A New Channel and QoS Aware Scheduler Algorithm for Real-time and Non-real-time Traffic in 5G Heterogeneous Networks

Gabriel A. Queiroz, and Éderson R. da Silva

***Abstract*—5G mobile communication systems have increasing demands related to Quality of Service (QoS) parameters integrated with high user densification in heterogeneous network scenarios. In this sense, 5G networks are expected to handle a wide range of applications and services. Therefore, scheduling algorithms that can benefit users of real-time (RT) and non-real-time (NRT) applications are necessary. In this sense, the main novelty of this work is the proposal of a new Channel and QoS Aware Scheduler (CQAS) in a heterogeneous network with multiple traffic models: full buffer (IoT), HTTP, vehicular, VoIP, gaming, and video. System-level simulations are carried out to analyze the performance of the CQAS and compare it to Round Robin (RR), best Channel Quality Indicator (CQI), and QoS Aware Scheduler (QAS) varying the number of users to stress test the network. The results show that CQAS presents significant overall throughput gains except for HTTP. For video users, CQAS achieves gains of up to 16.3%, 21.2%, and 163.6% when compared to QAS, best CQI and RR. As for the NRT applications, CQAS shows throughput gains over QAS of between 165.6% and 171.2% depending on the number of users. In addition, it meets the delay constraints of the 5G RT applications while performing well in reliability and the fairness index so that it outperforms the other algorithms overall.**

Link to graphical and video abstracts, and to code: https://latamt.ieeer9.org/index.php/transactions/article/view/8933

***Index Terms*—Channel and QoS Aware Scheduler, scheduling algorithm, real-time traffic, non-real-time traffic, 5G heterogeneous networks.**

I. Introduction

The demand for mobile communication services is growing as new technologies and applications are introduced into cellular networks, bringing with them the integration of more users and devices connected to operators' infrastructures. Thus, 5G systems encompass technologies such as Device-to-Device (D2D), Machine-to-Machine (M2M), Internet of Things (IoT), Vehicle-To-Everything (V2X), among others, which make up a highly connected world [1].

According to [2], 5G and beyond-5G systems must deal with this significant increase in users, services, and applications by implementing more efficient networks with higher data rates, greater spectral and energy efficiency, reduced latency, and increased network capacity. Of particular note is the capacity of cellular communication systems, which can be attributed to three main factors: the increase in the number of mobile infrastructure nodes, the growth of spectrum use, and greater channel efficiency [3].

Modern cellular systems evolved into a complex ecosystem consisting of base stations (BSs) that operate with different output powers and antenna locations due to the diverse cell sizes (i.e. macro, micro, pico, and femto), and variety of access technologies [4]. This description of a non-homogeneous environment is commonly characterized as a heterogeneous network (HetNet), and it is essential for 5G systems to deal with the diversification of the demands placed on the network.

The factors mentioned above can be described under the concept of network densification, which is divided into spatial densification and spectrum aggregation [5]. Spatial densification is achieved by increasing the number of antennas per node and the density of base stations deployed per region while ensuring a uniform distribution of users among all base stations. In turn, spectrum aggregation refers to the use of larger amounts of the electromagnetic spectrum, involving the millimeter wave bands (30-300 GHz). Both factors have an impact on increasing network capacity, which can be analyzed using the capacity equation for an additive white Gaussian noise (AWGN) channel:

$$C = m\left(\frac{W}{n}\right)\log_2\left(1+\frac{S}{I+N}\right). \quad (1)$$

In (1), W represents the bandwidth of the base station, the load factor parameter n indicates the number of users sharing the base station in question, the spatial multiplexing factor m describes the number of spatial streams between the base station and user devices, S is the desired signal power, while I and N characterize, respectively, interference and noise power at the receiver. An analysis of (1) shows how network densification, a characteristic of HetNets, impacts channel capacity and, consequently, total network capacity.

As such, improving the system's capacity means that several mobile terminals with different access techniques can coexist. To support massive connectivity, the various users and different power levels are distributed in a short space through multiple small cells, resulting in a denser network structure. In addition, the distribution of small cells reduces uncovered areas and expands the communication range due to the development of access points in areas with poor channel

G. A. Queiroz, and É. R. da Silva are with Federal University of Uberlândia, Minas Gerais, Brazil (emails: gabriel.andrade@ufu.br, and ersilva@ufu.br).

quality. Finally, small cell implementation in a wide-area communication scenario reduces link loss and delays between users and base stations, as the backhauling signals can be reached through a small path loss [6].

In addition to the complex structure of HetNets, there is a heterogeneity of applications and services offered to users, which involves their corresponding traffic models [7]. Each traffic model must meet criteria regarding throughput, latency, and data loss by Quality of Service (QoS) requirements. Thus, effective management of radio resources remains essential in network configuration, since the use of precise resource allocation, particularly scheduling techniques, configures the efficient use of resources, including bandwidth, and power of antennas, while mitigating interference between cells and users and ensuring that QoS requirements are met, increasing the quality of experience for heterogeneous users [3].

In this sense, in a wireless scenario, the packet schedulers have the fundamental role of maximizing spectral efficiency through an effective resource allocation policy that reduces or makes insignificant the impact of variations in channel quality [8]. However, scheduling algorithms such as Round Robin (RR) and Best Channel Quality Indicator (best CQI) do not consider buffer state information, so the study of Channel and Quality of Service Aware schedulers is necessary to meet real-time (RT) and non-real-time (NRT) requirements according to 5G and beyond-5G applications [9]. In addition, the proposed algorithm considers both channel conditions and QoS requirements, testing the variation in the number of users to approximate a diverse HetNet scenario such as current mobile networks.

The main contribution of this study is the proposal of a new Channel and QoS Aware scheduling algorithm with a better distribution of resources between network applications. Also relevant is the modeling of a HetNet scenario with multiple traffic models (6 in total, as described in Section III: full buffer/IoT, HTTP, video, VoIP, gaming, and vehicular). Finally, the proposed algorithm is compared with traditional schedulers in the literature in scenarios with a varied number of users and different performance metrics. It highlights the throughput, the QoS latency requirements, the Block Error Rate (BLER), and the fairness index. For this purpose, the Vienna 5G System Level Simulator was used [10]. We emphasize our interest in applying Reinforcement Learning (RL) techniques in future work to reduce the total simulation time. This technique allows protocols to observe network conditions and uses previously acquired knowledge to respond efficiently to the complex and dynamic operation of resource allocation [11].

The rest of this article is organized as follows: Section II provides a summary of the main related works. Then it is discussed the requirements of 5G systems, application scenarios, and the implementation of traffic models in Section III. Section IV describes the scheduling algorithm techniques used in this research, while Section V shows the simulated scenario, its results, and analysis. Finally, Section VI provides the conclusions of this work.

