Latency-Rate Downlink Packet Scheduler for LTE Networks

Fabian Maurice MALHEIROS FRANCO, Augusto FORONDA, and Emilio Carlos Gomes WILLE
Universidade Tecnológica Federal do Paraná (UTFPR), Av. Sete de Setembro 3165, 80230-901, Curitiba (PR), Brazil.
fabian@utfpr.edu.br, foronda@utfpr.edu.br, ewille@utfpr.edu.br

***Abstract*—The Long-Term Evolution standard is currently the leading technology used in mobile 4G networks. Under LTE user devices may request services available on the Internet such as voice/video streaming and Web pages. The access to such services is managed by a base station, which does resource scheduling through a multiple access network technology, ensuring Quality of Service. Hence, one of the main challenges regards responding to requests for services that require low latency and high bandwidth. Considering that resource scheduling in LTE networks is a difficult problem this work proposes the Latency-Rate Downlink Packet Scheduler (LR-DPS) for the scheduling of resources of downlink traffic, aiming to meet the maximum delay requirements for the input traffic. The proposal is divided into three hierarchical stages. In the first stage, traffic is bounded by a token bucket. In the second, time allocation and source rates are determined in order to meet the restrictions. In the third stage, the data is allocated in resource blocks in a balanced way, ensuring fairness. The results of the simulations considering different kinds of traffic show that the LR-DPS met the requirements when other known schedulers exceeded the maximum delay requested by up to 90%.**

***Index Terms*—wireless networks, channel allocation, 4G mobile communication, heuristic algorithms, streaming media.**

I. INTRODUCTION

In mobile networks, regardless of the technology used, video traffic represents over half of the data consumption and presents an annual growth of around 50% [1]. In this same scenario, social networks should also grow, but their relative traffic participation will decrease in the near future, as a result of the sharp increase in video consumption. Other application categories have lower growth rates and, thus, are decreasing in proportion compared to total traffic. The use of video incorporated to social networks and Web pages is also increasing, fed by larger device screens, higher resolutions, and new platforms that support real-time transmission.

The emergence of new applications and consumer behavior may change network traffic volume. Video streaming in different resolutions may impact data traffic consumption to a high degree. A high-definition video (1080p) usually increases the volume of data traffic by around four times compared to the same video at a standard resolution (480p). An emerging trend with the increase in the transmission of immersive video formats, such as the 360-degree video, would also affect data traffic consumption. For example, a 360-degree video in YouTube consumes four to five times more bandwidth than a normal video at the same resolution [2].

One of the primary challenges to ensure the quality of service in mobile networks of high density and limited resources is under the responsibility of the scheduling discipline, which consists in managing the packet transmission and reception queues [3]. Some schedulers have been proposed in the literature to ensure certain level of Quality of Service (QoS) for the users equipment (UE), however they do not guarantee a bounded delay for the input traffic for a specific number of UEs. In this sense, the present work proposes and evaluates the *Latency-Rate Downlink Packet Scheduler* (LR-DPS) for the scheduling of resources of downlink traffic for the Long-Term Evolution (LTE) network architecture. The main function of this scheduler is to provide the guarantee of a maximum delay for variable bit rate (VBR) traffic handled by a base station for a specific number of UEs estimated by the proposed model, even though the proposed scheduler can be applied for constant bit rate (CBR) traffic. For this purpose, the scheduler is composed of three hierarchical processing stages for incoming traffic conditioning, user data rate calculation, and resource blocks allocation, respectively.

The rest of this paper is organized as follows. Section II presents the works related to the theme. Section III presents the details of the LTE network architecture. The latency rate model is described in Section IV. Section V presents in detail the mechanisms used in the proposed scheduling algorithm. Section VI shows the settings of the simulation environment. Section VII presents the results obtained through computational simulations. Finally, conclusions and future work are presented in Section VIII.

II. RELATED WORK

To make this paper as self-contained as possible, we now present a succinct survey of works that have addressed scheduling algorithms in wireless networks. The work in [4] proposes a scheduling algorithm in the downlink direction with a guarantee of QoS. Initially, it is calculated the QoS Class Identifier (QCI) for each system user and, based on such values, compute the resource blocks necessary to meet their transmission rate requisites. Then, the users are put in decreasing priority order for resource allocation.

In [5], a solution is presented for the provision of QoS using particle swarm optimization (PSO) for downlink traffic in LTE networks. Evaluating the signal quality for each UE, the optimization seeks to guarantee different rates requested by the users and results in an allocation matrix.

The authors in [6] use a token bucket algorithm for the provision of QoS in downlink traffic for LTE networks. The

Digital Object Identifier 10.4316/AECE.2020.04007

proposed scheduling solution aims to meet the requirements of real-time video traffic and VoIP through a process divided into three stages. The first stage classifies the input flows for distinguishing the real-time services. In the second stage, a buffer manager stores the data and controls the data rate. In the third stage, the scheduler uses a token bucket to allocate real-time traffic with different rates for video and VoIP traffic, while a proportional fair scheduling algorithm handles the other traffic streams.

In [7], a packet scheduling algorithm seeks to meet QoS constraints of delay and bandwidth for real-time applications. The proposed model uses a packet prediction mechanism that consists of three phases. The first considers the frequency domain for the effective use of bandwidth; the second manages transmission queues and calculates expected delays for the packets. Lastly, the third phase occurs in the time domain just as the previous, starting a cutting process to meet the delay requirements.

The authors in [8] seek to meet the QoS constraints standardized by the 3GPP group for the QoS Class Identifier classes, at the same time they maximize the system performance in terms of fairness and throughput in the downlink direction. For this purpose, they adopt the Knapsack Algorithm in the time domain over the traffic overload patterns.

