A Weighted QoS Aware Scheduler algorithm for multiple traffic models in 5G heterogeneous networks

Gabriel A. Queiroz, and Éderson R. da Silva

***Abstract*— Fifth Generation (5G) networks are associated with Quality of Service (QoS) requirements in conjunction with the greater densification of heterogeneous network scenarios. Thus, the plurality of applications and services stands out, which involves real-time (RT) and non-real-time (NRT) traffic models for users. Thus, system-level simulations are performed to implement a Weighted QoS Aware Scheduler (WQAS) and compare it to the QoS Aware Scheduler (QAS), Round Robin (RR), and Best Channel Quality Indicator (best CQI) scheduling algorithms in a heterogeneous network with varying numbers of users and multiple traffic flows. In this sense, the main contribution of this work is the prioritization of RT applications, and the results show gains in throughput and high performance for reliability and latency.**

Keywords*— *5G heterogeneous networks, non-real-time traffic, real-time traffic, scheduling algorithm, Weighted QoS Aware Scheduler.

I. INTRODUCTION

Fifth Generation (5G) and beyond-5G mobile communication systems must deal with the implementation of a wide range of new technologies and applications, considering the 5G usage scenarios of enhanced Mobile Broadband (eMBB), Ultra Reliable and Low Latency Communications (URLLC), and massive Machine Type Communications (mMTC) [1].

This way, there is a significant increase in users, services, and applications and 5G networks must implement more efficient networks considering higher data rates, greater spectral and energy efficiency, reduced latency, and increased network capacity. This capacity increase can be attributed to three main factors: a higher number of mobile nodes, the growth of spectrum use, and greater channel efficiency [2].

It is worth highlighting the increase in network capacity through network densification, which consists of a heterogeneous network (HetNet) of multiple base stations (BSs) that operate with different output powers and diverse cell sizes. In addition, the complex structure of HetNets involves the diversification of users through multiple traffic models, which must meet throughput, latency, and data loss criteria following Quality of Service (QoS) requirements [3].

Considering the requirements of latency, user experienced data rate and reliability, we highlight the use cases related to video, work and gaming on the cloud, augmented reality, and industry automation, which are found among the usage scenarios of eMBB and URLLC [2], as they encompass access to multi-media content and time-sensitive applications. Thus, in a heterogeneous network where traffic is close to eMBB and URLLC services, there are requests for large amounts of data linked to latency and reliability, so it is crucial to make efficient use of network resources through scheduling algorithms [4].

The objective of this paper is to propose a Weighted QoS Aware Scheduler (WQAS) based on the QoS Aware Scheduler (QAS) algorithm proposed in the study by A. Shiyahin, et al. [5], highlighting real-time (RT) applications. In addition, the simulated scenario is a HetNet with multiple traffic models. For RT services, we have vehicular, Voice over IP (VoIP), gaming, and video streaming, while for non-real-time traffic models, we consider Hypertext Transfer Protocol (HTTP) and full buffer, which also serves as the basis for Internet of Things (IoT) users.

The novelty of this paper is the analysis of a greater variety of traffic models, as well as the variation in the number of users in a heterogeneous scenario, both in terms of network structure and the diversity of services provided. Hence, the proposed scheme enables RT applications to be prioritized through the allocation of Resource Blocks (RBs) and this provides flexibility and better predictability of the expected behavior in multi-traffic scenarios.

The remainder of this article is laid out as follows: Section II presents a summary of the main related works. The scheduling algorithm techniques are described in Section III, while Section IV presents the simulation scenario, its results, and analysis. Finally, the research conclusions are presented in Section V.