II. Related Works

A survey about 5G usage scenarios and traffic models [12] associates existing traffic models with the most significant use cases, analyzing the performance of 5G systems. To do so, it considers attributes such as traffic volume, network deployments, and main performance targets. In addition, this survey brings together the main references from Standards Development Organizations (SDOs) and industry associations on the subject.

In [13], the authors propose to investigate the state of the art of 5G focused on mechanisms for coexistence between enhanced mobile broadband (eMBB) and ultra-reliable, low-latency communications traffic (URLLC) for resource scheduling. Thus, the paper presents a classification of works according to the following approaches: multiplexing, QoS provisioning, network slicing, machine learning, and Centralized/Cloud Radio Access Network (C-RAN). In the case of approaches based on QoS provisioning, the importance of the scheduling algorithm in considering both the QoS framework policy and user requirements is highlighted. It also emphasizes fairness as one of the important metrics to consider when designing a QoS-aware strategy.

The authors of [14] propose a new scheduling policy called Channel Aware Optimized Proportional Fair (CAOPF) aimed at optimizing the channel behavior based on CQI. The performance of the scheduler is analyzed and compared to schedulers such as RR and Proportional Fair (PF). The simulation results indicate that CAOPF presents better QoS performance considering that it provides no pending data of users in good channel conditions, higher average cellular throughput, higher user throughput in good and average channel conditions, and optimized fairness index. Despite considering various parameters and being a recent work, the algorithm was proposed based on LTE networks.

The authors of [15] consider channel conditions to improve the throughput performance of guaranteed bit rate and non-guaranteed bit rate traffics. To this end, they introduce a delay control mechanism that considers the QoS requirements of the varying traffic classes, improving the average throughput. However, the Prioritized QoS-Aware downlink scheduling algorithm is proposed for LTE networks and is not evaluated considering delay and packet drop ratio.

Another metrics-based scheduler was proposed in [16]. A 5G eMBB scheduling algorithm aware of throughput and CQI was developed, observing performance gains over the best CQI and PF algorithms in terms of both throughput and fairness.

A QoS-Aware joint component carrier selection and resource allocation scheme for carrier aggregation in 5G is proposed in [17]. The study addresses component carriers selection and resource allocation considering 5G QoS identifiers, maximizing the average throughput of users, and satisfying QoS users in terms of delay. It considers three index classes according to packet delay budget and packet loss ratio. Thus, the proposed scheme maximizes proportionally fairness average throughput of different service classes of users while meeting the delay and rate requirements. This study shows the significance of studying 5G schemes that consider delay constraints.

In [18], the authors developed two new policies for RT and NRT traffic, namely Adjusted Largest Weighted Delay First (ALWDF) and Fair Throughput Optimized Scheduler (FTOS), and then joined them to introduce the Advanced Fair Throughput Optimized Scheduler (AFTOS). It aims to maximize spectral efficiency and user throughput considering fairness, delay, and packet loss ratio. Although the study highlights a wide range of performance metrics and the results prove that AFTOS outperforms Maximum Throughput (MT), PF, and Modified Largest Weighted Delay First (MLWDF), the implementation relies on LTE systems with small cells.

Finally, in [9] the authors consider multiple traffic models and propose a QoS Aware Scheduler (QAS) to achieve the QoS requirements imposed on network performance. Round Robin (RR), best CQI, and QAS scheduling algorithms are compared in terms of average throughput, sum throughput, BLER, and latency per traffic model.

Based on the papers mentioned above, despite being a subject widely explored by academia and the telecommunications industry, the literature lacks studies that explore the combination of scheduling algorithms, mixed traffic models, and 5G HetNets. Therefore, this work intends to implement a new algorithm called Channel and Qualify of Service Aware Scheduler (CQAS) and compare its performance to the RR, best CQI, and QAS algorithms applied to HetNets in conjunction with full buffer, HTTP, video, VoIP, gaming, and vehicular traffic models. In addition, network stress is studied by varying the number of users and evaluating the metrics of average throughput, BLER, fairness index, and latency.

The novelty of this work to the related works is the analysis of a larger set of traffic models in the face of varying users in a heterogeneous scenario, both in terms of network structure and the plurality of RT and NRT applications. Furthermore, unlike other algorithms, CQAS considers both the channel conditions and QoS requirements of 5G systems, as well as being tested for a wider range of performance metrics.

III. 5G Usage Scenarios and Traffic Models

The International Telecommunication Union (ITU) describes the framework and overall objectives for the development of "IMT for 2020 and beyond", which includes 5G systems, defining the 5G usage scenarios [3].

Firstly, the eMBB scenario derives from the increased demand for mobile broadband services, which provide access to multimedia content, services, and user data. It therefore considers coverage in large areas and access points. In the case of large area coverage, there is a need for uninterrupted coverage, medium to high levels of mobility, and high data rates. In contrast, in the case of access points, there is a need for high user density, high traffic capacity, low levels of mobility, and higher data rates.

Concerning the URLLC, there are strict requirements regarding throughput, latency, and availability. Some examples include remote medical procedures and automation of smart grid systems. Finally, the massive Machine Type Communications (mMTC) use case is characterized by the large number of connected devices which, in general, transmit a low volume of data that is not sensitive to delay. The requirements for these devices are low cost and long battery life.

The heterogeneous structure of 5G networks encompasses various user types, making it necessary to study the traffic models for each usage scenario. According to [12], the main factors influencing the increase in 5G traffic are video use, as video-on-demand services will account for around two-thirds of all mobile traffic; the proliferation of devices, as it is esteemed that there will be an increase of around 1.4 billion smartphones and tablets from 2020 to 2030; application uptake, as it expects that will be more than 270 billion downloads of applications.

In this sense, applications related to VoIP and real-time gaming tend to take up even more space in network consumption. In addition, everyday access to web pages reinforces the importance of studying HTTP traffic. As for the proliferation of devices, the growing demand for internet access includes the growing application of IoT in industrial sectors and those related to people's convenience. Finally, vehicular traffic is associated with the use of cell phones in traffic or even applications of IoT devices in vehicles, such as monitoring and enabling intelligent functions. Then the traffic models studied and simulated in this work are described below according to [12] and [19].

Users have an infinite amount of data to transmit in full buffer traffic. The implementation is made so that each user transmits a single packet of infinite size in the model in Fig. 1. It is important to state that there is no record of transmission latency because all packets are not fully transmitted at the end of the simulation. IoT users also follow this model.