In [9], the authors propose an optimization aiming to maximize the quality of the user experience (QoE) for scheduling video-stream traffic in the downlink direction. The proposal applies an integration of Random Neural Networks (RNN) with a Genetic Algorithm (GA). The process starts with the identification of the input parameters of the network and application, such as the transmission rate and delay, as well as their respective limits. This way, the RNN must output a fitness function to maximize the QoE. Then, a possible solution goes through an evolution process using the GA, and the process continues until a stopping criterion is met or a maximum number of generations is reached.

In [10], the authors propose a scheduling process with the provision of QoS for uplink traffic using a GA. The scheduling solution is divided into three steps. In the first step, the users are organized in a list by priority and urgency of their packets. In the second step, resources are first allocated to users with packets close to a delay limit, and then other users are selected for allocation. Finally, in the third step, the resources are allocated using the allocation solution found from the genetic algorithm.

The work in [11] proposes the Channel and QoS Aware (CQA) scheduling algorithm. Its discipline prioritizes traffic considering the delay experienced by the UE, a guaranteed bitrate, and the quality of the channel in different sub-bands. CQA performs the scheduling according to different criteria in the time domain (TD) and the frequency domain (FD), aiming to reach greater spectral efficiency at the same time it seeks to satisfy the traffic delay requirements.

The above related work tries to ensure certain level of QoS to the input traffic. However, they do not guarantee the maximum delay for a certain number of UEs. Therefore, the proposed scheduler is the first to provide a delay bound and estimate the number of UEs for LTE network, which is very important to provide QoS for video traffic transmission.

III. LTE Network Architecture

The LTE network architecture consists of control components, base stations, and mobile user devices (Fig. 1). The access network is formed only by the eNodeB base station, which belongs to the *evolved UTMS terrestrial radio access network* (E-UTRAN) and provides access, through a wireless network, to the *user equipment* (UE), which may use data services. The eNodeB is responsible for the communication between the UE and the *evolved packet core* (EPC) [3].

The logic components of the EPC are described next [12]:

- *Serving Gateway* (SGW): its primary functions include the routing and forwarding of packets, as well as the necessary support so that UEs move in areas serviced by different eNodeBs;
- *Packet Data Network Gateway* (PGW): its main function includes the control of the traffic exchange with the external network;
- *Mobility Management Entity* (MME): its main functions include the signaling and control functions to manage the access of the UE to the network connections, such as the recording of network resources and the mobility management function.

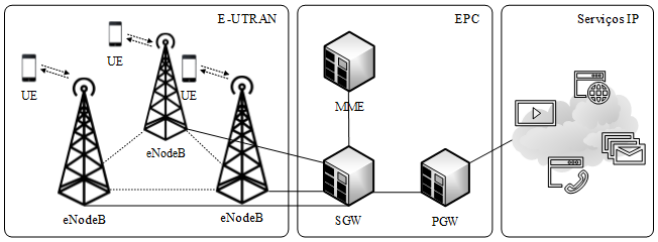

Figure 1. LTE network architecture

The E-UTRAN is formed by a series of eNodeB base stations distributed non-hierarchically and usually connected among each other. The eNodeB is responsible for managing the radio resources, which include the carrier control, admission control, connection mobility, and resource allocation for the UEs [12]. Resource allocation occurs in the uplink direction for receiving the data transmitted by the UE, and in the downlink direction for transmission of data to the UE [13]. Each subframe of one millisecond duration uses the frequency base band (typically 1.4-20MHz) which is further divided into orthogonal narrow-band subcarriers. Resource blocks (RBs), which are the minimum allocation units, are composed of subcarriers [14]. LTE downlink scheduling is run in subframes by allocating RBs.

The LTE downlink scheduler decides how RBs are allocated to UEs. The number of UEs connected to the eNodeB and the channel qualities between eNodeB and UEs dynamically change due to the UEs' mobility and the characteristics of the wireless medium such as multipath fading, shadowing, etc.

IV. Latency Rate Model

The scheduling model proposed in this work aims to meet the quality of service of individual transmissions. For this purpose, we use the *Latency Rate* (LR) [15] model together with a token bucket mechanism. This approach allows calculating constrained limits in the end-to-end delay of individual sessions. Hence, as a measuring parameter of quality of service, we use the meeting of a maximum delay

for each concurrent link in an eNodeB.

The LR server theory, proposed in [15], allows calculating limits for the maximum delay in data communication networks. The server nomenclature is used to depict a combination of a scheduler and a transmitter that exists in an output port of a base station or router. Such servers can support different scheduling disciplines and different traffic models. In the case of an LR scheduler, its behavior is determined by two parameters: *latency* and *allocated rate*. All the servers that guarantee rates to their clients display this property and, thus, may be modeled as LR servers. The latency of an LR server may be considered as the worst case of delay of the first packet in the period occupied by a flow. In general, the latency parameter depends on the scheduling algorithm implemented, as well as on the allocated rate and the traffic parameters of the session that is being serviced. For a specific scheduling algorithm, parameters such as the transmission rate in the output link, the number of sessions that share the link, and the attributed rates may influence the latency.