II. RELATED WORKS

Considering traffic models and 5G usage scenarios, J. Navarro-Ortiz, et al. [6] proposed a survey about 5G systems and the association of traffic models with the most important 5G use cases. It also brings together a vast amount of information on the subject, taking into account references from various Standards Development Organizations (SDOs) and industry associations. Thus, it analyzes performance targets, network deployments, and traffic volume. As for scheduling algorithm techniques, the survey proposed in [1] stands out. This literature review compares different scheduler techniques, considering metrics, performance and highlighting the considerations of each proposed algorithm in relation to 5G networks.

Regarding algorithms based on network metrics, we highlight the study [7], which highlights the heterogeneous traffic offered to the 5G network and the need for scheduling algorithms to consider QoS requirements. It evaluates the Maximum Rate (MR), Round Robin (RR), Proportional Fair (PF), and a proposed UE-based Maximum Rate (UEMR) scheduler in terms of throughput and fairness.

Also noteworthy is the work [8], which studies the densification of 5G networks and the resource allocation considering QoS in terms of fairness and perceived throughput. It compares the PF, Exponential PF (EXP-PF), and Maximum Largest Weighted Delay First (MLWDF) algorithms,

Gabriel A. Queiroz, and Éderson R. da Silva, Department of Electrical Engineering, Federal University of Uberlândia – UFU, Uberlândia, MG, Brazil, e-mails: gabriel.andrade@ufu.br, and ersilva@ufu.br. This work was supported by CNPq (process 131026/2022-4).

considering the fairness index, useful rate, and spectral efficiency.

The authors of [4] propose a survey that addresses the packet scheduling algorithms in URLLC for 5G and beyond-5G systems, highlighting the perspectives of decentralized, centralized and joint scheduling techniques. Then, it analyzes the performance of some algorithms and points out the main challenges in the area. Finally, the authors of [5] study multiple traffic models and propose a QoS Aware Scheduler (QAS) to meet the QoS requirements imposed on 5G systems. To this end, they compare the QAS to the RR and best CQI algorithms, considering average throughput, sum throughput, Block Error Rate (BLER), and latency.

Considering the works mentioned above, it can be seen that the literature lacks studies on the combination of scheduling algorithms, multiple traffic models, and heterogeneous networks, even though these fields are highly explored by academia and the telecommunications industry. According to [6], one of the main influences in the increase in 5G traffic is video use, since video-on-demand services will represent around two-thirds of overall mobile traffic.

Therefore, this work implements a modification of the QAS algorithm, called Weighted QoS Aware Scheduler (WQAS) and compares its performance to the RR, best CQI and QAS algorithms in a heterogeneous network scenario with mixed traffic models, highlighting RT applications, mainly video streaming. So, a resource allocation scheme is proposed which, by assigning RBs based on scheduling weight, can prioritize these traffic classes.

III. SCHEDULING ALGORITHMS

Scheduling algorithms are Radio Resource Management (RRM) techniques responsible for the efficient resource allocation. The operation of a scheduling algorithm can be described according to Eq. (1) [9]:

$$m_{j,k} = max_{\cdot i} \{m_{i,k}\}. \quad (1)$$

The k-th Resource Block (RB) is allocated to the j-th user if its $m_{j,k}$ metric is the highest. In this way, this metric indicates the transmission priority of each user given a particular RB.

A. Round Robin (RR)

The Round Robin (RR) is a channel-unaware scheduler, as it performs a fair division of time resources between all users following a random list of users, guaranteeing equality in terms of the time each user occupies the channel, but it is not fair in terms of throughput, which depends on channel conditions [9]. The RR metric is calculated according to Eq. (2), where t indicates the current time interval and T_i represents the last time interval in which the user was scheduled.

$${m_{i,k}}^{RR} = t - T_i. \quad (2)$$

B. Best Channel Quality Indicator (best CQI)

In turn, the best CQI scheduler considers the quality of the channel to allocate resources to users. In Eq. (3), the CQI value is indicated by ς and calculated for each user i. Thus, users with the highest CQI value are prioritized in the resource allocation mechanism [10].

$${m_{i,k}}^{best-CQI} = \varsigma_i(\tau). \quad (3)$$

C. Quality of Service Aware Scheduler (QAS)

As for the QoS Aware scheduler, it was proposed by A. Shiyahin, et al. in [5] and performs scheduling based on a weighted sum throughput maximization problem following Eq. (4). The solution to this optimization problem depends on discipline convex programming [11] and, to this end, the MATLAB-based Gurobi Optimizer [12] is chosen.

