Quality of Service Aware Scheduling in Mixed Traffic Wireless Networks

Master's Thesis

for obtaining the academic degree

Diplom-Ingenieurin

as part of the study

Telecommunications

carried out by

Areen Shiyahin

matriculation number: 11930497

Institute of Telecommunications

at TU Wien

Supervision:

Associate Prof. Dipl.-Ing. Dr.techn. Stefan Schwarz

Univ. Prof. Dipl.-Ing. Dr.techn. Markus Rupp

Statement on Academic Integrity

Hiermit erkläre ich, dass die vorliegende Arbeit gemäß dem Code of Conduct – Regeln zur Sicherung guter wissenschaftlicher Praxis (in der aktuellen Fassung des jeweiligen Mitteilungsblattes der TU Wien), insbesondere ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel, angefertigt wurde. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In– noch im Ausland in gleicher oder in ähnlicher Form in anderen Prüfungsverfahren vorgelegt.

Vienna, September 2022

Areen Shiyahin

Abstract

Considering a traffic mix in a wireless network, the scheduler at the base station needs to be aware of the type of user packets and the state of each user buffer to satisfy the requirements of users. In such networks, the requirements vary in terms of latency, throughput, and reliability. Thus, a trade-off is required between the prior performance metrics to enhance the overall network performance. In this thesis, a Quality of Service Aware Scheduler (QAS) is proposed with tuning parameters to achieve balanced Quality of Service (QoS) delivery. Moreover, a moderate fairness is imposed among full buffer users that are assumed to have infinite amount of data in packet buffer. An open-source modeling tool is used to solve the Resource Blocks (RBs) optimization problem of QAS. System Level (SL) simulations are performed to investigate the performance of the proposed scheduler. In addition, QAS performance is compared to the standard scheduling strategies, namely Round Robin (RR) and Best Channel Quality Indicator (CQI) under different network loads and users positioning methods.

Contents

1 Introduction

1.1 Motivation

Nowadays, wireless networks are extremely complex due to the high number of users and the massive heterogeneity in terms of applications and their corresponding traffic models [34]. The number of connected devices and users is expected to reach 29.3 billion by 2023 [33]. Each application has some Quality of Service (QoS) requirements in terms of latency, throughput, and data loss [6]. In such large-scale and versatile wireless networks, the existence of different services poses a considerable challenge in achieving all QoS requirements simultaneously. It is essential to provide adequate QoS for all applications without comprising either of them. Therefore, Quality of Service Aware Scheduler (QAS) is definitely needed in network design to balance the requirements of Real Time (RT) and Non Real Time (NRT) applications and maximize user satisfaction. Standard scheduling strategies such as Round Robin (RR) and Best Channel Quality Indicator (CQI) do not consider buffer state information, or type of user data. Furthermore, the RR scheduler do not consider the experienced quality of each radio link in the cell. It distributes resources among users equally, allowing for a high level of fairness [13]. The Best CQI scheduler is opportunistic in nature. It allocates users with good channel conditions to resources allowing high system throughput to be achieved [13]. Thus, a trade-off must be achieved in resource allocation to guarantee enhanced system throughput and reliability, and reduction in latency.

Generally speaking, formulating the Resource Blocks (RBs) allocation problem with varying requirements leads to a non-convex optimization problem when the objective function of the problem is, e.g., maximizing the sum rate over all users satisfying some constraints [12]. The solution of a mathematical optimization problem depends on the type of objective and constraint functions and the number of variables and constraints involved in the underlying problem [5]. The general optimization problem is usually difficult to solve and may be intractable. However, there are some problem classes that can be solved reliably. A well-known example is the convex optimization problem. In addition, there are some transformations that can be used to translate some general optimization problems into convex ones.

As realistic measurements in wireless communications are costly, onerous and including significant overhead, the need for simulations arises [15]. In the context of cellular networks, the term System Level (SL) means an accurate abstract representation of the physical layer of a wireless system. This representation can be evaluated with a lower computational complexity than computing all of the algorithms involved in the processing of the physical layer. Generally, SL simulations are fundamental for evaluating new transmission techniques and the key parame-

ters of wireless systems (e.g., capacity, latency, etc) and gaining insights into the potential technologies. Therefore, in this work a SL abstraction is implemented for QAS.

1.2 Objective

In this thesis, we design a SL abstraction of QAS as an efficient solution for RBs allocation in 5th generation (5G) networks. As services in wireless networks are classified in terms of their delivery requirements to RT and NRT, the ultimate goal of introducing our QAS is to minimize latency for RT traffic while pledging a sufficient data rate for NRT traffic. Furthermore, it should enhance the performance of the network in a way that does not severely deteriorate the average reliability among users. Moreover, it introduces moderate fairness among full buffer users that are assumed to always want to transmit data during the simulation period. Complying with QoS demands of different traffic models, QAS takes into account the packet buffer state of each user (e.g., how long a packet has been in the buffer, packet size, etc.) by weighting the estimated throughput of each user with latency and reliability parameters to tune its behavior. This way, we maintain the balance between the essential aforementioned performance metrics. This thesis answers a couple of questions, which are as follows:

- Does QAS achieve trade-off between throughput, latency, and reliability in wireless networks?
- Where do traditional scheduling strategies fail? How can QAS overcome these challenges and provide an efficient resource allocation?
- Does QAS perform well compared to other standard scheduling strategies with different cell loads?
- What are the impairments of QAS?