Fig. 1. Full buffer traffic model [19].

As for Hypertext Transfer Protocol (HTTP), the user's interactions with the World Wide Web (WWW) web page structure give this model a bursty profile. A web page consists of a main object and several embedded objects so the number of embedded objects, size of all objects, and reading and parsing time for the main object are the parameters that characterize web browsing. Fig. 2 illustrates the HTTP model.

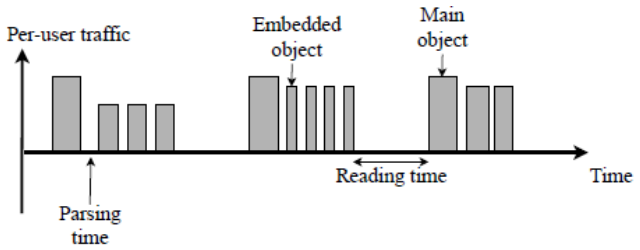

Fig. 2. Hypertext Transfer Protocol traffic model [19].

In relation to the video streaming traffic model, each video frame consists of several randomly sized packets arriving at regular time intervals T. Delay intervals called inter-arrival times are introduced between the packets of a frame by a video encoder. In this way, a video streaming session is characterized by the inter-arrival times between frame starts and between the packets of a frame, the number of packets per frame, and the size of each packet. A source video rate of 64 kbps is used. Fig. 3 illustrates the traffic model for video streaming.

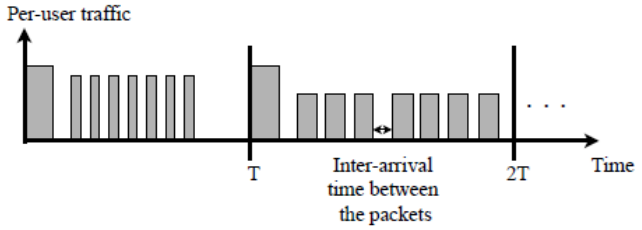

Fig. 3. Video streaming traffic model [19].

For VoIP traffic, it is used the Adaptive Multi-Rate (AMR) audio codec, a data compression scheme optimized for voice coding. A data rate of 12.2 kbps is used. There is only one VoIP packet generated every 20 ms during periods of activity, and the inter-arrival time between VoIP packets is named encoder frame length. In addition, every 160 ms during break times, a Silence Insertion Descriptor (SID) payload is generated. Fig. 4 illustrates the VoIP traffic model.

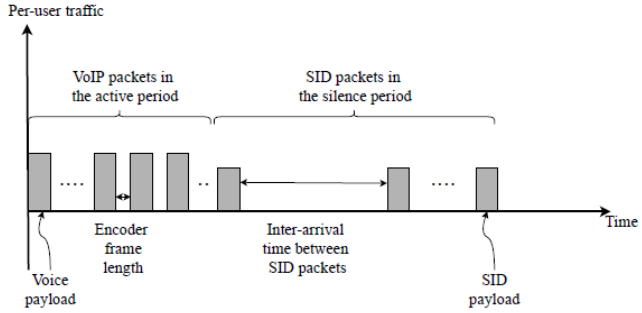

Fig. 4. VoIP traffic model [19].

Gaming traffic is generated with a uniformly distributed initial time to simulate the random timing relationship between client traffic packet arrival and uplink frame boundary. It is also characterized by parameters such as the inter-arrival time between packets, packet size, and the portion of the User Datagram Protocol (UDP) header to be added to the packet. It is worth noting that gaming packets are relatively small due to the interactive nature of games. Fig. 5 shows the gaming traffic model. Finally, the vehicular traffic model is based on [20]. Details and more information can be observed in [12].

IV. Scheduling Techniques

To achieve the benefits of a HetNet architecture, it is important to implement resource allocation or Radio Resource Management (RRM) techniques. Scheduling algorithms configure the efficient use of resources, which includes

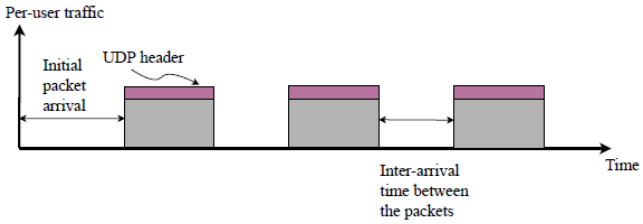

Fig. 5. Gaming traffic model [19].

bandwidth, power, and antennas, while mitigating interference between cells and users, and ensuring QoS levels for active users [2].

In general, the allocation of resources to a user follows this logic: the k-th Resource Block (RB) is allocated to the j-th user if its $m_{j,k}$ metric is the highest, i.e. if it meets (2) [8]:

$$m_{j,k} = max_{\cdot i} \{m_{i,k}\}. \quad (2)$$

This metric can be interpreted as the transmission priority of each user during a RB. The calculation of the metric varies according to some parameter related to the data flow, with a view to the desired performance of the system, such as:

- *Status of transmission queues*: the status of the UEs' transmission queues can be used to minimize delays in packet delivery. For example, the longer the queue, the higher the metric value.
- *Channel quality*: CQI can be used to allocate resources to users with better channel conditions. For example, the higher the flow rate, the higher the metric value.
- *Historical resource allocation*: information on previous performance can be used to increase equality in resource allocation. For example, the lower the last flow rate achieved, the higher the metric.
- *Buffer status*: conditions of the reception buffer can be used to avoid overloads. For example, the more space available in the receive buffer, the higher the metric.
- *QoS requirements*: the QoS Class Identifier (QCI) value of each data flow can be used to apply specific policies to meet QoS requirements.

The Round Robin scheduler performs fair sharing of time resources among all users according to an arbitrary list of users. This strategy guarantees equality in terms of the amount of time the channel is occupied by users, but it is not fair in terms of throughput, as this also depends on the experienced channel conditions. In this sense, the RR scheduling algorithm is a channel-unaware strategy [8]. Equation 3 shows the calculation of the RR metric, ${m_{i,k}}^{RR}$, where t represents the current time and T_i is the last time the user was served:

$${m_{i,k}}^{RR} = t - T_i. \quad (3)$$

In turn, best CQI scheduling provides the RB with the best channel quality to the users with the best link condition based

on (4). The term ς represents the CQI value, while i indicates the user for which the metric is calculated. Thus, the scheduler evaluates the CQI updates between the uplink and downlink of all users to give priority to active users that have a high CQI in the resource allocation process. As a result, this mechanism tends to provide the highest CQI value [21].

$$m_{i,k}{}^{best-CQI} = \varsigma_i(\tau). \tag{4}$$

The QoS Aware scheduling algorithm studied can be found in detail in [9]. The weighted sum throughput maximization problem can be written as (5).