The RBs allocated to each user i are indicated by the vector $b_i = [b_{1,i}, \dots, b_{n,i}]^T$, while ${t_i}^T$ indicates the throughput vector. $\alpha^{-\beta i}$ is the reliability parameter, which decreases exponentially with the base $\alpha = 2$, has β_i as an indicator of the average BLER on user i codewords, and the latency priority factor $\sigma = 1.05$. Thus, users with highly reliable traffic, indicated by $\alpha^{-\beta i}$, have priority in scheduling.

Another factor that impacts user priority is the proximity of the current delay of the user's packet to the delay constraint (DC) of the configured real-time (RT) traffic model. Thus, $d_{c,i} - d_i$ represents the difference between the characteristic delay constraint of user i, $d_{c,i}$, and the current delay of the user d_i.

Also, in Eq. (4), the first two constraints indicate that the RBs are binary, and that each RB is associated with one user at a time. The third constraint ensures that the amount of RBs allocated to a user is sufficient based on the total amount of bits in the buffer of user i indicated by γ_i. Thus, c guarantees viability by proportionally reducing the allocated RBs of all users. Finally, $\mathcal{J}_o$ represents the desired fairness index, which is a constraint on fairness implemented using Jain's fairness index [13].

$$\underbrace{\arg max}_{\{b_1,\dots,b_{i,c}\}} c + \left(\sum_{i=1}^{I} \zeta_i {t_i}^T b_i\right)$$

subject to:

$$b(n) \in \{0,1\}, \forall n$$
$$b_j{}^T b_k = 0, \forall k \neq j$$
$${t_i}^T b_i \geq c\gamma_i, \forall i \in \{non\ full\ buffer\ users\}$$
$$0 \leq c \leq 1$$
$$\sqrt{\mathcal{J}_o I}\|{t_i}^T b_i\|_2 \leq \sum_{i=1}^{I} {t_i}^T b_i, \forall i \in \{full\ buffer\ users\}. \quad (4)$$

D. Weighted Quality of Service Aware Scheduler (WQAS)

This study proposes a modification of QAS in the mixed traffic models scenario by implementing weights for RT traffic users. Thus, vehicular, VoIP, gaming, and video streaming users have a scheduling weight of 10, resulting in 10 consecutive RBs assigned to the user if they are scheduled. On the other hand, NRT users are assigned only one RB when scheduled. Algorithm 1 shows the summarized Weighted QoS Aware Scheduler.

To implement this scheduling weight parameter, slots are initially distributed following a Weighted Round Robin (WRR) queuing process. Thus, a vector of consecutive RBs assigned is created to determine different quantities of RBs for the RT and NRT services. Next, a second level of scheduling is carried out from the QoS Aware scheduler, in which the parameters relating to desired fairness, reliability, and latency are defined and the tuning parameter for calculating the tuning throughput per RB is configured. Finally, the integer binary optimization problem is

configured, and users are scheduled considering the number of RBs previously assigned by the WRR level.