Let $A_i(t)$ be the incoming traffic. The maximum delay on the delivery of a data packet is measured between the moment this packet is received by the LR server and its transmission. This delay considers the receiving time of the first bit in a data packet until the transmission of its last bit. Thus, the delay D_i for traffic i has an upper bound according to Eq. (1) where r_i is the data rate, $Lmax_i$ is the maximum packet size, σ_i is the maximum token bucket size and θ_i is the latency [13],

$$D_i \le \frac{\sigma_i}{r_i} + \theta_i - \frac{L\max_i}{r_i} \tag{1}$$

An LR server can provide a maximum requested delay $Dmax_i$ for each traffic following Eq. (2),

$$\frac{\sigma_i}{r_i} + \theta_i - \frac{L\max_i}{r_i} \le D\max_i \tag{2}$$

In this way, an LR packet scheduler can ensure a delay limit if the input traffic is regulated by a token bucket, as shown in Fig. 2. Let R be the transmission rate of the physical medium, thus the time necessary for transmitting the data for traffic i, is given by Eq. (3), where T_T is the total time to allocate all the users,

$$\theta_i = T_T + \frac{L\max_i}{R} \tag{3}$$

According to Foronda et al. [16, 17], the constraint regarding the maximum delay is given by:

$$\frac{T_T.(\sigma_i - L\max_i)}{r_i.T_T - L\max_i} + T_T + \frac{L\max_i}{R} \le D\max_i \tag{4}$$

By isolating r_i in Eq. (4), we have Eq. (5),

$$\frac{\sigma_i - L\max_i}{D\max_i - \frac{L\max_i}{R} + T_T} + \frac{L\max_i}{T_T} \le r_i \tag{5}$$

Furthermore, Eq. (6) allows calculating the rate allocated by the server and the time to allocate all the users. It is verified, therefore, that the rate of the token bucket (ρ_i) added to the rate of transmitting a packet of size $Lmax_i$ must be smaller than the rate allocated by the server.

$$\rho_i + \frac{L\max_i}{T_T} \le r_i \tag{6}$$

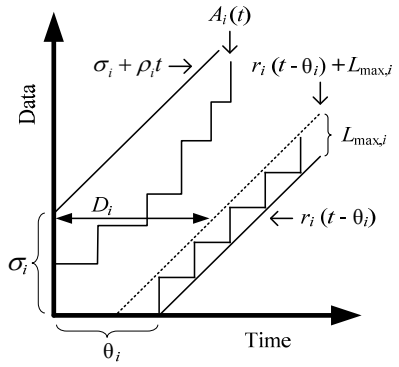

Figure 2. Latency Rate scheduler model

V. Latency Rate Downlink Packet Scheduler

The architecture of the resource scheduler for LTE networks proposed in this work is presented in Fig. 3. Starting from requirements by the UEs, traffic sources forward data to the E-UTRAN access network and a maximum delay must be met for each type of traffic. In this case, the maximum delay required for a traffic is identified by $D\max_i$ and must be known beforehand by the eNodeB. When data are received by the eNodeB, they are stored in the buffer of the *radio link control* (RLC) layer and submitted to a token bucket algorithm. Then the total time (T_T) for the transmission of all the traffic is calculated based on the bandwidth configuration of the eNodeB and the reception capacity of the UEs. Hence, the total number of resource blocks groups (TRBG) to be transmitted in interval T_T is determined. The TRBG takes into account the transmission capacity per RBG (*RBGmax*) based on the information of the channel quality indicator. Finally, an allocation module selects the traffic that will be forwarded and allocates them in RBGs in the time/frequency space for transmission via a wireless network interface.

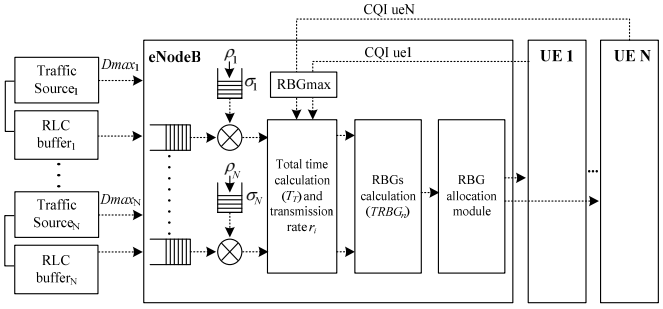

Figure 3. Architecture of the resource scheduler LR-DPS

In this model, a token bucket algorithm limits the incoming traffic, and the LR model provides an allocation rate for a maximum delay requested. The incoming traffic received by the eNodeB is requested by an UE. The requirement, in this specific case, is met by a traffic source available on the external network, which performs the transmission of data through the LTE network core (EPC).

The application of the proposed model may be segmented into three stages described in the following. Among the stages mentioned, the second and third ones refer specifically to the application of the scheduling model in LTE networks.

A. Stage 1 – Application of the Token Bucket

The token bucket controls arriving packets as follows. Upon arrival, a packet will be sent out with the token bucket size decreased by the packet size in bytes provided there are enough tokens for the packet. In our model, the token bucket always has enough tokens, and the arrival packets are not discarded. Upon being received by the eNodeB the incoming traffic is remodeled, being limited by the size and the rate of the token bucket. Thus, the first stage refers to the treatment of the incoming traffic by the token bucket mechanism and is subject to Eq. (7), where σ_i corresponds to the size of the token bucket and ρ_i represents the rate of the token bucket in interval t for each traffic,

$$A_i(t) \le \sigma_i + \rho_i t \tag{7}$$

With the application of the token bucket for each incoming traffic A_i, we have a distinct $\sigma_i + \rho_i$ set, which represents a maximum limit for the incoming traffic. The size of the token bucket must accommodate the data received from the traffic source and the rate, which must be sufficient to forward the data without packet losses.