$$\begin{gathered}\underbrace{\arg max}_{\{b_1,\ldots,b_{i,c}\}} c + \left(\sum_{i=1}^{I} \zeta_i t_i{}^T b_i\right) \\ \text{subject to:} \\ b(n) \in \{0,1\}, \forall n \\ b_j{}^T b_k = 0, \forall k \neq j \\ t_i{}^T b_i \geq c\gamma_i, \forall i \in \{non\ full\ buffer\ users\} \\ 0 \leq c \leq 1 \\ \sqrt{\mathcal{J}_o I}\|t_i{}^T b_i\|_2 \leq \sum_{i=1}^{I} t_i{}^T b_i\,, \forall i \\ \in \{full\ buffer\ users\}.\end{gathered} \tag{5}$$

The vector $t_i{}^T$ indicates the throughput of each user i, while $b_i = \left[b_{1,i}, \ldots, b_{n,i}\right]^T$ is the vector of RBs allocated to each user i. Note that $\zeta_i = \alpha^{-\beta i}\sigma^{-\max\{d_{c,i}-d_{i,0}\}}$, where: $\alpha^{-\beta i}$ is the reliability parameter, decreasing exponentially with the base $\alpha = 2$; β_i indicates the average BLER over user i codewords; $\sigma = 1.05$, indicating the latency priority factor; $d_{c,i} - d_i$ refers to the difference between the characteristic delay constraint (DC) of user i, $d_{c,i}$, and the current delay of that user d_i. The values of the characteristic delay constraints are predefined for each RT traffic model. In this sense, a user's priority is determined based on how close the current delay of a user's packet is to the delay constraint. In addition, users with highly reliable traffic, $\alpha^{-\beta i}$, have priority.

The first constraint indicates that the RBs are binary, while the second indicates that each RB is associated with one user at a time. The value γ_i represents the total amount of bits in user i's buffer, so the third constraint ensures that the amount of RBs allocated to a user is sufficient for their use. Thus, the variable c is included to achieve feasibility since it proportionally reduces the assigned RBs of all users. Finally, there is a constraint regarding fairness between full buffer users by implementing Jain's fairness index [22] so that $\mathcal{J}_o$ indicates the desired fairness index.

As mentioned in [9], this optimization problem is called mixed binary integer programming, and an open-source MATLAB tool for disciplined convex programming [23] is used alongside Gurobi Optimizer to solve it [24].

A Channel and Quality of Service Aware Scheduler (CQAS) is proposed, integrating the CQI with the tuning parameter used by the QAS. Algorithm 1 represents the summarized CQAS algorithm and Fig. 6 illustrates the summarized flowchart of this scheduler.

The algorithm starts by initializing the parameters of the RBs, redefining the resource grid for the slot, and defining the desired fairness. It then obtains the active users and measures the CQIs for each of them. Originally in the best CQI scheduler, the range of CQI values goes from 1 to 15. To match the reliability parameter, $\alpha = 2$, and the latency priority factor, $\sigma = 1.05$, the CQI values are set to the range $I = [1.1,1.15]$.

Next, the tuning parameter is configured based on the multiplication of the latency, reliability, and CQI parameters. It then creates an array of estimated throughput per RB for all users and, finally, the tuned throughput per RB is obtained as the multiplication of the tuning parameter by the estimated throughput. It also obtains the number of bits stored in each user's buffer and defines the integer binary optimization problem, as explained in the QAS algorithm. Finally, it schedules the users.

Algorithm 1: Summarized CQAS

Input: active users
Output: scheduled users
1. Define RB parameters: *currentTime*
2. Reset the resource grid for this slot
3. Define parameters: *desiredFairness, reliabilityParameter,* and *latencyParameter*
4. Get active users
  if there are no active users **then**
    return to Step 1
  else
    proceed to Step 5
  end
5. Creates an array of users and an array to associate each user with the corresponding CQI value
activeUsers = [user$_i$, user$_{i+1}$, ..., user$_n$]
CQIParameter = [user$_i$CQI, user$_{i+1}$CQI, ..., user$_n$CQI]
6. Measure CQIs and allocates CQI values to *CQIParameter*
7. Set the constraint that imposes that every RB is assigned to one user at a time
8. Set CQI values of *CQIParam.* to the range *I = [1.1,1.15]*
9. Set array of tuning parameter by multiplying latency, reliability, and CQI tuning parameters for every user:
*tuningParameter = reliabilityPar.*latencyPar.*CQIPar.*
10. Set array of estimated throughput for all users
11. Set array of tuned throughput per RB:
*tunedThroughput = tuningParameter*estimatedThroughput*
12. Set binary integer optimization problem
13. Schedule users

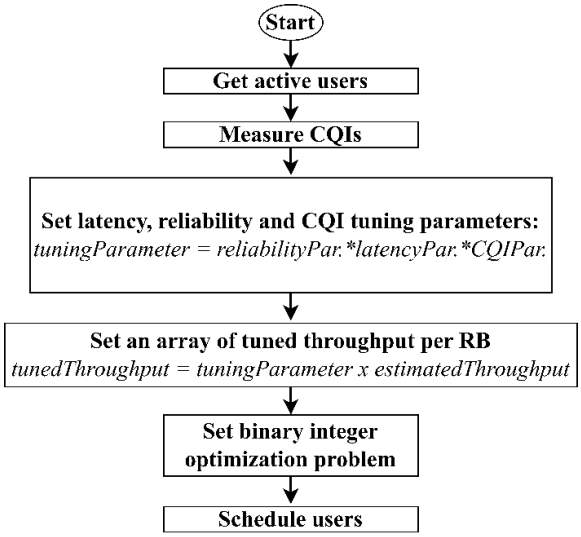

Fig. 6. CQAS scheduling summarized flowchart.

V. Simulation Results Discussion and Analysis

The Vienna 5G System Level Simulator MATLAB-based [10] is used to implement the CQAS scheduler and compare it to the RR, best CQI, and QAS schedulers. A HetNet scenario was implemented with mixed traffic models for RT and NRT applications and indoor and outdoor users: full buffer, which is also applied to IoT users, HTTP, video, VoIP, gaming, and vehicular. To do this, the number of users of each type was varied by 50, 100, 150, and 200, i.e. the total number of users was 350, 700, 1050 or 1400. Thus, the results presented were obtained through the average of ten simulation runs for each scheduler to consider the confidence of possible variations. However, this variation was very small, so it was decided not to represent the confidence interval in the figures. Fig. 7 illustrates the scenario that includes 350 total users and Table I shows the main simulation parameters.