Algorithm 1: Summarized Weighted QoS Aware Scheduler

Input: active users
Output: scheduled users
1. Define RB parameters: $currentTime$
2. Reset the resource grid for this slot
3. Define fairness parameter: $desiredFairness$
4. Get active users
5. **WRR level**
   if $user == RT\ user$ **then**
   $weightWRR = 10$
   else
   $weightWRR = 1$
   end if
6. Set the constraint that imposes that every RB is assigned to one user at a time
7. Set tuning parameter by multiplying latency and reliability tuning parameters:
   $tuningP = reliabilityParam.x\ latencyParam.$
8.Set array of tuned throughput per RB:
$tunedThroughput = tuningP\ x\ estimatedThroughput$
9. Set binary integer optimization problem
10. Schedule users

IV. Simulation Results Discussion And Analysis

The Vienna 5G System Level Simulator [14] was chosen to compare the RR, best CQI, QAS, and WQAS algorithms. Ten simulations were carried out for each scenario, varying the scheduler and the number of users, so that the results analyzed were obtained as an average. A HetNet was implemented by deploying diverse base stations and users with multiple traffic models. For RT traffic models, video, VoIP, gaming, and vehicular were implemented, while for NRT models, HTTP and full buffer (which is also applied to IoT users) were used. In addition, the total number of users was 350, 700, 1050, or 1400, i.e. each type of user varied by 50, 100, 150, and 200.

Table I shows the main simulation parameters. Of relevance is the implementation of 7 macro BSs, 5 pico BSs, and 16 femto BSs. The macro BSs are arranged in a hexagonal grid structure, following the urban path loss model found in the study [15]. In turn, the pico BSs are located along the streets, serving vehicular users and presenting the free-space path loss model [16]. Finally, femto BSs are located in the center of user clusters and allocate resources mainly to IoT users, following the indoor or Street Canyon (outdoor) path loss model used in the study [15]. The metrics analyzed were throughput, BLER (reliability), fairness index and latency.

A. Throughput

Fig. 1 shows the average throughput per traffic model for 1400 total users, while Fig. 2 illustrates the variation in the average throughput of video users as a function of the number of users. The average BLER for 1400 total users is shown in Fig. 3.

Moreover, Fig. 1 shows that WQAS performs best for RT traffic models, while NRT users are neglected and have lower average throughput when compared to QAS. Users with lower average BLER are expected to have higher average throughput due to greater reliability. However, despite having the highest overall values, full buffer, and IoT users had the highest negative variations when comparing QAS and WQAS. This shows that there is a trade-off when reserving RBs for RT traffic models, to the detriment of NRT traffic models. In general, RT users have similar average BLER, but video users stand out with higher average throughput due to the greater number of larger packets transmitted.

TABELA I. *MAIN SIMULATION PARAMETERS.*

Parameters	Values/Meaning
Simulation duration	2000 time slots
Time slot duration	1 ms
Schedulers	RR, best CQI, QAS, and WQAS
Traffic models	RT – vehicular, VoIP, gaming, and video NRT – HTTP, and full buffer (also IoT)
Delay Constraints (DCs)	Vehicular – 20 ms VoIP – 40 ms Gaming – 60 ms Video – 100 ms
Number of users	350, 700, 1050 or 1400
Number of BSs/ Transmit power	7 macro BSs/46 dBm 5 pico BSs/43 dBm 16 femto BSs/30 dBm
Path loss model	Macro BSs – UrbanMacro5G [15] Pico BSs – free space [16] Femto BSs – indoor or Street Canyon (outdoor) [15]
Channel model	IoT users – Rayleigh Other users – vehicular or pedestrian as in [17]

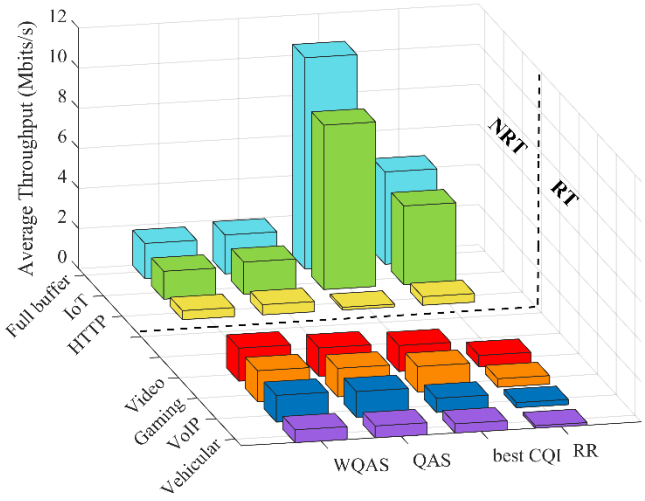

Fig. 1. Average Throughput per scheduler and traffic model for 1400 total users.