1.3 Outline

In light of the prior questions, the thesis is divided into chapters that answer the questions in organized manner. We start with introducing the state of the art scheduling strategies and their drawbacks in Chapter 2. Chapter 3 serves as a reference for some expressions used in this work. Chapter 4 covers the structure and key functionalities of 5G System Level Simulator (SLS) that belongs to the Vienna Cellular Communications Simulators (VCCS) family. We use this simulator for designing and evaluating the performance of QAS. Chapter 5 provides an introduction to the theory of mathematical optimization and convex optimization in particular. The latter allows us to solve RBs optimization problems efficiently.

Chapter 6 is divided into two main sections. In the first section 6.1, we provide an overview of several QoS classes and their main features. In the second section 6.2, we formulate the RBs allocation problem of QAS and explain the impact of the used tuning parameters. Chapter 7 is entirely focused on the network deployment considered in this work and the performance of QAS in terms of SL simulations, focusing on its benefits over classic scheduling strategies. We conclude this chapter with a discussion giving a critical reflection of our work.

1.4 Notation

The following notation is used throughout this thesis:

$\mathbb{R}$	real numbers
$\mathbb{R}^+$	non-negative real numbers
$\boldsymbol{X}$	matrix
$\boldsymbol{X}^{-1}$	inverse matrix
$\boldsymbol{X} \in \mathbb{R}^{N \times K}$	real-valued matrix with N rows and K columns
x	scalar
$\boldsymbol{x}$	vector
$\boldsymbol{x} \in \mathbb{R}^{N \times 1}$	length N real-valued column vector
$\mathcal{D}(\cdot)$	domain
$\|\cdot\|_2$	Euclidean norm
$\mathrm{vec}(\cdot)$	vec operation
$\boldsymbol{x}^\intercal, \boldsymbol{X}^\intercal$	vector/ matrix transpose
$\cdot^*$	optimal value
$\arg\max$	argument of the maxima
$\arg\min$	argument of the minima

2 State of the Art

This chapter provides a review of the literature of SL simulations and scheduling strategies, highlighting their advantages and potential downsides.

System Level Simulations

Measurement-based analysis in wireless communications becomes time consuming, challenging, and prohibitively expensive [31]. Therefore, SL simulations are a relatively fast, economical, and inevitable tool for understanding the interactions between network nodes and evaluating promising transmission techniques. They become crucial for evaluating the performance of developed radio access technologies such as Long Term Evolution (LTE) and 5G, testing, and optimizing algorithms before implementing to prevent deterioration of network performance. Accordingly, many SL simulation tools have been developed and widely used by academia and industry, some of them are mentioned below. Two popular discrete-event network simulators are OMNeT++ which is a modular simulation library and framework that can be used for building network simulators [39] and ns-3 that is used for Internet systems [10]. In [31] a simulator that focuses on vehicular communications is introduced. An open platform and a tractable testbed are offered in [35] to evaluate the SL performance of the 5G standard. In [26] the authors introduce a simulator for 5G mobile networks. Last but not least, VCCS offers a resilient simulation suite free of charge for academic use. It offers a 5G SLS [28] and a 5G Link Level Simulator (LLS) [29]. These simulators are implemented in MATLAB using object-oriented programming.

Classic Schedulers

Within the scope of wireless communications, user scheduling means the process of dynamically assigning radio resources to users according to a scheduling algorithm [14]. It plays a major role in maximizing the spectral efficiency and network capacity of the communications system. Scheduling strategies are carried out on the time and frequency grid, the so-called RBs.

In the following, we outline two of the most popular scheduling strategies. The RR scheduler distributes resources evenly among users without taking into account their channel conditions. In a circular order per scheduling cycle, users are scheduled over time as depicted in Figure 1. This scheduling strategy is simple and easily deployed. It offers fairness among users with respect to the amount of resources assigned to each of them but degrades the throughput performance [13]. The RR scheduler can be seen as a special case of the Weighted Round Robin (WRR) scheduler in which integer weights are configured for users. These weights deter-