The macro BSs are positioned in a hexagonal grid structure, while the pico BSs are positioned along the street where the vehicle users are, and the femto BSs are displayed in the center of the user clusters, mainly serving the IoT users. All other users are distributed according to a 2D Poisson distribution function [10].

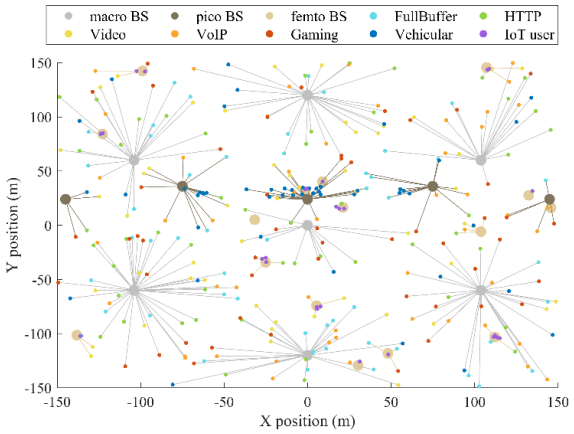

Fig. 7. Simulation scenario for the HetNet with multiple traffic models for 350 total users.

The 5G study on channel model for frequencies from 0.5 to 100 GHz [25] was the reference for the system-level simulations used to develop the path loss models in urban, indoor and street canyon macrocells, respectively, for macrocells and femtocells. In turn, the picocells follow the free-space path loss model presented in [26].

Furthermore, the Technical Specification Group Radio Access Network brings together the specifications for High Speed Downlink Packet Access: UE Radio Transmission and Reception (FDD) [27], defining the pedestrian and vehicular channel models chosen as the primary models for testing purposes. Finally, the channel model for IoT users is of the Rayleigh type.

Femtocells are favored for cell association and are responsible for allocating resources for IoT users. In turn, vehicular users are mostly allocated to picocells but are also served by macrocells. It should be noted that the movement model used is called random direction [10] since a random direction is assumed in the first slot so that the user moves in this direction at a constant speed. It is also important to point out that indoor/outdoor and Line-Of-Sight (LOS)/Non Line-Of-Sight (NLOS) decisions are defined by user type, Power Delay Profile (PDP) channel models are used for pedestrian and vehicular users, and an AWGN channel is assumed for users in clusters around femto base stations [10].

TABLE I
Main Simulation Parameters

Parameters	Values
Time slot duration	1 ms
Simulation duration	2000 time slots
Number of users	350, 700, 1050 or 1400
Traffic Models	Full buffer (IoT users are also configured as full buffer), HTTP, video, VoIP, gaming and vehicular
Number of BSs	7 macro BSs/46 dBm, 5 pico BSs/43 dBm,
Transmit Power	and 16 femto BSs/30 dBm
Path loss model	Macrocells – UrbanMacro5G [25] Picocells – Free space [26] Femtocells – Indoor or Street Canyon (outdoor) [25]
Channel model	Rayleigh (IoT users), vehicular and Pedestrian (remaining users) [27].

A. Throughput Gain

The average throughput per traffic model and scheduler for 350 total users is illustrated in Fig. 8, while the average throughput as a function of number of users for video, HTTP, and IoT is shown in Fig. 9, Fig. 10, and Fig. 11. Moreover, the overview of the average BLER for 350 total users can be seen in Table II.

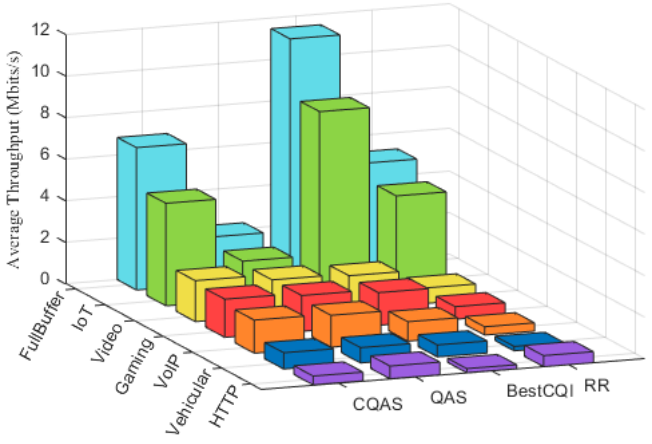

Fig. 8. Average Throughput per traffic model and scheduler for 350 total users.

Fig. 8 shows how CQAS outperforms QAS for all traffic models except for HTTP. It is noticeable how the implementation of the CQI tuning parameter in CQAS circumvents the limitation of QAS for full buffer and IoT users. Users with full buffer traffic, including IoT users, have higher average throughput values since their lowest average BLER. Also, Table II shows that video users have slightly better reliability than VoIP users since they produce larger amounts of larger packets, which results in higher average throughput. In turn, the gaming users have similar average BLER and average throughput to video users due to the higher data flow characteristic of this traffic model.

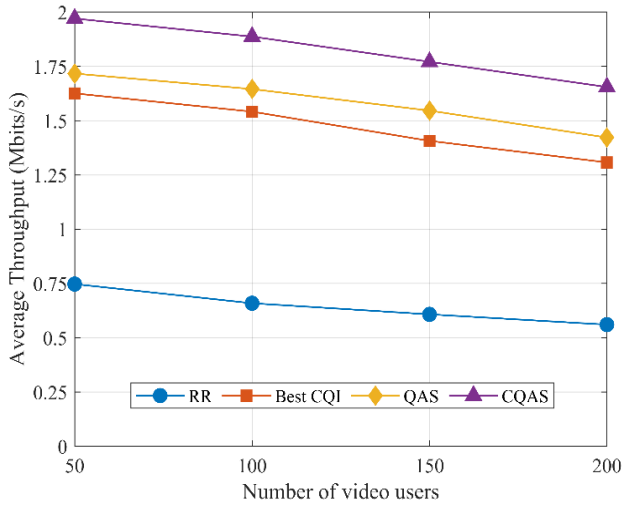

Fig. 9. Video users' Average Throughput as a function of number of users.

