quality of Service (QoS) requirements and to optimize system performance of Long Term Evolution (LTE) network. The LTE standard does not specify a scheduling algorithm and some schedules have been proposed in the literature. Most of them consider the channel state or some QoS metric, however, they do not consider the PF characteristics and the delay required by the user. Then, an enhanced scheduling algorithm was proposed based on Latency-Rate (LR) server theory and system characteristics specified by the LTE standard. Properties of this proposal have been investigated theoretically and through simulations. A simulation is presented with video traffic and also performance comparisons with Proportional Fairness (PF) and Modified Largest Weighted Delay First (MLWDF) schedulers. The results show that the proposed schedule has a better performance compared with the above schedulers.

Keywords: LTE, QoS, Latency-Rate, Scheduling.

1. Introduction

Long Term Evolution (LTE) was developed with the aim to support the Quality of Service (QoS) requirements of various multimedia applications available on the Internet such as 3D video conferencing and mobile HD TV [1]. These applications have heterogeneous Quality of Service (QoS) requirements such as transmission delay, jitter, packet loss rate, packet error rate etc., to provide better user experience. The LTE Radio Resource Management (RRM) block located at the base station, called the evolved NodeB (eNB), performs some tasks such as Call Admission Control and Packet Scheduling (PS), which are responsible for accepting a certain number of users to guarantee the required QoS and for distributing radio resources among user equipment (UEs), respectively. Both are an open issue for designers. As a consequence, the LTE Call Admission Control and Packet Scheduling have attracted the attention of researchers from both industry and academy. These scheduling algorithms do the radio resource allocation either by time domain approach (TD) or frequency domain approach (FD). In the Time domain approach, the LTE system assigns all the system resources to one UE during the particular transmission time interval (TTI). In frequency domain the resource is allocated to UE based on frequency and time domains [2]. The scheduling in both downlink and uplink is carried out by scheduler present at the Medium Access Control (MAC) sub-layer of eNB. Many scheduling algorithms have been proposed in the literature and system performance is significantly different. PF scheduler tries to allocate user in a fairer way considering the quality of the link. MLWDF is an enhanced PF scheduler, which considers the user delay and the quality of the link. MLWDF was developed because the video traffic on the Internet has increased and it is important to decrease delay for users. MLWDF has a better performance than PF, however, it is necessary to consider another metric to select a user. This paper presents an enhanced PF scheduling algorithm for LTE networks, which considers the delay required by the user and the traffic characteristics. We have developed an analytical model based on Latency-Rate (LR) server theory [3] and the rate for each user is calculated with the traffic characteristics and delay required which will be used to enhance the PF scheduler. After developing this model, a set of simulations is presented for video traffic. The performance of the proposed model was compared with some well-known schedulers, such as PF and MLWDF.

The paper is structured as follows. In Section 2, related research is described. In Section 3, a brief description of the LTE standard is presented. Our analytical model for packet scheduling is proposed and explained in Section 4. Evaluation of the capacity of the new model is shown in Section 5. Conclusions are presented in Section 6.

2. Related Work

Schedulers are resource allocation mechanisms that are responsible for distributing resources among different user types. Schedules are located in eNB more precisely on the MAC layer, dynamically allocating resources in both downlink and uplink. Moreover, the scheduling algorithms provide different characteristics for each user to be able to transmit in each time slot, whether they are continuous or not, defining different metrics for each user depending on the type of application that the UE is using at any given time. The PF selects the user with the best instantaneous data rate in relation to their average data rate, requiring a greater eNB effort to inform the UEs about their slot positions [4]. This scheduler tries to allocate users in a fairer way considering the state of the link.