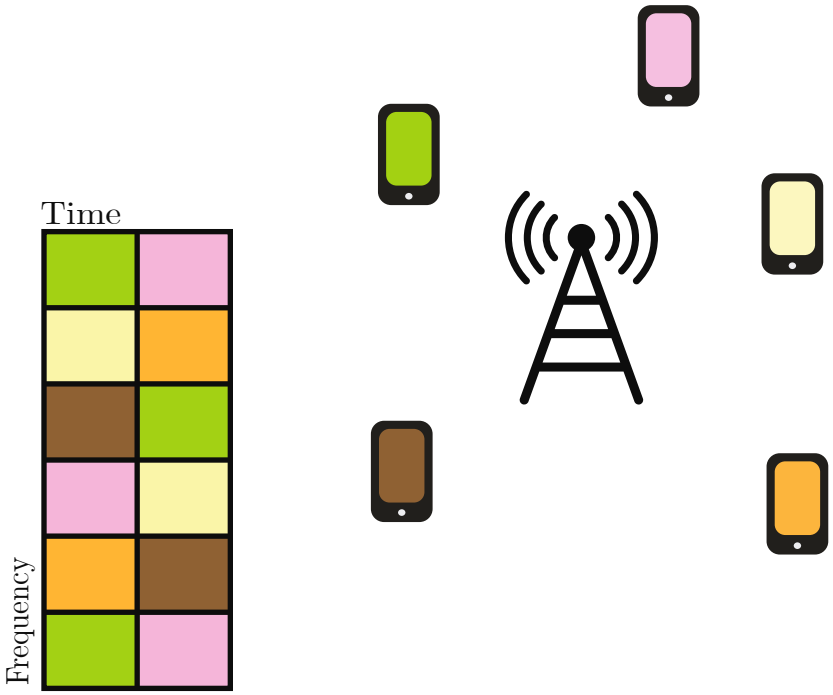


Figure 1: Illustrative example of RBs allocation by a RR scheduler.

mine the number of service opportunities for each user during a scheduling cycle [3].

The Best CQI scheduler assigns RBs to users with the best channel conditions; therefore, users close to the Base Station (BS) are more likely to be scheduled [13]. Due to the scheduling principle, users send CQI feedback to the BS, which uses this information to select the scheduled users, as illustrated in Figure 2. As the channel has high quality, a high data rate can be used. This results in high system throughput and network capacity at the expense of fairness in terms of resource division between users.

Quality of Service Aware Schedulers

There is a lot of previous work on QoS aware scheduling; for instance in [18], the Proportional Fair (PF) scheduler was modified to reduce end-to-end delay for RT users; however, this comes at the expense of throughput for NRT users. In [27], authors aim to minimize the latency for RT traffic while still offering a good level of throughput by proposing a Multiple Access Channel (MAC) scheduler. However, prioritizing users according to their Block Error Ratio (BLER) performance was

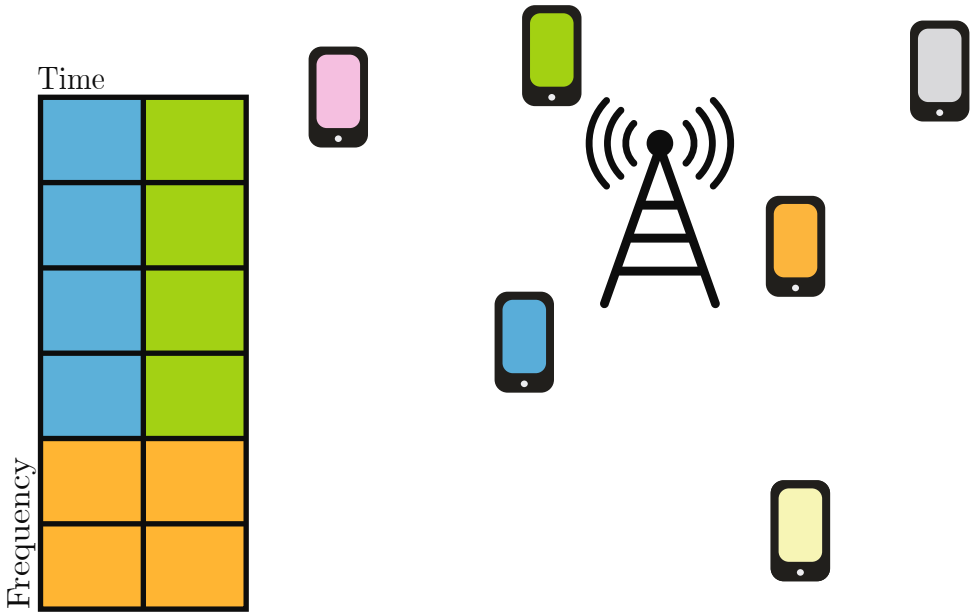


Figure 2: Illustrative example of RBs allocation by a Best CQI scheduler.

not addressed in the latter work. BLER refers to the ratio between the number of erroneous received code blocks and the total number of received code blocks in a data transmission. Furthermore, an extended Modified Largest Weighted Delay First (MLWDF) scheduling algorithm was developed with position-based parameters to enhance QoS of cell-edge users in [23]. The performance of the proposed algorithm deteriorates under high cell load. The resource allocation problem was formulated as Markov Decision Process (MDP) in [30]; therefore, a value iteration algorithm could be introduced to maximize the throughput of RT traffic. The work does not focus on improving the NRT throughput as well. Eventually, authors of [37], exploit channel conditions to improve the throughput performance of guaranteed bit rate and non-guaranteed bit rate traffics. In spite of this, they do not evaluate the performance of the proposed algorithm in terms of other performance metrics like the latency. In light of these limitations, the development of new QoS aware scheduling strategy emerges. An essential difference between our proposed scheduler and the works mentioned above, is the adoption of the linearized model in [12] which simplifies the RBs allocation optimization problem.

3 Terminology

For convenience, this chapter provides definitions of some terms used in this thesis. Summaries about these terms can be found in Section 4.7.

In wireless communications, fading is defined as the attenuation of a wireless signal due to various parameters that influence the quality of transmission over the communication link [14]; hence the Signal to Interference and Noise Ratio (SINR). Two popular types of fading a communication channel may experience are explained in the following and are depicted in Figure 3.