Considering the performance of RT traffic, the proposed WQAS has a better overall performance, as there is a higher average throughput while maintaining average BLER values practically the same as those observed in the QAS implementation. Thus, Fig. 2 highlights the gains of WQAS concerning the other schedulers for video users, which have the highest average throughput among the RT models. Comparing WQAS to QAS, there are gains of 12.3%, 11.3%, 14.7% and 16.2% for 50, 100, 150 and 200 users. Also, for the 200-user scenario, in which there is greater network stress, WQAS outperforms the RR and best CQI by 192.6% and 26.9%. In addition, there is a tendency for the average throughput to decrease as the number of users increases.

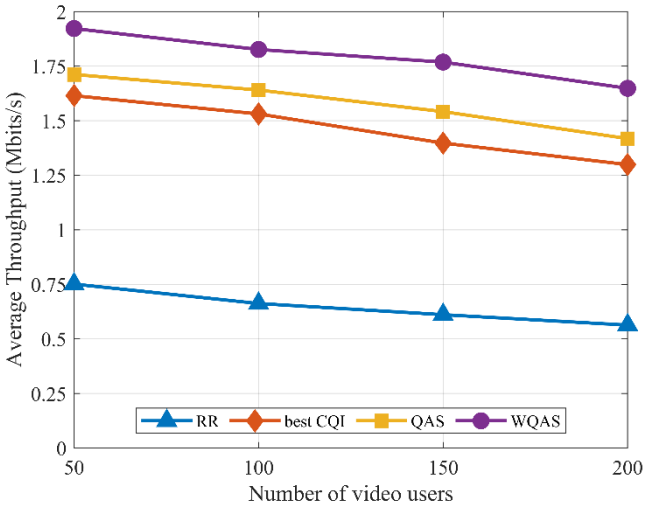

Fig. 2. Video users' average throughput as a function of number of users.

B. Reliability

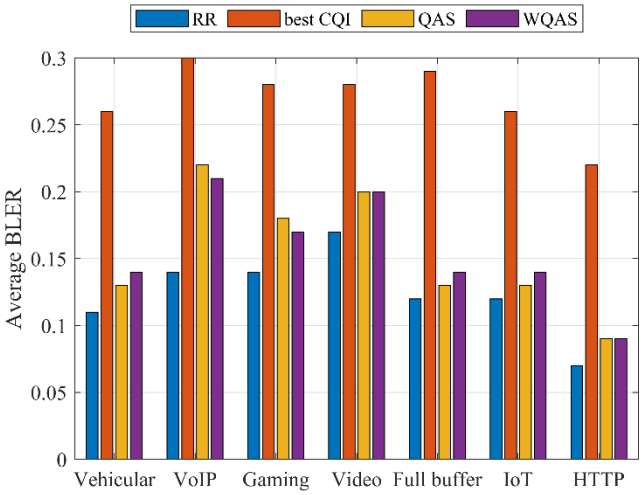

Fig. 3. Average BLER for 1400 total users.

As explained in Eq. (4), the lower the average BLER value, the higher the reliability, which means a higher priority for user scheduling. In this way, Fig. 3 shows that both QAS and WQAS have intermediate values between RR and best CQI, but the proposed WQAS stands out because it performs better in the average throughput metric while maintaining similar average BLER values to QAS. It should be noted that the lowest values are attributed to RR due to its scheme for allocating resources to all users, while best CQI has the highest number of transmission failures due to the evaluation of channel conditions and therefore has the lowest reliability.