The MLWDF is used to support multiple users with different QoS, prioritizing a user with a longer packet delay and with better channel conditions than the average. Although this scheduler considers the delay for user selection, it is not the latency specified by the user [5]. The author in [6] uses game theory to optimize the allocation of resource blocks. The primary key to improve the performance when using 4G is resource allocation. A solution based on game theory was provided by improving the allocation of 4G radio resources and optimizing the system as presented in simulations comparing the two schedulers PF and MLWDF. The metrics of simulation to validate the game theory approach in the LTE network were throughput, delay, packet loss and fairness index.
In [7], the author proposes an algorithm to schedule users to provide justice among the users based on Physical Resource Blocks (PRBs). Once the scheduler knows the number of users to be schedule and the bandwidth, the proposed scheduler is able to provide a fairness of resource blocks equal to all users and even poor channel users can increase the number of allocated times.

The FLS scheduler aims to improve the real-time service quality for LTE downlink [8]. The algorithm consists of two distinct levels that interact together by dynamically scheduling radio resources to users, considering channel condition conditions, maximum delay, and data source behavior. When using the algorithm at the highest level, FLS makes the frame-by-frame definition to know the amount of data the data source needs to satisfy the delay constraint in order to be able to transmit in real time to the UE, generating a linear control cycle. At the lowest level, the PF scheduler algorithm assigns the RBs to each TTI considering the bandwidth requirements calculated at the high level by the FLS. The approach of this scheduler defines in a single time the amount of data that must be transmitted by each data source.

The EXP-PF scheduler has been designed to increase the priority of real time flows with respect to non-real time ones. The metric used to select real time flows considers the head of a line packet delay and the metric for non-real time flows is the one of the PF [9].

In [10], the author proposes a scheduling algorithm for the LTE downlink aiming to improve the delay and the throughput of the users in the system. The GLWD (Generalized Largest Weighted Delay First) scheduler algorithm is an improvement over the conventional MLWD scheduler algorithm. The scheduler differentiates packet-level service based on real-time packet priority within a video stream, improving delay and throughput. The results presented by the author in the simulations in comparisons to conventional scheduling algorithms including PF, EXP-PF and especially MLWD, show that the proposed scheduler outperforms in terms of throughput, real-time delay, packet loss rate, and delay loss restriction, the scheduling algorithms compared.

The author [11], presents a proposal to improve the performance of the LTE downlink scheduler The scheduler algorithm implements a "Dynamic Multi-traffic Scheduler" using bandwidth allocation and resource blocks based on the number of classes of services. The proposed scheduler considers the different channel conditions of users and creates a balance between QoS guarantees and fairness for multi-users The author presents the results of the comparisons between the proposed dynamic and static scheduler algorithm. The results show that the proposed algorithm outperforms the other schedulers improving performance and fairness compared to static schedules.

Some of the schedulers consider the channel quality to select users and some of them also consider a QoS metric, such as delay. However, none of them consider the delay required by the user or the traffic characteristics. For example, a user with a stricter delay requirement should be selected more frequently. And a user with higher traffic rate should also be selected more frequently. Our proposed model aims to improve LTE downlink scheduler which will consider channel quality, the delay required by the user and traffic characteristics.

3. The LTE Standard Overview

The LTE Standard currently in release 8, has been developed and specified by Third Generation Partnership Project (3GPP) named LTE 3.9G, the technology created to work with a high data packet transfer rate both downlink and uplink. The standard works with high-speed wireless communication for mobile phones and data terminals to transfer the information reducing the latency down to 10ms and the time of access. The use of both OFDMA in the downlink and SCFDMA in the uplink allows to guarantee a better spectral efficiency, more specification on release 8 are summarized in Table 1.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak data rate</td>
<td>Downlink: 300 Mbps</td>
</tr>
<tr>
<td></td>
<td>Uplink: 75 Mbps</td>
</tr>
<tr>
<td>Modulation (Uplink/Downlink)</td>
<td>QPSK, 16QAM and 64QAM</td>
</tr>
<tr>
<td>Bandwidth (MHz)</td>
<td>1.4, 3, 5, 10, 15, and 20</td>
</tr>
<tr>
<td>TTI</td>
<td>1ms</td>
</tr>
<tr>
<td>Sub-carrier spacing</td>
<td>15KHz</td>
</tr>
<tr>
<td>Access schemes</td>
<td>OFDMA (Downlink)</td>
</tr>
<tr>
<td></td>
<td>SC-FDMA (Uplink)</td>
</tr>
<tr>
<td>Duplex schemes</td>
<td>Frequency Division Duplex (FDD)</td>
</tr>
<tr>
<td></td>
<td>Time Division Duplex (TDD)</td>
</tr>
<tr>
<td>Mobility</td>
<td>Low speeds (0 - 3 Km/hr)</td>
</tr>
<tr>
<td></td>
<td>High speeds (30 - 120 Km/hr)</td>
</tr>
<tr>
<td>Supported antenna</td>
<td>Downlink: 4x4, 4x2, 2x2, 1x2 and 1x1</td>
</tr>
<tr>
<td>configurations</td>
<td>Uplink: 1x2 and 1x1</td>
</tr>
</tbody>
</table>