Large-Scale Fading

Large-scale or macroscopic fading represents the power attenuation of a wireless signal over distances large compared to the wavelength due to location dependent fading mechanisms, i.e., it is impacted by positioning of network elements and terrain layout. Macroscopic fading is modeled by four characteristics:

- Path loss: These models describe the attenuation of strength of a wireless signal as it propagates between the transmit and receive antennas.
- Antenna gain: As an antenna has a radiation pattern, its gain models the ability of the antenna to emit more or less in any direction.
- Shadow fading: Due to large objects blocking the propagation paths between the transmitter and the receiver, the signal strength is attenuated by the so-called shadow fading.
- Wall loss: The signal strength may be attenuated by the wall penetration loss of blockages obstructing a communication link.

Small-Scale Fading

Due to the existence of different propagation mechanisms, such as scattering, many wireless signals travel along different paths from the transmit to the receive antenna. This phenomenon is known as multipath propagation. Due to the interference of the different mulitpath components, a radio signal experience a rapid time-dependent fading over short distances in the order of the wavelength which is known as small-scale fading. Examples of small-scale fading models are Pedestrian A (PedA) and Vehicular A (VehA) [4].

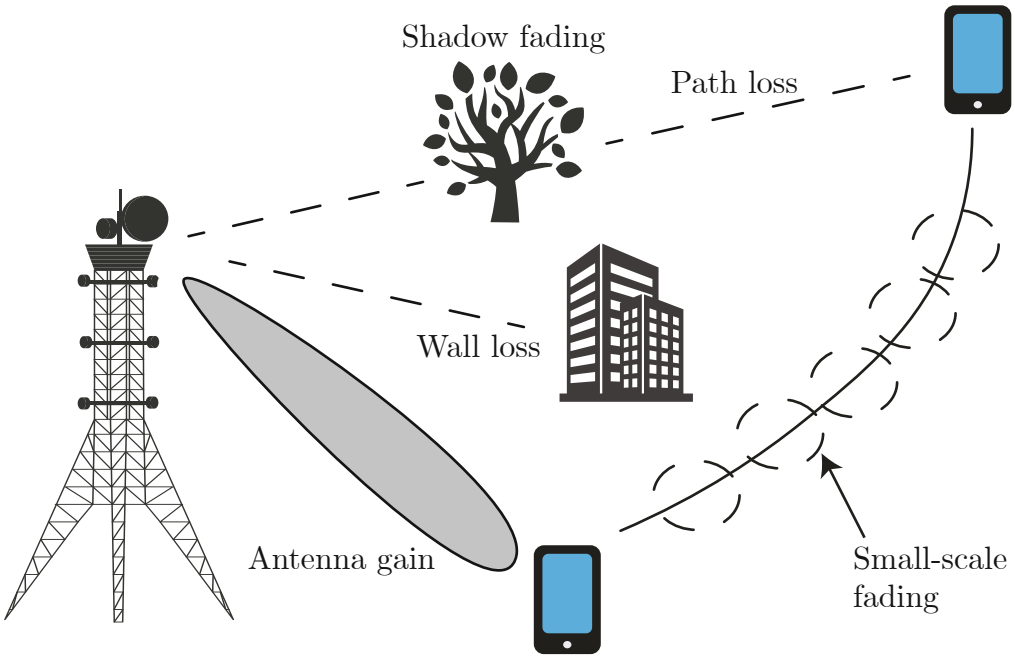


Figure 3: Sketch of large- and small-scale fading.

4 System Level Simulations

In this chapter, we describe the flow of SL simulations using the 5G SLS [28] as depicted in Figure 4. In particular, we introduce the simulation of Single User Multiple-Input Multiple-Output (SU-MIMO) Orthogonal Multiple Access (OMA) transmissions and characterize the main components of the simulator, to which we point out in Chapter 7. Simulations are defined in scenario files. These files contain a set of desired parameters that are related to the simulated region and time, the type of users and BSs, the large- and small-scale fading models, the type of transmission, and the storage option for the results.

4.1 Network Elements Generation

The simulator allows investigating the performance of large-scale wireless networks that are diverse in terms of network layouts and type of users and BSs. After the initialization part of the simulation is completed, during which the scenario parameters are chosen, an arbitrary number of users and BSs are defined, and the network geometry is created.

Blockages

The first step in network generation is placing walls and buildings within the simulated region. Cities' layouts are designed from buildings created by combining multiple walls. These walls have a loss property that determines the attenuation of a link passing through them. The term link refers to the radio connection between a transmitter and a receiver. Buildings specify whether a network element is indoor or outdoor, while streets allow for more realistic user positioning. Couple of city types can be created in the simulator, namely Manhattan and arbitrary cities from OpenStreetMap [40]. The latter type is used in this work. Figure 5 illustrates an example of a Manhattan city.