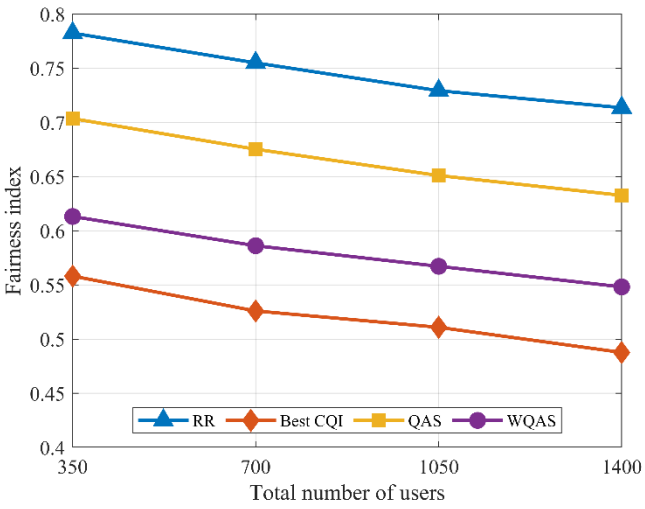

Fig. 4. Fairness Index as a function of total number of users.

C. Fairness Index

The fairness index is calculated from the Jain's Fairness Index metric [13] and determines the fairness of resource distribution. Fig. 4 shows the fairness index as a function of the total number of users, indicating a reduction in the fairness index as the total number of users increases.

The worst performance is linked to the best CQI scheduler, as it is an algorithm that takes channel conditions into account and tends not to serve users with poor channel conditions, for example, users located on the edges of cells. On the other hand, RR has the highest fairness index, as it serves all users on a first-come, first-served basis. For its part, QAS maintains values close to the desired fairness index parameter of 0.7 imposed in (4) in accordance with [5]. Finally, WQAS shows a drop in fairness performance compared to QAS, the price is due to the reservation of RBs for RT-type users, reducing the chance of allocating resources to NRT users and lowering the overall fairness index.

D. Latency

Concerning latency, the proximity of each user's latency value to the delay constraints (DCs) defined for each traffic model is evaluated. Thus, the closer the DC imposed, the higher the user's priority. The DC values adopted were 20, 40, 60, and 100 ms for the vehicular, VoIP, gaming, and video traffic models, according to Navarro-Ortiz, Jorge, et al [6].

Both QAS and WQAS achieved the desired performance for latency following the order of the DCs: vehicular, VoIP, gaming, and video. On the other hand, the RR and best CQI algorithms extrapolate all the DCs. For RR, for example, only 88% of vehicular users fall below a maximum of 42 ms for 1400 total users. As for the best CQI, this same scenario shows 72% of vehicular users below 2035 ms. Also noteworthy for the best CQI: 81% of video users are under 1875 ms, 75% of VoIP users are under 1954 ms, and 78% of gaming users are under 1992 ms.

Fig. 5 illustrates the Latency Empirical Cumulative Distribution Function (ECDF) per RT traffic model for WQAS under 1400 total users, indicating the highest network stress. Note that 100% of users achieved values lower than the DCs for

all traffic models, following the sequence of imposed DCs. The following values stand out: 100% of vehicular users are under 17 ms, 100% of VoIP users are under 38 ms, 100% of gaming users are under 56 ms, and 100% of video users are under 96 ms. Therefore, the results show that WQAS achieves the QoS requirements demanded by 5G applications, as well as having high average throughput and reliability. In addition, WQAS performs better in terms of latency, since in the most stressed network scenario, it has a lower average latency than QAS by 2 ms for vehicular users, 1.5 ms for VoIP users, 3.2 ms for gaming users, and 2.8 ms for video users. Hence, the results indicate that WQAS tends to maintain the DCs of RT applications in scenarios of high network stress for more situations than QAS.