B. Stage 2 – Total Time Allocation

Each UE must satisfy two constraints regarding the transmission rate. Also, if the transmission rates are minimal, a higher number of UEs may be allocated in the spectrum. Therefore, in this work, Stage 2 is formulated as an optimization problem. The objective of the Total Time Allocation (TTA) problem is to determine the total time (T_T) that minimizes the sum of all data rates (r_i) of the UEs. Clearly the unknown variables are: T_T and r_i. Let N be the number of UEs. The TTA problem is formulated as follows:

$$r_{TTA} = \min_{T_T} \sum_{i \in N} r_i \tag{8}$$

subject to:

$$\rho_i + \frac{RBG\max_i}{T_T} \le r_i, \quad i \in N \tag{9a}$$

$$\frac{\sigma_i - RBG\max_i}{D\max_i - \frac{RBG\max_i}{R} + T_T} + \frac{RBG\max_i}{T_T} \le r_i, \quad i \in N \tag{9b}$$

$$0 < T_T < D_l \tag{9c}$$

Eq. (8) is the objective function. Eqs. (9a-b) correspond to the constraints regarding the data rate of the stations. Eq. (9c) is the constraint regarding the total time to allocate all the UEs with active traffic, where $D_l = min_{i \in N}(Dmax_i)$. One may observe that $r_{TTA} \le R$, where R is the transmission rate of the system. With the objective function being linear and the constraint set being convex, the TTA problem may be classified as a constrained multivariable convex optimization problem. The convex condition ensures the existence of a single solution, i.e., a global optimum. In general, it is a problem of difficult analytical solution. In our scheduler, T_T and r_i are estimated again if new stations enter (leave) the network or if some station requires a new maximum delay.

1) Solving the TTA problem

In particular, if all UEs have the same characteristics, the solution of the TTA problem is straightforward and it is found by equating constraints (9a) and (9b). Therefore, the optimal total time is given by Eq. (10) - where we drop the subscript *i*:

$$T_T = D\max - \frac{RBG\max}{R} - \frac{\sigma - RBG\max}{\rho} \tag{10}$$

and the optimal data rate corresponds to:

$$r = \rho + \frac{RBG\max}{T_T} \tag{11}$$

In this case, $r_{TTA} = N.r$, and the maximum number of UEs that may be allocated corresponds to $\lfloor R/r \rfloor$. However, in a more general case (UEs with different characteristics), it is of utmost importance that the TTA problem be solved in polynomial time. In this sense, this work suggests a simple heuristic to obtain an approximate solution for the problem. This heuristic, denominated *decomposition heuristic* (proposed in Wille et. al [18, 19]), corresponds to a subdivision of the original problem into N problems with easy solutions and the subsequent final composition, as described next. Let $T_T^{(j)}$ be the solution for the total time considering that only session j is active. This value is given by Eq. (12),

$$T_T^{(j)} = D\max_j - \frac{RBG\max_j}{R} - \frac{\sigma_j - RBG\max_j}{\rho_j}, \quad j \in N \tag{12}$$

Let $r_i^{(j)}$ be the data rate of source i obtained when considering the total time $T_T^{(j)}$. Such values are given by:

$$r_i^{(j)} = \max\left\{ \rho_i + \frac{RBG\max_i}{T_T^{(j)}}, \frac{\sigma_i - RBG\max_i}{D\max_i - \frac{RBG\max_i}{R} + T_T^{(j)}} + \frac{RBG\max_i}{T_T^{(j)}} \right\}, \quad i \in N, j \in N \tag{13}$$

Therefore, the optimal total time (T_T^*) will be the one that produces the smallest total data rate according to:

$$T_T^* = \left\{ T_T^{(j)} \mid \arg\min_{j \in N} \sum_{i \in N} r_i^{(j)} \right\} \tag{14}$$

Finally, we noticed that the computational effort of the decomposition heuristic required to find solutions to the TTA problem is $O(N)$.

C. Stage 3 – Resource Block Group Allocation

In the LTE network architecture, the services used in the downlink direction generate data flows that are forwarded from the EPC to the eNodeB. Upon receiving such data, the scheduler is in charge of allocating them in the form of resource block groups. The configuration of the RBGs meets the definition of *Resource Allocation Type 0*, which determines the number of blocks per group in function of the bandwidth available for data transmission. Next, the RBGs are allocated to the UE by the physical layer,

respecting the division by time and frequency defined in this stage by the resource allocation module.

Initially, the values of r_i and T_T, calculated in Stage 2, are used to calculate TRBG. This value represents the total number of RBGs allocated for a UE in an interval T_T, and is defined by Eq. (15), where $RBGmax_i$ represents the transmission capacity in bits per RBG for UE_i,

$$TRBG_i = \frac{r_i.T_T}{RBG\max_i} \quad (15)$$

Therefore, we have a set of TRGBs distributed in the time and frequency domains for interval T_T that ensure the r_i rate and, consequently, the requested delay $Dmax_i$.

Once the set of TRBGs is found, the allocation process described in *Algorithm 1* is carried out sequentially and cyclically to promote fairness for all users. The algorithm runs until all the traffic is transmitted. From the set of TRBGs, we identify among the requested traffic streams the smallest TRBG value for allocation in total time T_T, i.e., the one that requires the lowest number of RBGs for transmission. This value, identified as TR, is used as a reference for the allocation of all traffic. Hence, for each iteration of the scheduler, all the UEs have their data allocated in RBGs and, for each UE, some RBGs are allocated proportionally to the value of TR. Thus, when the $TRBG_i$ value is equal to that of TR, one RBG is allocated for the selected UE, given that the proportion will be of 1:1.

The proportional amount of RBGs allocated at each iteration is identified by A_{RBG} (Allocated RBG). However, it should be considered that the proportion may result in non-integer values and in order for the allocation to be done efficiently the RBGs must be filled in full whenever possible. In this case, the integer values are allocated to each iteration, and if there is traffic in the transmission queue, the fractional values are assigned to an overrun variable $S_{RBG,i}^{(n)}$.

Algorithm 1 Allocation Module

Algorithm parameters:

S_{RBG}: RBG overrun variable for allocation.
A_{RBG}: Allocation of contiguous RBGs for the iteration.
$N_{\text{UE:}}$ Number of UEs to be allocated.

Input Data:

CTRBG [ ] = Set of TRBGs.