The RR scheduler allocates resources more fairly, leading to similar average throughput values that are lower than the other algorithms as well as lower average BLER values. Thus, the CQAS and QAS schedulers present a better combination of average throughput values while maintaining interesting BLER values. We can see the tendency for average throughput to decrease as the number of users increases observing Fig. 9,

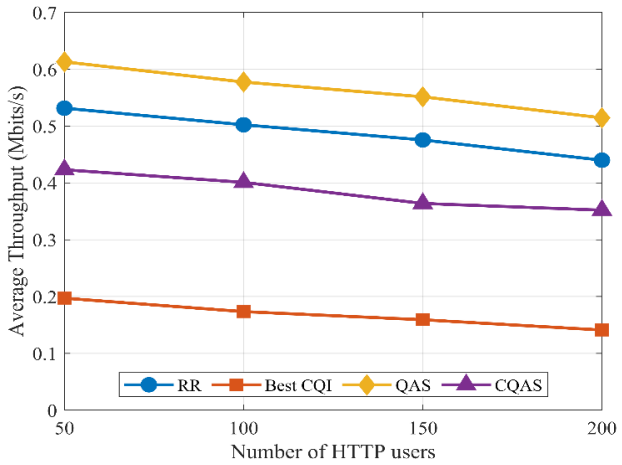

Fig. 10. HTTP users' Average Throughput as a function of number of users.

Fig. 10, and Fig. 11. Also notable is the gain in average throughput of CQAS for video users compared to the other schedulers, reaching values 14.8%, 14.7%, 14.5%, and 16.3% higher than QAS for 50, 100, 150, and 200 users, respectively. Moreover, there are gains of 163.6% and 21.2% for CQAS compared to RR and best CQI when analyzing 50 users. This shows the contribution of CQAS to improving the throughput performance of RT applications compared to other schedulers.

Fig. 10 shows the limitation of the CQAS algorithm concerning HTTP traffic, as it is worse than QAS so the throughput values are higher only when compared to the best CQI. This highlights one of the negative effects of implementing the CQI tuning parameter: CQAS is worse than QAS by 30.9%, 30.6%, 34%, and 31.6% for 50, 100, 150 and 200 users. On the other hand, Fig. 8 and Fig. 11 illustrate how CQAS is superior to QAS for the other NRT applications. Of particular note is Fig. 11, which shows the limitations of QAS for IoT traffic, so that even the RR algorithm obtained better results. The contribution of CQAS to the throughput of IoT users compared to QAS is seen in gains of 165.6%, 163.4%, 166.7%, and 171.2% for 50, 100, 150, and 200 users. Also according to Fig. 11, CQAS is superior to RR by 11.2% and inferior to best CQI by 46.6% for 200 users.

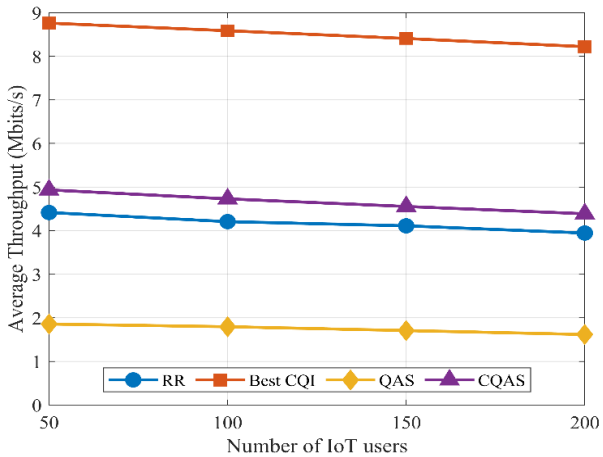

Fig. 11. IoT users' Average Throughput as a function of number of users.

B. Reliability Gain

As mentioned in the explanation of (5) in Section 4, lower average BLER values indicate high reliability and therefore a higher priority for user scheduling. In this sense, the resource allocation of the CQAS is effective, with BLER values intermediate between the RR and the best CQI scheduler, but slightly higher than the QAS values as shown in Table II. This is because the RR scheduler allocates resources to all users, while the best CQI scheduler evaluates the channel conditions so that some transmissions may fail and, consequently, reliability is reduced. In addition, CQAS shows a slight increase in BLER values compared to QAS due to the influence of the CQI parameter on its structure but shows an increase in overall throughput.

HTTP users have fewer active users per interval and, therefore, more users with zero average BLER, which lowers its average BLER value. Finally, vehicle users tend to have lower BLER values than other real-time traffic models due to the association of vehicles with picocells, which makes them less bottlenecked.

C. Fairness Index Gain

The fairness index is used to determine whether users are receiving resources fairly. This metric is calculated from the vector containing the data rate values according to [22]. Fig. 12 illustrates the variation in the fairness index as a function of the total number of users. The fairness index decreases as the total number of users increases. In addition, the highest values come from the RR scheduler, as it allocates resources to all users on a first-come, first-served basis. On the other hand, the best CQI is based on channel conditions, so users with poor channel conditions tend not to be benefited from scheduling. Finally, the QAS and CQAS algorithms obtained values close to the desired fairness parameter of 0.7 imposed in (5). However, the implementation of the CQI parameter in CQAS generated the cost of reducing the fairness index compared to QAS, while enabling greater average throughput.

TABLE II
AVERAGE BLER FOR 350 TOTAL USERS

	RR	best CQI	QAS	CQAS
Full Buffer	0.15	0.33	0.18	0.22
HTTP	0.10	0.25	0.12	0.18
Video	0.19	0.33	0.25	0.27
VoIP	0.18	0.35	0.24	0.29
Gaming	0.16	0.34	0.21	0.25
Vehicular	0.13	0.30	0.17	0.19
IoT	0.15	0.32	0.19	0.22

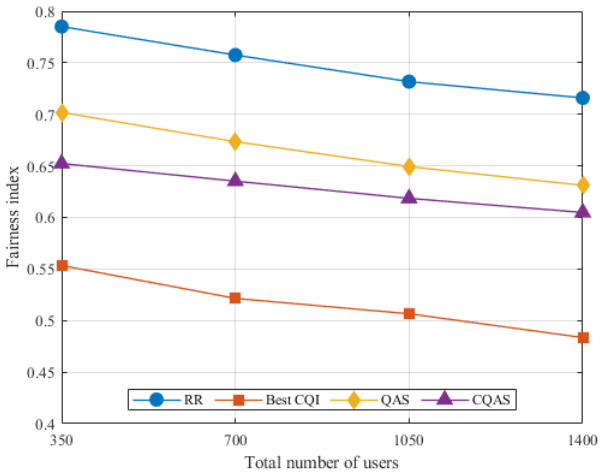

Fig. 12. Fairness Index as a function of total number of users.

D. Latency Gain

Regarding latency, the closer the latency value experienced by the user is to the delay constraint imposed on the traffic model, the higher the user's priority. In this sense, values of 20, 40, 60, and 100 ms are considered delay constraints, respectively, for the vehicular, VoIP, gaming, and video traffic models. Among the algorithms, only QAS and CQAS follow the expected performance following the latency sequence below the delay constraints (DCs): vehicular, VoIP, gaming, and video.