Users

Users are defined as a receiving or transmitting end point of the communication link. As the scope of our discussion is limited to the downlink, users represent the receiver. They are placed in the Region Of Interest (ROI) or the interference region according to multiple placement methods, such as stochastic Poisson Point Process (PPP) or at some predefined positions. As the simulated areas are finite, the region within the area is called ROI while users at its border experience the network interference from an interference region added to the simulation area to avoid border effects. Some of the main properties of a user are the number of receive

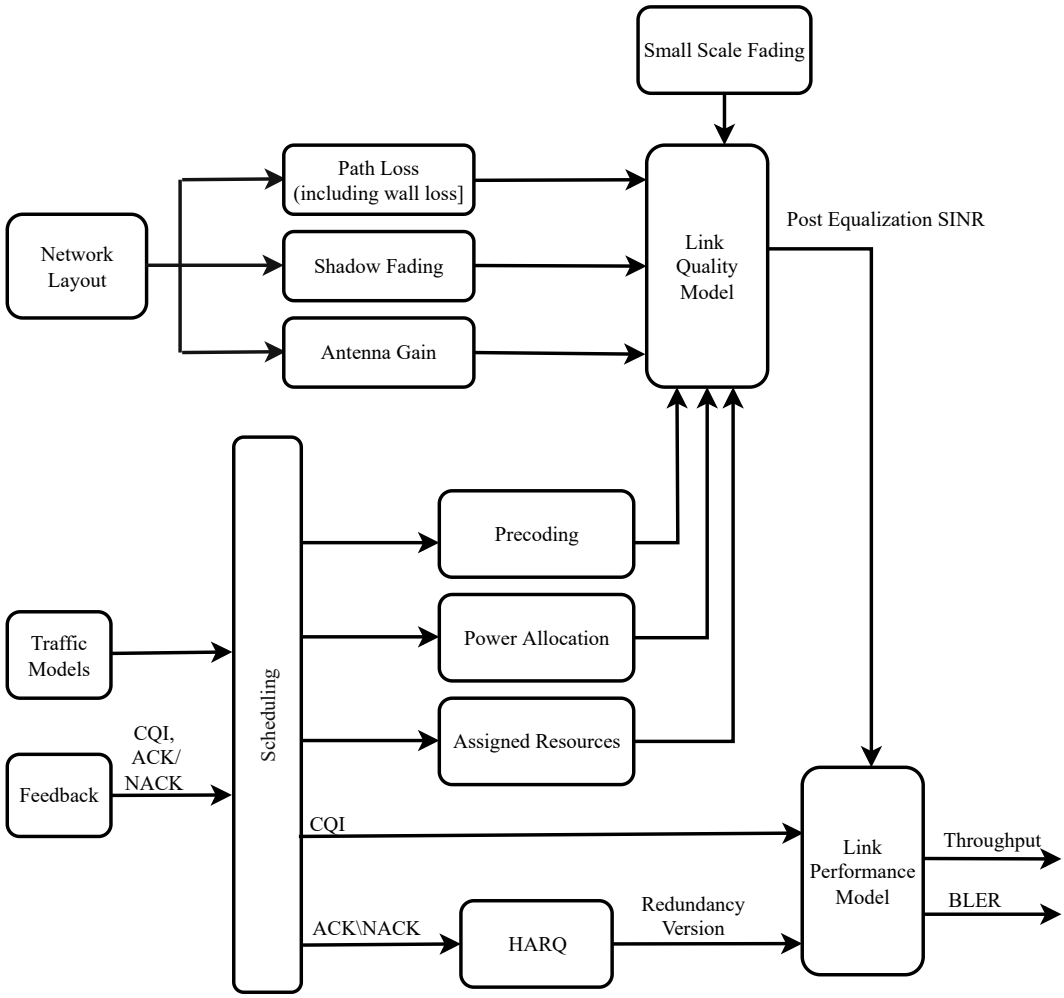


Figure 4: Schematic diagram of the 5G SLS of VCCS.

antennas denoted by n_{RX}, the small-scale fading model, and the data traffic model.

Base stations

The BS is the other end point of a communication link. Two essential properties of a BS are the transmit antenna denoted by n_{TX} and the baseband or digital precoder used for this BS. Example of antenna radiation patterns supported in the simulator are omnidirectional and three-sector. BSs are positioned according to a hexagonal grid, on top of buildings, in predefined positions or in random positions according to PPP. Analogously to the users, BSs are placed in the ROI or the interference region. Each BS has a one-to-one relationship with a scheduler

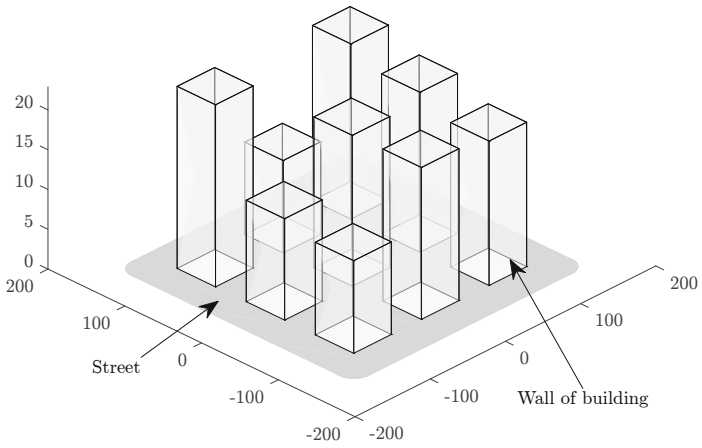


Figure 5: Placement of walls and streets in a Manhattan city.