Algorithm:

1: *TR* = The smallest TRBG in CTRBG.
2: $S_{RBG,i}^{(0)} = 0$
3: **For each** iteration i = 1 to TR
4: **For each** UE_i to N_{UE}
5: $A_{RBG,i}^{(n)} = \text{int}\left[\frac{TRBG_i}{TR} + S_{RBG,i}^{(n-1)}\right]$
6: $S_{RBG,i}^{(n)} = frac\left[\frac{TRBG_i}{TR} + S_{RBG,i}^{(n-1)}\right]$
7: Allocate $A_{RBG,i}$ for the selected UE
8: **End for each**
9: **End for each**

Output:

Allocation of UEs in the time/frequency space for T_T.

Thus, with each new iteration, the overrun variable is incremented with decimal values until its value is equal to or greater than 1, that is, the overrun variable can occupy a RBG in full. When there is no more traffic in the transmission queue, the overrun data is sent in the last iteration. Finally, it is observed that the value of TR also corresponds to the number of iterations necessary to allocate all the RBGs in the T_T interval; however, Algorithm 1 is executed whenever there is data for transmission.

VI. Simulation Settings

Considering the factors of simplicity, reliability, documentation and customization, the Ns-3 (version 3.26) was chosen as the instrument to validate the proposed model. Ns-3 is a discrete event simulator for network systems, aimed primarily for educational and research purpose since it addresses several data communication technologies such as Ethernet, LTE, Wi-Fi, WiMAX, among others. It is a free software written in C++ and publicly available under the GNU GPLv2 license for use, research, and development [19].

The setting of the parameters used in the simulation scenario is shown in Table I. Such parameters were configured considering the *LTE-EPC Network Simulator* (LENA) module [21, 22].

TABLE I. Ns-3 Simulation Parameters

Parameter	Value
eNodeB	1
UEs	30, 40, 50
eNodeB TX Power	46 dBm
Attenuation Model	EPA 3 km/h
Distance	500 m
Type of Traffic	Video (VBR)
Number of Traffic Streams	2
Simulation Time	30 s
Number of Simulations	50
Bandwidth	5 MHz, 10 MHz, 15MHz
Modulation Coding Scheme	28
RBs	25
RBG Size	2 RBs
TTI (Transm. Time Interval)	1 ms
RB Bandwidth	180 kHz
Modulation	64-QAM
Duplexing Mode	FDD

Two distinct traffic traces (each with 30s long) were considered for simulations. They were acquired from a variable bit rate video, generated based on *EvalVid* [23], a framework for the evaluation of the quality of videos transmitted through a real or simulated communication network. In order to use it in Ns-3, it was necessary to configure the *evalvid-ns3* module, which enables the transmission of video from a client/server model.

Two schedulers were selected for comparison purposes: *Round Robin* (RR) [24, 25], due to its simplicity and widespread utilization; and *Channel and QoS Aware* (CQA) [10], for having a particularity that seeks to minimize the delay in packet delivery for voice and video traffic. Both have an implementation in the Ns-3 simulator.

VII. Evaluation Results

In this section, we present an exhaustive series of tests demonstrating the effectiveness of our proposal. First, we consider a scenario with 30, 40 and 50 UEs. One UE requests traffic 1 and the others UEs request traffic 2. The token bucket parameters (obtained in Stage 1) are shown in Table II, as well as the time and rate attribution (Stage 2) are

also shown (considering *RBGmax* = 1480 bits and *R* = 17760 kbps). In order to obtain exact results, we opted to use the *fmincon* function of MATLAB® software to solve the TTA problem (in practical use one must resort to the decomposition heuristic).

TABLE II. PARAMETERS FOR TWO DIFFERENT TRAFFIC STREAMS

Token bucket parameters		
	Traffic 1	**Traffic 2**
Token bucket size (bits)	30560	49472
Token bucket rate (kbps)	306	680
Time and rate attribution		
Dmax (ms)	150	150
T_T (ms)	55	55
r_i (kbps)	333.28	706.91
TRBG	12.39	26.27

We considered as metrics of interest the maximum delay and throughput and, as variable parameters, the number of UEs and bandwidth. To determine the delay generated in the scheduling process, we consider the time interval that the eNodeB consumes to deliver a specific packet to a UE. This measurement is performed by the analysis of the *Packet Data Convergence Protocol* (PDCP) header data considering the worst case, i.e., we consider the longest delay that occurred in the entire transmission to confirm if the delay requested by the user was met.

A. Maximum Delay for Different Number of UEs

Fig. 4 shows that with the utilization of LR-DPS it is possible to guarantee the maximum delay requested for two distinct VBR traffic streams. We observe that, in the most demanding situation (50 UEs), the results demonstrate that the RR scheduler exceeded the limit of 150 ms by 81% and the CQA scheduler by 90%. Therefore, one may notice a significant advantage of the proposed scheduler compared to the RR and CQA schedulers insofar as the density of UEs increases to 40 and 50 UEs, and that the maximum delay is met for the cases presented in Fig. 4.

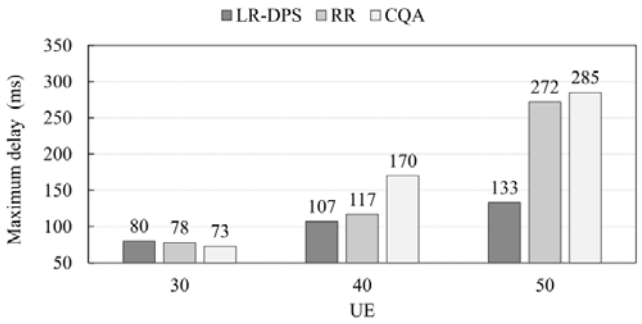

Figure 4. Maximum delay for two VBR traffic in a bandwidth of 5 MHz

The way that traffic 2 is treated in the scheduling process has evidenced the advantages of the proposed model. Once the incoming traffic is known and modeled by a token bucket, parameterized by a mathematical model, we have as a practical result the reservation of more RBs for UEs with more data to be allocated in the same period. On the other hand, if the process is carried out agnostically in regards to the incoming traffic, different traffic streams will have the same data rate, resulting in longer delays for traffic sources with higher rates.