Further, this standardization aims to support devices from previous technologies by applying both Frequency Division Duplex (FDD) and Time Division Duplex (TDD), to support estimated throughput flows at 300Mbit/s downlink and 75Mbit/s uplink, as shown in Figure 1.

![Figure 1. Time division duplex structure](image)

The high-level LTE network architecture can be exemplified by the three main components, which consist of Evolved Packet Core (EPC), Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) and User Equipment (UE) as shown in Figure 2 [2].

The EPC is a framework for providing converged voice and data on a 4G, the core consists of three logical nodes in their functionalities, such as Packet Data Network Gateway (PGW), Serving Gateway (S-GW) Mobility Management Entity (MME). The network core is IP-based also providing access to both 3GPP and non-3GPP technologies, in other words, means that their accesses were not specified in the 3GPP standard.

E-UTRAN is a simple network of eNBs, its main function is to manage the Radio Resource Management (RRM) mechanism that is responsible for schedules, relay protocols, coding, handover, call admission control and power control.
eNB consists of making the point-to-point connection with the users’ terminals in the data transmissions between the radio and the EPC network [12].

Figure 2. Architecture LTE

4. Analysis of The Analytical Model

Figure 3 illustrates an LTE network with the proposed schedule, which is based on a modified LR schedule and the token bucket algorithm. The basic approach consists of the token bucket limiting input traffic and the LR schedule providing rate allocation for each user. The token bucket size and token bucket rate are calculated according to the input traffic characteristics. And the rate allocated for each user is estimated according to the delay required by each user. Then, the rate allocated to each user will be used to enhance the PF schedule, named Proportional Fair - Latency Rate Scheduler (PFLR) schedule.

Figure 3. LTE network with new scheduler

A scheduler that provides guaranteed bandwidth can be modeled as an LR scheduler. The behavior of an LR scheduler is determined by two parameters for each session $i$: latency ($\theta_i$) and allocated rate ($\tau_i$). The latency ($\theta_i$) of the scheduler may be seen as the worst-case delay and depends on network parameters. In the new scheduler, the latency $\theta_i$ is

$$\theta_i = TF + \frac{L_{\text{max},i}}{R} \quad (1),$$

where $TF$ is time to allocate all the users, and $L_{\text{max},i}$ is the maximum size of a packet in session $i$ and $R$ is the capacity of the output channel.

Now, we show how the allocated rate $\tau_i$ for each session $i$ may be determined.

An LR scheduler can provide a bounded delay if input traffic is shaped by a token bucket. A token bucket is a non-negative counter, which accumulates tokens at a constant rate $i_{\text{rate}}$ until the counter reaches its capacity $i_{\text{size}}$. Packets from session $i$ can be released into the queue only after removing the required number of tokens from the token bucket. In an LR scheduler, if the token bucket is empty, arriving packets are dropped. However, our model ensures that there will always be tokens in the bucket and packets will not be dropped. If the token bucket is full, a maximum burst of $i_{\text{size}}$ packets can be sent to the queue. When the flow is idle or running at a lower rate as the token size reaches the upper bound $i_{\text{size}}$, accumulation of tokens will be suspended until the arrival of the next packet. We assume that the session starts out with a full bucket of tokens.