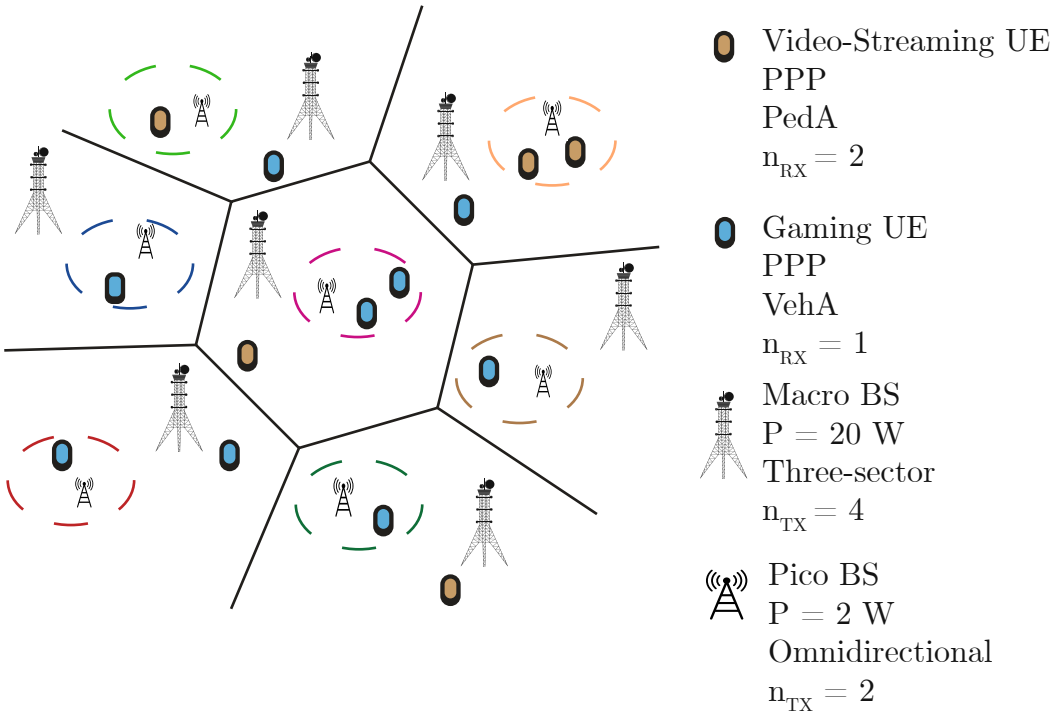


Figure 6: Sketch of cellular networks consisting of two types of BSs. The transmit power, the antenna type and the number of transmit antennas are the main properties emphasized for each BS. In addition, the network involves two types of users with different traffic models, user placements, small-scale fading channel models and number of receive antennas.

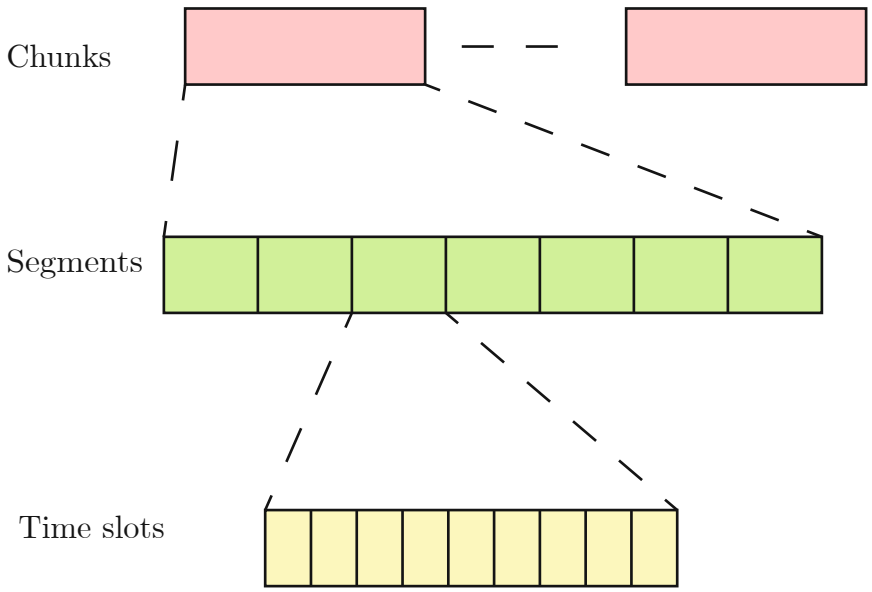


Figure 7: Simulator timeline consisting of time slots, segments, and chunks.

that allocates radio resources to the users attached to it. Macro, pico, and femto BSs with different transmit powers can be configured. The transmit power, denoted by P, is the maximum power used by a BS to transmit signals to a user. Figure 6 provides an example of a cellular network with different types of network elements.