B. Minimum and Average Throughput

Observing the data received in the PDCP layer of the UEs, we measured the minimum throughput value, which corresponds to the lowest data flow verified in the downlink direction in a given period. In this scenario, the calculation is performed for 1-second intervals when, for each UE, thirty measurements are carried out referring to the thirty seconds of transmission. Then, we select the smallest value among the measurements. Fig. 5 shows that the three schedulers do not have a significant difference in minimum flow when compared to each other and when there are different UE densities.

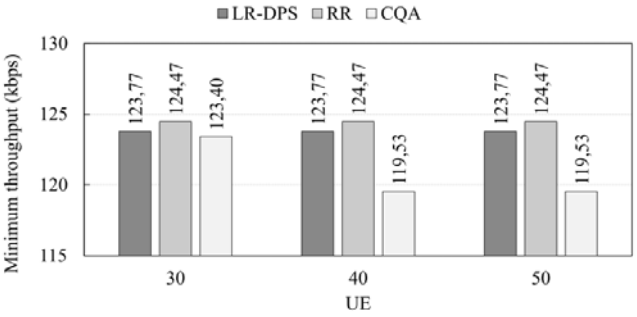

Figure 5. Minimum throughput for two VBR traffic (5 MHz)

From the same simulation, we extracted the average flow values with three different densities. The calculation, just as for the minimum throughput, considers the flow measures from 1-second intervals, from which the average values found for each UE density are calculated. The values present in Fig. 6 show that the average flow values remained close, just as the minimum values verified. The values observed for average flow present a small drop in the rate as the number of users increase. This is because, in the case of finite resources for allocation, fewer resources are allocated per user in the same period if more UEs are disputing the transmission medium.

C. UEs Allocated in Different Bandwidth Settings

From the results obtained for different bandwidths, it is possible to analyze the limits of users serviced, taking into consideration a maximum delay of 150 ms for 5, 10, and 15 MHz. The plot in Fig. 7 shows that, for the three bandwidths tested (5 MHz, 10 MHz, and 15 MHz) the proposed scheduler allocated more users compared to the RR and CQA schedulers.

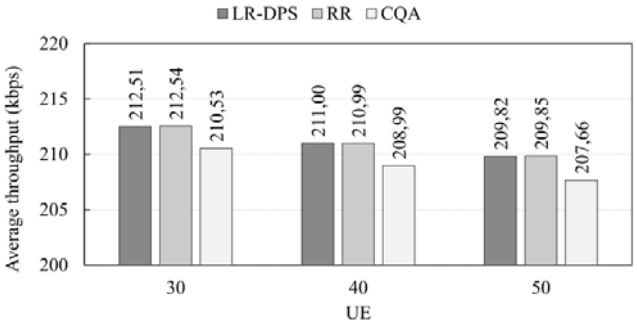

Figure 6. Average throughput for two VBR traffic (5 MHz)

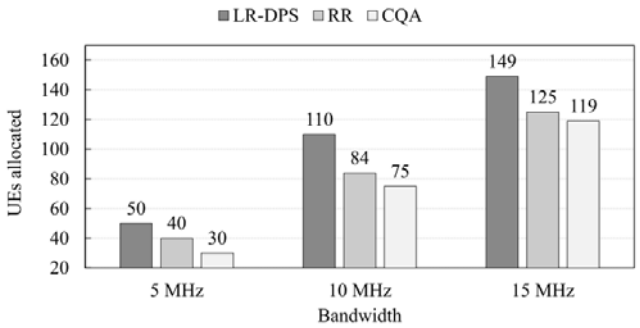

Figure 7. Maximum number of UEs allocated for two VBR traffic

D. Delay and Fairness for Different Traffic Proportions

Considering that different kinds of traffic result in distinct setting parameters for the LR-DPS scheduler, we sought to study its performance in three test scenarios with different traffic proportions. The values used for the three proposed tests may be observed in Table III. In the first test, 42 UEs request traffic 1, while 5 UEs request traffic 2. This proportion changes to the other two tests.

TABLE III. PARAMETERS FOR TWO DIFFERENT TRAFFIC STREAMS

Test	1		2		3	
Traffic	1	2	1	2	1	2
UEs	42	5	32	10	21	15
r_i (kbps)	333.28	706.91	333.28	706.91	333.28	706.91
TRBG	12.39	26.27	12.39	26.27	12.39	26.27
T_T (ms)	55		55		55	
R (kbps)	17532		17734		17603	

The performance of LR-DPS is compared with that of the RR and CQA schedulers. Fig. 8 shows that the maximum delay of 150 ms was only met using the LR-DPS scheduler in the three proposed tests.

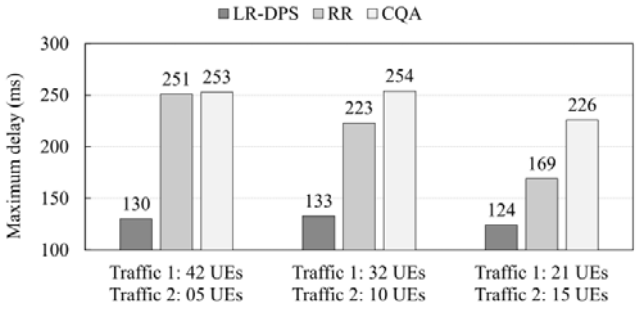

Figure 8. Maximum delay for different traffic settings requested (5 MHz)

The LR-DPS looks for a better distribution of the individual maximum delays through the manner of distribution of the RBGs in Stage 3.