In our model, we consider LTE standard overhead for each packet. Then, as we will show below, the token bucket size will decrease by both packet size and overhead.

The application session $i$ declares the maximum packet size $L_{\text{max},i}$ and the maximum allowable delay $D_{\text{max},i}$, which are used by the LTE scheduler to calculate the service rate for each session so as to guarantee the required delay. The service rate for each session will be considered to modify the original PF scheduler. Then, the user selection will consider the channel conditions and the delay required by each user.

Incoming traffic $A_i(t)$ from session $i$ ($i = 1, \ldots, N$) passes through a token bucket inside the user terminal during the time interval $(0, t)$, as shown in the Figure 4.

Figure 4. Inbound Traffic with the Token Bucket

The data traffic is bounded by

$$A_i(t) \leq \sigma_i + \rho_i t \quad (2).$$

Then, the packet is queued in the station until it is transmitted by the wireless medium. Queue delay is measured as the time interval between the arrival of the last bit of a packet and its
transmission. In the new scheduler, queuing delay depends on token bucket parameters, network latency and allocated rate. In [3] and [13], it is shown that if input traffic $A_i(t)$ is shaped by a token bucket and the scheduler allocates a service rate $\tau_i$, then an LR scheduler can provide a bounded maximum delay $D_i$:

$$D_i \leq \frac{\sigma_i}{\tau_i} + \theta_i - \frac{l_{\text{max},i}}{\tau_i} \quad (3),$$

where $\tau_i$ is the service rate, $\sigma_i$ is the size of the token bucket, $\theta_i$ is the latency, and $l_{\text{max},i}$ is the difference between the lower and upper limits, as shown in Figure 5.

![Figure 5. Maximum delay $D_i$](image)

Equation (3) is an improved bound on the delay for LR schedulers. Thus, the token bucket rate plus the overhead transmission rate must be smaller than the service rate to provide a bounded delay. The upper bound $D_{\text{bound}}$ should be smaller than or equal to the maximum allowable delay:

$$D_{\text{bound}} = \frac{\sigma_i}{\tau_i} + \theta_i - \frac{l_{\text{max},i}}{\tau_i} \quad (4).$$

Therefore, three different delays are defined. The first is the maximum delay $D_i$, the second is the upper bound on the delay $D_{\text{bound}}$ and the third is the required maximum allowable delay $D_{\text{max},i}$. The relation between them is $D_i \leq D_{\text{bound}} \leq D_{\text{max},i}$. So, the delay constraint condition of the new scheduler is

$$\frac{\sigma_i}{\tau_i} + \theta_i - \frac{l_{\text{max},i}}{\tau_i} \leq D_{\text{max},i} \quad (5).$$

The second delay constraint condition is the token bucket rate plus the rate to transmit overhead and a maximum-sized packet must be smaller than the service rate to guarantee a bound on delay. Thus, the second constraint condition is

$$\rho_i + \frac{l_{\text{max},i}}{T_P} \leq \tau_i \quad (6).$$

This analytical model is used to calculate the user rate according to the delay required by the user and traffic characteristics. This user rate will be used to modify the PF scheduler metric. The PFLR scheduler is presented in Equation (7) and the control parameters are presented in Equation (8).

$$W_{i,j} = \frac{\gamma_i}{K} \times X \quad (7),$$

$$X = 0.5 \text{ if } \gamma_i > \tau_i,$$

$$X = 1.5 \text{ if } \gamma_i < (\tau_i \times 0.7) \quad (8),$$

where $W_{i,j}$ is the PFLR scheduler metric, $\gamma_i$ is the user throughput and $X$ is the control parameter.

As seen previously, the analytical model has two constraints. The 1st constraint shown in Figure 6 illustrates Equation 3 and how the maximum delay $D_i$ is calculated to restrict the delay requested by the user. The first term is the size of the token bucket divided by the service rate $\frac{\sigma_i}{\tau_i}$ the 2nd term which is the network latency $\theta_i$ the 3rd term is $\frac{l_{\text{max},i}}{\tau_i}$.