Fig. 13 illustrates the Latency Empirical Cumulative Distribution Function (ECDF) per real-time traffic model for CQAS and 1400 total users, which is the network's most stressed scenario. It is noticeable that 99% of users achieved latency values lower than the delay constraint for all cases, but the expected sequence was still followed as the delay constraints increased. CQAS and QAS reached 100% of real-time users with latency below their respective delay constraints for the scenarios of 350, 700, and 1050 total users. QAS also achieved 100% for all real-time users in the 1400 total users' scenario.

TABLE III
AVERAGE % OF USERS UNDER AVERAGE MAXIMUM LATENCY VALUES FOR 1400 TOTAL USERS

	RR		best CQI		QAS		CQAS	
	%	$\overline{max.\ lat.}$	%	$\overline{max.\ lat.}$	%	$\overline{max.\ lat.}$	%	$\overline{max.\ lat.}$
Vehicular	90	41	69	2080	100	20	99	22
VoIP	100	22	80	1909	100	37	99	41
Gaming	100	32	77	1936	100	53	99	61
Video	100	35	75	2039	100	95	99	101
HTTP	86	1996	50	1967	88	932	66	1536

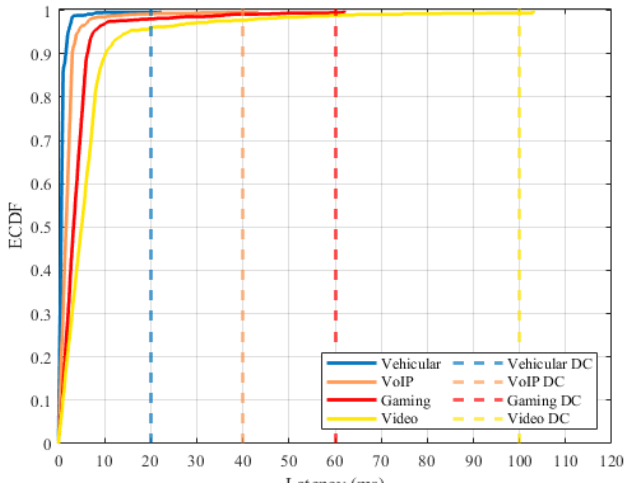

Fig. 13. Latency ECDF per real-time traffic model for CQAS and 1400 total users.

Table III shows the average % ($\overline{(\%)}$) of users under the average maximum latency values ($\overline{max.lat.}$) for 1400 total users. Concerning RR, all the DCs were exceeded, with, for example, 90% of the vehicular users under a maximum of 41 ms in the scenario of 1400 total users. As for the best CQI, it also didn't show good results for the RT applications. Considering the 1400 total users' scenario it is noticeable that 69% of vehicular users are under 2080 ms, 80% of video users are under 1909 ms, 77% of VoIP users are under 1936 ms, and 75% of gaming users are under 2039 ms. Hence, the results show that CQAS is more balanced between its commitment to serving applications with QoS requirements and providing more transmissions, and consequently, higher average throughput for RT and NRT applications. However, the QAS algorithm penalizes NRT traffic to the detriment of RT, while the best CQI scheduler is unable to handle RT traffic.

E. Simulation Time

Finally, as the QAS and CQAS optimization problem involves integer programming, they fall into the nonlinear programming (NLP) problems category, so they use computing resources intensively. Therefore, it is important to compare the simulation time information for each algorithm, given the variation in the total number of users. Table IV illustrates the average simulation time in minutes considering the 10 runs for each scenario.

TABLE IV
AVERAGE SIMULATION TIME IN MINUTES

Total users	RR	best CQI	QAS	CQAS
350	19	28	48	62
700	35	52	91	116
1050	66	98	175	221
1400	126	193	343	445

CQAS shows an increase in a total simulation time of 225%, 131%, and 30% when compared with the RR, best CQI, and QAS algorithms for the highest network-stressed scenario. Also noteworthy is the average value of 445 minutes to simulate a 1400-user scenario for CQAS, which makes it interesting to apply reinforcement learning methods to reduce the total simulation time for more stressful scenarios.

VI. CONCLUSION

Looking at recent publications on scheduling, there is a lack of metric-based schedulers in HetNet scenarios with multiple traffic models and varying numbers of users. Therefore, the novelty of this study is the proposal of the implementation of an algorithm based on the best CQI and QAS: CQAS, which considers both channel conditions and QoS. A comparative study was carried out between the RR, best CQI, QAS, and CQAS algorithms for full buffer, HTTP, video, VoIP, Gaming, and vehicular traffic in a HetNet made up of macrocells, picocells, and femtocells, as well as varying the number of users.

The results conclude that the CQAS scheduler implemented shows a significant improvement in the network's performance, highlighting the gains in overall throughput without greatly affecting the average reliability of users and the fairness index. In addition, it was possible to obtain latency values compatible with the delay constraints imposed on each traffic related to real-time applications, which is not achieved by the RR and best CQI schedulers.

The gain in the average throughput of CQAS for video users compared to the other schedulers stands out, reaching values up to 16.3% higher than QAS, and 21.2% and 163.6% more when compared to best CQI and the RR scheduler. Meanwhile, for NRT applications, the gains of CQAS over QAS for IoT traffic stand out: 165.6%, 163.4%, 166.7%, and 171.2% for 50, 100, 150, and 200 users.

However, for the same range of users, there is a limitation of CQAS: it is worse than QAS by 30.9%, 30.6%, 34%, and 31.6%. Note that there is a significant gain in throughput for NRT applications when comparing QAS and CQAS except for HTTP where it got a drop of between 30.6% and 34% depending on the number of users. It indicates that CQAS can maintain adequate performance for RT applications with QoS requirements and improve the throughput of NRT applications. Therefore, the network can serve more users and applications as required by 5G systems.

For future research, it is interesting to analyze heterogeneous networks under conditions of higher stress, given that for 1400 total users, CQAS achieved 99% of RT users with latency below the delay constraints. Furthermore, one of the points to be improved regarding CQAS is the simulation time, so the reinforcement learning methodology is a strong candidate to be studied and incorporated into solving the problem and improving the results of this work. Finally, the authors highlight the interest in applying techniques to improve a selected RT traffic or set of RT models to the detriment of other traffic models.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support provided by CNPq (process 131026/2022-4).

REFERENCES

[1] A. Mamane, M. Fattah, M. El Ghazi, M. El Bekkali, Y. Balboul, and S. Mazer, "Scheduling algorithms for 5G networks and beyond: Classification and survey," *IEEE Access,* vol. 10, pp. 51643-51661, 2022, doi: 10.1109/ACCESS.2022.3174579.