4.2 Simulation Timeline

After the network elements are generated, the main simulation loop starts. The simulator timeline consists of three units, namely Time Slots (TSs), segments, and chunks. The TS is the smallest unit of length 1 ms. It represents the time during which the small-scale fading is assumed to be constant and corresponds to a LTE-A subframe. Multiple TSs form a segment during which the macroscopic fading is assumed to be constant. The duration of a segment depends on the correlation distance of macroscopic fading values, i.e., the distance a network element travels while large-scale fading values are constant. The user-to-BS association is set for every segment, since macroscopic fading parameters are needed to achieve the cell association. The association decision is made after evaluating a cell association metric. In this thesis, we use the macroscopic received power metric. The longest unit is the chunk which consists of a fixed number of TSs and may contain one or more segments. Considerably long distances are assumed between chunks in order to obtain uncorrelated user positions among chunks. Therefore, their simulations are assumed to be independent which allows for parallelization.

4.3 Simulation Loop

Within the main simulation loop over TSs of a chunk, most of the key functionalities reside. In the following, we explain the functionalities briefly.

4.4 Traffic Models

Various types of users with different applications are involved in cellular networks. Each of these users has a data traffic model according to its application. Thus, the simulator offers 3rd Generation Partnership Project (3GPP)-compliant traffic models that are described below. During the main simulation loop, the necessity of new data generation is checked for each user according to its traffic model statistics. Then, the packet buffer of the user is updated according to the number of bits a user would transmit, i.e., throughput. As a result, packet transmission latency is obtained at the end of the simulation. Below, we describe the traffic models shown in Figure 8:

- *Constant Rate*: This model generates packets of fixed size that are generated according to a packet arrival rate. It allows modeling various RT and NRT applications, e.g., vehicular communications, by specifying additional parameters such as the delay constraint of the application.
- *Full Buffer*: It is assumed that users have an infinite amount of data in their buffers at any time [22].
- *File Transfer Protocol (FTP)*: In this best-effort model, a sequence of file transfers separated by reading times is generated [6].
- *Hypertext Transfer Protocol (HTTP)*: In this interactive model, a web page consists of a main file and several embedded files [6].
- *Video Streaming*: At a regular time interval, frames of video data arrive. Each frame consists of random sized packets. Encoding delay intervals between the packets are introduced by the video encoder [6].
- *Gaming*: Users engaged in interactive gaming have packets separated with interarrival times and associated with User Datagram Protocol (UDP) headers [6].
- *Voice over IP (VoIP)*: Two-state Markov model represents the voice activity model of VoIP [6]. Adaptive Multi-Rate (AMR) is the audio codec used with a source rate of 12.2 kbps. During voice activity periods, VoIP packets are separated with encoder frame length. While Silence Insertion Descriptor (SID) packets are generated during silence periods.

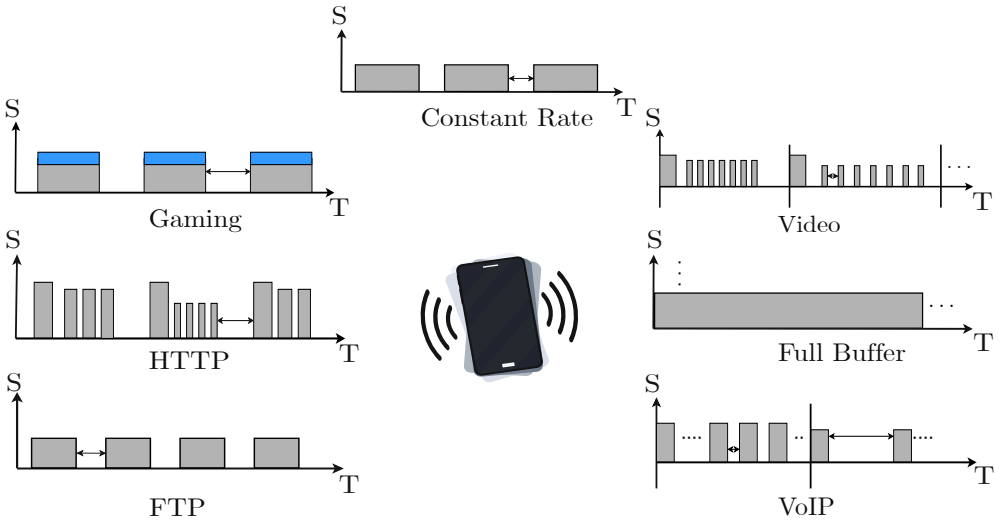


Figure 8: Illustrative sketch of traffic models of users where S denotes the traffic size generated over time which we refer to with T.

4.5 Scheduling

The scheduler embraces the functionality of the MAC layer. It utilizes the information from the user buffer provided by its traffic model statistics to determine resource allocation, namely transmit power allocation, RBs and appropriate Modulation and Coding Scheme (MCS) for transmission on each RB. Each scheduler is mapped to a BS exclusively which is responsible for allocating its users at every TS. The user allocation, as an output of the scheduler, is delivered to the Link Quality Model (LQM) of the simulator, which we elaborate in the next section, in order to calculate a performance measure, that is, post-equalization SINR. The scheduler framework is developed encompassing Hybrid Automatic Repeat Request (HARQ). The latter initiates retransmissions from a BS to its attached user in case of transmission failure. Transmission errors are declared by Acknowledged (ACK) or Non-Acknowledged (NACK) in the user feedback.

Classic Schedulers

Two standard schedulers are available in the simulator, namely the RR and the Best CQI that are described in Chapter 2.