![Figure 6. First constraint for maximum delay $D_i$](image)

The second constraint of the model is presented in Figure 7, where 3 different situations are exemplified.

![Figure 7. Second constraint for maximum delay $D_i$](image)

- The situation (A) is ideal where Equation (6) has equality $\rho_i + \frac{l_{\text{max},i}}{T_P} = \tau_i$;
- Situation (B) has the second constraint met but is not ideal $\rho_i + \frac{l_{\text{max},i}}{T_P} < \tau_i$;
- Situation (C) does not guarantee the restrictions because $\rho_i + \frac{l_{\text{max},i}}{T_P} > \tau_i$;

5. Performance Evaluation

In this paper, a performance evaluation of the algorithm proposed in the analytical model is presented. The tool used to perform the simulations was the LTE-Sim [14]. LTE-Sim is a simple LTE network event simulator, developed C++
language. Moreover provides support for multiple scenarios, both simple and multi-cell environments, QoS management, user mobility, handover, including implemented scheduling algorithms such as PF and MLWDF. Integrated with self-generated graphical delay, throughput, loss packet rate. In order to perform these simulations in a realistic scenario we used the parameters composed of single cell with interference, eNB number equal to 1, radius equal to 1 km, UE range 1 → 13 evenly distributed for the cell, UE mobility 0, 3, 30 and 120 km speed, flow Video, bandwidth 5 MHz in downlink, other parameters are summarized in Table 2. Table 3 describes the input traffic characteristics.

Table 2. Simulation parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth</td>
<td>5 MHz</td>
</tr>
<tr>
<td>PHY</td>
<td>OFDMA</td>
</tr>
<tr>
<td>Slot Time Duration</td>
<td>1 ms</td>
</tr>
<tr>
<td>Delay required by the user</td>
<td>0.2 s</td>
</tr>
<tr>
<td>Simulation duration</td>
<td>46 s</td>
</tr>
<tr>
<td>Frame Structure</td>
<td>TDD</td>
</tr>
<tr>
<td>Frame Length</td>
<td>10 ms</td>
</tr>
<tr>
<td>Cell Radius</td>
<td>1 km</td>
</tr>
<tr>
<td>UEs</td>
<td>1 → 13</td>
</tr>
<tr>
<td>User Speed (km)</td>
<td>0, 3, 30 and 120</td>
</tr>
<tr>
<td>eNodeB</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3. Description of the input traffic

<table>
<thead>
<tr>
<th>User</th>
<th>Application</th>
<th>Arrival Period (ms)</th>
<th>Packet size (max) (bits)</th>
<th>Sending rate (kb/s) (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 → 13</td>
<td>Video</td>
<td>40</td>
<td>1490</td>
<td>242</td>
</tr>
</tbody>
</table>

First, token bucket parameters are estimated as described in Section 3. Since we want to find the minimum token bucket parameters for packet loss rate in the token bucket, we choose the token bucket size and the token bucket rate much larger than maximum packet size from video (1490 bits). After this, for each user input traffic, we decrease the token bucket size until the token bucket start to drop packets and then, we decrease the token bucket rate until the token bucket start to drop packets. The tokens bucket parameters are described in Table 4.

Table 4. Token bucket parameters

<table>
<thead>
<tr>
<th>Video</th>
<th>Bucket size (bits)</th>
<th>33000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bucket rate (Kbps)</td>
<td>280</td>
</tr>
</tbody>
</table>

Furthermore, with the parameters of the Bucket size and Bucket rate, it was possible to establish metrics of the user’s video rate for delay required, as shown in Table 5.