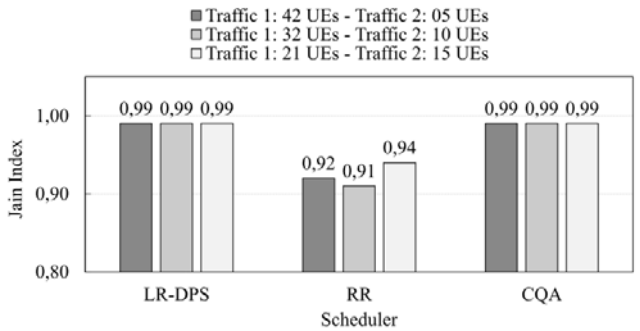

Figure 9. Fairness index for the individual delay in different traffic

Hence, this same test scenario is used to verify the fairness among the three schedulers and validate one of the aims of Stage 3. For this purpose, we use Jain's index [26], which varies between zero and one, where one represents the case of greatest fairness. From the maximum individual delays measured in each test scenario, we confirmed that the LR-DPS and CQA schedulers presented better results, as shown in Fig. 9.

E. UEs Allocated in Different Requested Delays

Lastly, we analyzed the performance of the LR-DPS scheduler for different delay requirements, namely 100 ms and 150 ms for the maximum delay requested. Thus, we examined three test scenarios in which the numbers of UEs that request a maximum delay of 100 ms are 10, 15, and 20, and all the others UEs request a service with 150 ms. The values for the three tests with traffic 1 may be observed in Table IV.

TABLE IV. PARAMETERS FOR TWO DIFFERENT DELAYS

Test	1		2		3	
Dmax (ms)	100	150	100	150	100	150
UEs	10	36	15	30	20	24
r_i (kbps)	474.42	352.25	461.09	357.03	450.34	362.92
TRBG	10.26	7.62	9.03	7.00	7.91	6.38
T_T (ms)	32		29		26	
R (kbps)	17425		17627		17717	

Fig. 10 presents the obtained results considering the two constraints of maximum delay. The results show that, for the three proposed tests, the LR-DPS scheduler met the Dmax of 100 ms and 150 ms. Considering a transmission rate of 17760 kbps for the bandwidth of 5 MHz, we verified that it was possible to service 46 UEs for test 1, 45 UEs for test 2, and 44 UEs for test 3.

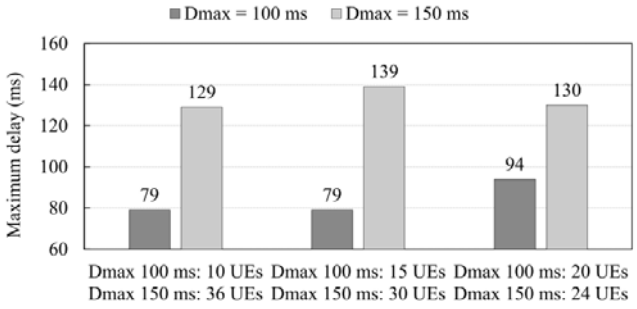

Figure 10. Maximum delay for different configurations of maximum delay

VIII. CONCLUSION AND FUTURE WORK

In this paper, we have presented a new packet scheduling algorithm, named LR-DPS, for traffic in the downlink direction in LTE networks. The main function of this scheduler is to provide the guarantee of a maximum delay for variable bit rate (VBR) traffic handled by a base station. For this purpose, the scheduler is composed of three hierarchical processing stages. Stage 1 has the role of conditioning incoming traffic; Stage 2, of calculating a data rate to guarantee a maximum delay requested; and stage 3 performs the allocation of the traffic in resource blocks.

To verify the performance of LR-DPS, we built scenarios using network simulator Ns-3 which, with the configuration of additional modules, enabled the simulation of VBR video traffic transmission, bringing more realism to the simulations. We chose the RR and CQA schedulers for

comparison purposes, the first for being a widely known reference and the latter for having characteristics closer to the proposal of this work. The simulations carried out explored several configuration sets of the simulation environment, therefore ensuring the evaluation of the proposed model in all its possibilities.

Analyzing the results got through the simulations, we have found that the LR-DPS model met its purpose and guaranteed the maximum delay requested in several situations. Lastly, we verified that the number of users in the system was more substantial compared to the other two schedulers analyzed (RR and CQA).

Future studies include the development and the analysis of a more extensive group of simulations considering a mix of traffic (video, voice, data), and the study of approaches which may simplify the working of Stage 3 of the proposal.

REFERENCES

[1] Cisco Visual Networking Index, 2018. Global mobile data traffic forecast update 2017-2022. White Paper.

[2] S. Afzal, J. Chen, and K. K. Ramakrishnan, "Characterization of 360-degree videos," in Proceedings of the Workshop on Virtual Reality and Augmented Reality Network - VR/AR Network '17, Los Angeles, CA, USA, 2017, pp. 1-6, doi:10.1145/3097895.3097896.

[3] F. Capozzi, G. Piro, L.A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: key design issues and a survey," IEEE Commun. Surv. Tutorials, vol. 15, no. 2, pp. 678-700, 2013, doi:10.1109/SURV.2012.060912.00100.

[4] Na Guan, Y. Zhou, Lin Tian, Gang Sun, and Jinglin Shi, "QoS guaranteed resource block allocation algorithm for LTE systems," in 2011 IEEE 7th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Shanghai, China, Oct. 2011, pp. 307-312, doi:10.1109/WiMOB.2011.6085383.

[5] Lin Su, Ping Wang, and Fuqiang Liu, "Particle swarm optimization based resource block allocation algorithm for downlink LTE systems," in 2012 18th Asia-Pacific Conference on Communications (APCC), Jeju, Korea (South), Oct. 2012, pp. 970-974, doi:10.1109/APCC.2012.6388227.