[2] E. Hossain and M. Hasan, "5G cellular: key enabling technologies and research challenges," *IEEE Instrumentation & Measurement Magazine,* vol. 18, no. 3, pp. 11-21, 2015, doi: 10.1109/MIM.2015.7108393.

[3] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The next generation wireless access technology*. Academic Press, 2020, doi: 10.1016/C2017-0-01347-2.

[4] A. Osseiran, J. F. Monserrat, and P. Marsch, *5G mobile and wireless communications technology*. Cambridge University Press, 2016, doi: 10.1017/CBO9781316417744.

[5] N. Bhushan *et al.*, "Network densification: the dominant theme for wireless evolution into 5G," *IEEE Communications Magazine,* vol. 52, no. 2, pp. 82-89, 2014, doi: 10.1109/MCOM.2014.6736747.

[6] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, "A survey on resource allocation for 5G heterogeneous networks: Current research, future trends, and challenges," *IEEE Communications Surveys & Tutorials,* vol. 23, no. 2, pp. 668-695, 2021, doi: 10.1109/COMST.2021.3059896.

[7] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE network,* vol. 34, no. 3, pp. 134-142, 2019, doi: 10.1109/MNET.001.1900287.

[8] 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 communications surveys & tutorials,* vol. 15, no. 2, pp. 678-700, 2012, doi: 10.1109/SURV.2012.060912.00100.

[9] A. Shiyahin, S. Schwarz, and M. Rupp, "Quality of Service Aware Scheduling in Mixed Traffic Wireless Networks," in *2022 IEEE 27th International Workshop on Computer Aided Modeling and Design* Networks," in *2022 IEEE 27th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2022: IEEE, pp. 159-165, doi: 10.1109/CAMAD55695.2022.9966904.

[10] M. K. Müller *et al.*, "Flexible multi-node simulation of cellular mobile communications: the Vienna 5G System Level Simulator," *EURASIP Journal on Wireless Communications and Networking,* vol. 2018, pp. 1-17, 2018, doi: 10.1186/s13638-018-1238-7.

[11] R. Buenrostro-Mariscal, P. C. Santana-Mancilla, O. A. Montesinos-López, J. I. Nieto Hipolito, and L. E. Anido-Rifon, "A review of deep learning applications for the next generation of cognitive networks," *Applied Sciences,* vol. 12, no. 12, p. 6262, 2022, doi: 10.3390/app12126262.

[12] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A survey on 5G usage scenarios and traffic models," *IEEE Communications Surveys & Tutorials,* vol. 22, no. 2, pp. 905-929, 2020, doi: 10.1109/COMST.2020.2971781.

[13] R. Kumar, D. Sinwar, and V. Singh, "QoS aware resource allocation for coexistence mechanisms between eMBB and URLLC: Issues, challenges, and future directions in 5G," Computer Communications, 2023, doi: 10.1016/j.comcom.2023.10.024.

[14] B. S. Monikandan, A. Sivasubramanian, S. Babu, G. Prasanna Venkatesan, and C. Arunachalaperumal, "Channel aware optimized proportional fair scheduler for LTE downlink," Peer-to-Peer Networking and Applications, vol. 13, pp. 2135-2144, 2020, doi: 10.1007/s12083-019-00826-z.

[15] A. Abdulazeez, M. M. Yahaya, I. B. Yabo, A. Bello, M. M. Umar, and A. Mohammed, "Prioritized quality of service-aware downlink scheduling algorithm for LTE network," in 2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON), 2022: IEEE, pp. 1-5, doi: 10.1109/NIGERCON54645.2022.9803123.

[16] M. I. Saglam and M. Kartal, "5G enhanced mobile broadband downlink scheduler," in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, 2019: IEEE, pp. 687-692, doi: 10.23919/ELECO47770.2019.8990378.

[17] R. Joda et al., "QoS-aware joint component carrier selection and resource allocation for carrier aggregation in 5G," in ICC 2021-IEEE International Conference on Communications, 2021: IEEE, pp. 1-6, doi: 10.1109/ICC42927.2021.9500923.

[18] D. H. Taha, H. Haci, and A. Serener, "Novel Channel/QoS Aware Downlink Scheduler for Next-Generation Cellular Networks," Electronics, vol. 11, no. 18, p. 2895, 2022, doi: 10.3390/electronics11182895.

[19] TU Wien Institute of Telecommunications, "The Vienna 5G System Level Simulator: User Manual" [Online]. Available: https://owncloud.tuwien.ac.at/index.php/s/izVNiuNNwnw1VS8. [Accessed: 4-June-2024].

[20] 3GPP, "Service requirements for V2X services," *3rd Generation Partnership Project (3GPP),* 2010.

[21] P. Thienthong, N. Teerasuttakorn, K. Nuanyai, and S. Chantaraskul, "Comparative study of scheduling algorithms in lte hetnets with almost blank subframe," *Engineering Journal,* vol. 25, no. 8, pp. 39-50, 2021, doi: 10.1109/iEECON45304.2019.8938835.

[22] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measurement of fairness and discrimination for resource allocation in shared computer system," *Eastern Research Laboratory, Digital Equipment Corporation: Hudson, MA, USA,* vol. 2, 1984, doi: 10.48550/arXiv.cs/9809099.

[23] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," ed, 2014.

[24] *Gurobi Optimizer*. (April 2010). Gurobi Optimization, Inc.

[25] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," *3rd Generation Partnership Project (3GPP),* 2017.

[26] A. F. Molisch, *Wireless communications*. John Wiley & Sons, 2012.

[27] 3GPP, "High Speed Downlink Packet Access (HSDPA); User Equipment (UE) radio transmission and reception (FDD)," *3rd-Generation Partnership Project (3GPP), 2002.*

Gabriel A. Queiroz obtained his Bachelor in Electronic and Telecommunications Engineering from the Federal University of Uberlândia – UFU, Minas Gerais, Brazil (2022). He is currently a MSc student in the Department of Electrical Engineering with an emphasis on computer networks at the Federal University of Uberlândia – UFU. His areas of interest include mobile communications, scheduling algorithms, and heterogeneous networks.

Éderson R. da Silva is an Associate Professor (Ph.D) in the Department of Electrical Engineering at the Federal University of Uberlândia – UFU. He holds a degree (2007) and a Ph.D (2010) in Electrical Engineering from the Federal University of Uberlândia – UFU. He has experience in Electrical Engineering and Telecomunnications, with an emphasis on computer networks, working mainly on the following topics: mobile wireless networks, simulation and quality of service.