Table 5. User rate (τ)

<table>
<thead>
<tr>
<th>Delay (s)</th>
<th>Video (kbps)</th>
<th>user rate</th>
<th>Nº of users</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>300</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

The users 1 to 13 requires 0.2s of delay. The results of the simulations with different user speeds are presented below. These results are the average user delay rate comparing the PF, MLWDF and PFLR schedulers. We noticed that for the users the delay decreased considerably when compared with PF and MLWDF schedulers. After the user rate and requested delay parameters were calculated, another parameter evaluated was the control parameter. This parameter provides a control for user selection by modifying the PF scheduler metric. Below is shown how these values for the control parameter were calculated through numerous simulations.

Figure 8 shows how the control parameter was calculated when the user rate is greater than 300 bits. In simulations the average user rate ranged from 160 to 400 bits. The PFLR scheduler attempts to maintain a user rate balance for all users in the system to ensure an average rate between 200 to 300 bits and decreasing the delay. When the user rate is greater than 300 bits, the control parameter that best suits is 0.5 in all simulations made.

Figure 9 shows the user rate being less than 200 bits. When the user rate is less than 200 bits, the ideal control parameter is 1.5.

Now we explain how the delay decreases with the proposed model. Figure 10 shows the results of the simulation output file running the PF scheduler. It is noted that a packet of 3820 bytes is sent and it is fragmented in 3 different sizes of 1490, 1490 and 840 bytes generating a delay of 0.005, 0.010 and 0.013 seconds taking in average 0.010 seconds to receive the complete package.
The output file of the simulator running the PFLR scheduler is shown in Figure 11. Note that the delay decreases to 0.005, 0.005, 0.003 seconds, taking in average 0.005 seconds to receive the complete package.

| TX Video | ID 0 | B 0 | SIZE 1490 | SRC -1 | DST 1 | T 1.26422 |
| RX Video | ID 0 | B 0 | SIZE 1490 | SRC -1 | DST 1 | D 0.0057717 |
| TX Video | ID 1 | B 0 | SIZE 1490 | SRC -1 | DST 1 | T 1.30622 |
| RX Video | ID 1 | B 0 | SIZE 1490 | SRC -1 | DST 1 | D 0.0057717 |
| TX Video | ID 2 | B 0 | SIZE 840 | SRC -1 | DST 1 | T 1.34822 |
| RX Video | ID 2 | B 0 | SIZE 840 | SRC -1 | DST 1 | D 0.0037717 |

**Figure 11.** Output file with PFLR scheduler

The delay decreases because the token bucket modifies the input traffic from VBR to CBR as show in Figure 10 and 11. Also, the new metric tries to keep the user throughput estimated by the analytical model to guarantee the required delay by the user.

Figure 12 shows the result for static users. Figure 13 shows the result for walk users. Figures 14 and 15 shows the result for vehicular users.

The average delay of all the scheduling users in the system was a 0.19s delay, guaranteeing a delay below 0.2s, maintaining a high throughput and validating the analytical model so that the delay requested by the UE is less than 0.2s.

It is important to note that the proposed scheduler PFLR presents a lower delay when compared to the other schedulers maintaining similar throughput. This is because PFLR attempts to maintain a balanced throughput.

As the user’s speed is increased, the delay is decreased, because the chance of being selected more often is greater as the channel quality improves.

The throughput was also analyzed in simulations with users with different speeds. Figure 16 shows the result for static users. Figure 17 shows the result for walk users. Figures 18 and 19 shows the result for vehicular users. The throughput estimation follows the same approach as in [15].

The disadvantage of the proposed scheduler is that throughput decreases a little when compared to the other schedulers, because the video stream has its transmission time changed by the token bucket as shown in Figure 10. The PFLR scheduler takes just a few more seconds to complete the transmission of all packets, interfering in the throughput calculation since the sending time of each packet is longer in comparison to the PF.
6. Conclusions
An improvement PF scheduler was proposed for the LTE downlink. Performance analysis between the proposed scheduler and the PF scheduler were done through simulations with the LTE-Sim event-simulator platform for the LTE network. The results presented show that the PFLR scheduler performed better when compared with the PF and MLWDF schedulers to meet the delay requested by the users in different types of simulations.

7. Acknowledgement
The work reported here is supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References