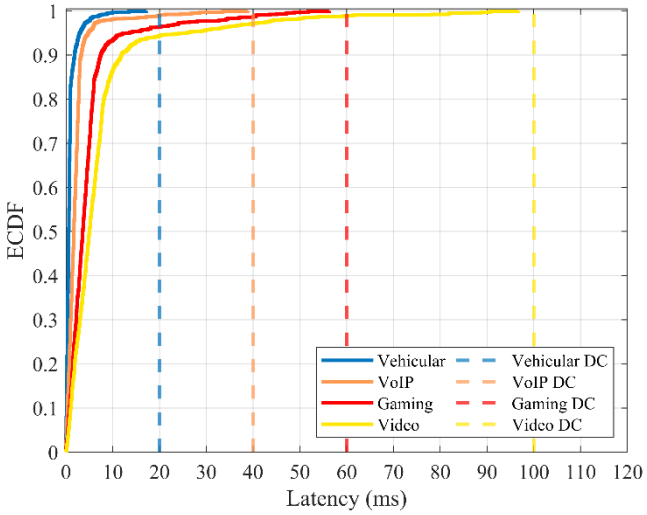

Fig. 5. Latency ECDF per RT traffic model for WQAS under 1400 total users.

V. CONCLUSION

Considering recent publications on scheduling algorithms, there is a lack of studies that encompass metric-based schedulers, HetNet scenarios, and multiple traffic models together. Therefore, this study proposes the implementation of a Weighted QoS Aware scheduler in a HetNet given the variation in the number of users and considering multiple traffic models: vehicular, VoIP, gaming, video, HTTP, and full buffer.

The results show that the implemented WQAS presents considerable improvements in network performance when considering the average throughput of real-time users, as well as the average BLER and latency requirements imposed through delay constraints on traffic models. However, there was a significant reduction in the fairness index, because of the distribution of RBs being based on weights, which is detrimental to NRT applications. Regarding average throughput, WQAS obtained gains of 192.6%, 26.9%, and 16.2% when compared to RR, best CQI, and QAS considering video users in the most stressed network scenario with 1400 total users. In addition, the latency values for WQAS and QAS were compatible with the sequence of delay constraints implemented.

Regarding the variation in users, it was observed that both the average throughput and the fairness index decrease as the number of users increases. Reliability decreases as the number of users increases because BLER also increases. Finally, for all user variations, the latency results were achieved for 100% of RT users. Therefore, WQAS is a potential algorithm for RT traffic models and shows moderate results for NRT services. As for future work, it is worth analyzing other proportions in the distribution of RBs, in order to achieve better performance for real-time services that include eMBB and URLLC scenarios.

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.

[2] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The next generation wireless access technology*. Academic Press, 2020.

[3] 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.

[4] M. E. Haque, F. Tariq, M. R. Khandaker, K.-K. Wong, and Y. Zhang, "A survey of scheduling in 5g urllc and outlook for emerging 6g systems," *IEEE access*, 2023.

[5] 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 of Communication Links and Networks (CAMAD)*, 2022: IEEE, pp. 159-165.

[6] 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.

[7] C. F. Müller, G. Galaviz, Á. G. Andrade, I. Kaiser, and W. Fengler, "Evaluation of scheduling algorithms for 5g mobile systems," *Computer Science and Engineering—Theory and Applications*, pp. 213-233, 2018.

[8] A. Mamane, M. El Ghazi, S. Mazer, M. Bekkali, M. Fattah, and M. Mahfoudi, "The impact of scheduling algorithms for real-time traffic in the 5G femto-cells network," in *2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC)*, 2018: IEEE, pp. 147-151.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), 2017.

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

[17] 3GPP, "High Speed Downlink Packet Access (HSDPA); User Equipment (UE) radio transmission and reception (FDD)," 3rd-Generation Partnership Project (3GPP), 2002.