[6] Liqun Zhao, Yang Qin, Maode Ma, Xiaoxiong Zhong, and Li Li, "QoS guaranteed resource block allocation algorithm in LTE downlink," in 7th International Conference on Communications and Networking in China, Kun Ming, Aug. 2012, pp. 425-429, doi:10.1109/ChinaCom.2012.6417520.

[7] W. K. Lai and C.-L. Tang, "QoS-aware downlink packet scheduling for LTE networks," Computer Networks, vol. 57, no. 7, pp. 1689-1698, May 2013, doi:10.1016/j.comnet.2013.02.017.

[8] N. Ferdosian, M. Othman, B. M. Ali, and K. Y. Lun, "Throughput-aware resource allocation for QoS classes in LTE networks," Procedia Computer Science, vol. 59, pp. 115-122, 2015, doi:10.1016/j.procs.2015.07.344.

[9] T. Ghalut, H. Larijani, and A. Shahrabi, "QoE-aware optimization of video stream downlink scheduling over LTE networks using RNNs and genetic algorithm," Procedia Computer Science, vol. 94, pp. 232-239, 2016, doi:10.1016/j.procs.2016.08.036.

[10] S. H. da Mata and P. R. Guardieiro, "Resource allocation for the LTE uplink based on genetic algorithms in mixed traffic environments," Computer Communications, vol. 107, pp. 125-137, Jul. 2017, doi:10.1016/j.comcom.2017.04.004.

[11] B. Bojovic and N. Baldo, "A new channel and QoS aware scheduler to enhance the capacity of voice over LTE systems," in 2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14), Castelldefels-Barcelona, Spain, Feb. 2014, pp. 1-6, doi:10.1109/SSD.2014.6808890.

[12] C. Cox, "An Introduction to LTE: LTE, LTE-Advanced, SAE, VoLTE and 4G mobile communications", 2nd Ed., pp. 21-46, Wiley, 2014.

[13] E. Dahlman, S. Parkvall, J. Skold, "4G, LTE – Advanced pro and the road to 5G," 3rd Ed., pp. 75-91, Academic Press, 2016.

[14] Harri Holma, Antti Toskala, "LTE for UMTS: OFDMA and SC-FDMA based radio access," ISBN: 978-0-470-74547-2, April, 2009.

[15] D. Stiliadis and A. Varma, "Latency-rate servers: a general model for analysis of traffic scheduling algorithms," IEEE/ACM Trans. Networking, vol. 6, no. 5, pp. 611-624, Oct. 1998, doi:10.1109/90.731196.

[16] A. Foronda, Y. Higuchi, C. Ohta, M. Yoshimoto, and Y. Okada, "Service interval optimization with delay bound guarantee for HCCA in IEEE 802.11e WLANs," IEICE Transactions on Communications, vol. E90-B, no. 11, pp. 3158-3169, Nov. 2007, doi:10.1093/ietcom/e90-b.11.3158.

[17] Y. Higuchi, A. Foronda, C. Ohta, M. Yoshimoto, and Y. Okada, "Delay guarantee and service interval optimization for HCCA in IEEE 802.11e WLANs," in 2007 IEEE Wireless Communications and Networking Conference, Kowloon, China, 2007, pp. 2080-2085, doi:10.1109/WCNC.2007.390.

[18] E. C. G. Wille, M. Mellia, E. Leonardi, and M. Ajmone Marsan, "A Lagrangean relaxation approach for QoS networks CFA problems," AEU - International Journal of Electronics and Communications, vol. 63, no. 9, pp. 743-753, Sep. 2009, doi:10.1016/j.aeue.2008.06.006.

[19] E. C. G. Wille and C. R. da C. Bento, "Metaheuristic methods for solving the capacity and flow assignment problem in TCP/IP networks," IEEE Latin America Transactions ,vol. 9, no. 5, pp. 851-859, Sep. 2011, doi:10.1109/TLA.2011.6031000.

[20] I. Bisio, S. Delucchi, F. Lavagetto, M. Marchese, G. Portomauro, S. Zappatore, "An Ns-3 based simulative and emulative platform." In: Modeling and Simulation of Computer Networks and Systems. pp. 555-575. Editor(s): M. S. Obaidat, P. Nicopolitidis, F. Zarai. 2015. doi:10.1016/B978-0-12-800887-4.00019-5.

[21] N. Baldo, M. Requena, J. Nin and M. Miozzo, "A new model for the simulation of the LTE-EPC data plane," In: Proc. of Workshop on NS-3 (WNS3'12), 2012.

[22] N. Baldo, M. Miozzo, M. Requena-Esteso, and J. Nin-Guerrero, "An open source product-oriented LTE network simulator based on ns-3," in 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems - MSWiM'11, Miami, Florida, USA, 2011, p. 293, doi:10.1145/2068897.2068948.

[23] J. Klaue, B. Rathke, A. Wolisz. "EvalVid - A framework for video transmission and quality evaluation". In: Kemper P., Sanders W.H. (eds), Computer Performance Evaluation. Modelling Techniques and Tools. Lecture Notes in Computer Science, vol. 2794. Springer, Berlin, Heidelberg. 2003. doi:10.1007/978-3-540-45232-4_16.

[24] J. B. Nagle, "On packet switches with infinite storage," In: Innovations in Internetworking. Artech House, 1988, pp. 136-139.

[25] H. Zhu and R. Hafez, "Scheduling schemes for multimedia service in wireless OFDM systems," IEEE Wireless Communications, vol. 14, no. 5, pp. 99-105, Oct. 2007, doi:10.1109/MWC.2007.4396949.

[26] R. Jain, W. Dah-Ming, W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," ACM Transactions on Computer Systems, 1984. https://arxiv.org/abs/cs/9809099.