Resource Grid

In the downlink of 5G networks, the structure of the radio frame is defined by Orthogonal Frequency Division Multiplexing (OFDM), which divides the fre-

quency selective channel into a set of orthogonal frequency flat channels. OFDM parameters, such as symbol duration, number of subcarriers, and subcarrier spacing, are defined by the resource grid. In OFDM, a physical RB, i.e., time-frequency resource, is allocated to a user and used for its transmission. Each RB is made up of a number of consecutive subcarriers in the frequency domain and OFDM symbols in the time domain. Different transmission numerologies are supported in the simulator; thus, different resource grids can be generated from a base grid. A numerology, which means a subcarrier spacing and symbol length, specify the size of a RB. Downlink transmissions are organized into radio frames that comprise 10 subframes, each of duration 1 ms similar to LTE. In 5G, each subframe is further divided into slots depending on the subcarrier spacing while in LTE each subframe consist of two slots. The term TS defined in Section 4.2 is actually the same as the subframe, namely the fundamental transmission time interval during which all data of one subframe/TS is jointly passed through the downlink signal processing chain of the system. Figure 9 illustrates a radio frame structure.

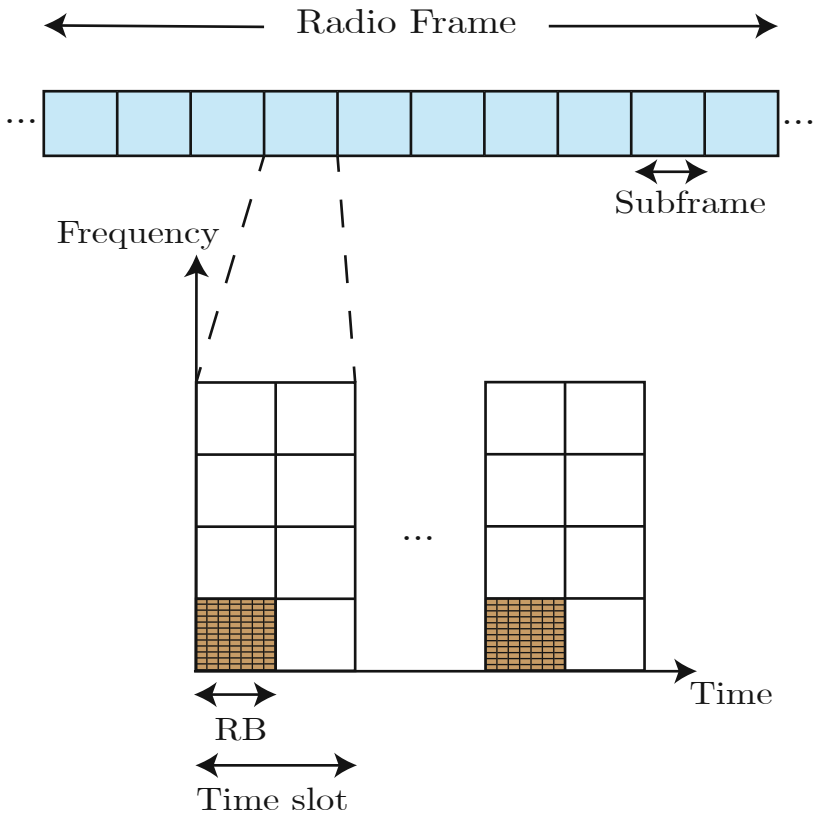


Figure 9: Resource grid structure used in the simulator.

4.6 Link Quality Model

Several steps in the Multiple-Input Multiple-Output (MIMO) downlink signal processing chain are abstracted in the LQM, which are the precoders, the channel, and the receive filter. The latter is assumed to be a zero forcing filter which minimizes the inter layer interference. As the name suggests, LQM is used to quantify the quality of the received signal after equalization [17]. LQM yields the post-equalization SINR calculated for each layer of each RB at which a user transmits. As multi-antenna transmission techniques are supported in LTE and 5G, this allows for simultaneous transmission for up to two parallel data streams, called codewords. These codewords of coded data bits go through several steps in the signal processing transmission chain and then are mapped to spatial transmission layers by means of layer mapping [25]. One codeword can be mapped on up to four transmission layers. In case of Single-Input Single-Output (SISO) transmission, one codeword and one layer are used.

4.7 Propagation Effects

Position Dependent Effects

Path loss models consider all possible combinations of BSs type, Line Of Sight (LOS)/Non Line of Sight (NLOS) connection between the transmitter and the receiver, and indoor/outdoor users. In the setup phase of each chunk before the main simulation loop, the path loss values are calculated. These values are then updated for each segment. Examples of path loss models that can be utilized in simulations are free space [14] and urban macro [19]. Both of these models are employed in this work. Each of the other macroscopic propagation effects namely wall loss, antenna pattern gain and shadow fading [24] are determined separately.

Time Dependent Effects

A power delay profile defines the strength of the signal received through a multipath channel as a function of time delay and is used to define channel models. These models describe small-scale fading in the simulator. Channel traces are generated before the main simulation loop for each of these channel models. Traces include all possible compositions of number of receive antennas, number of transmit antennas, channel model, maximum user speed, and carrier frequency. Channel models vary from stochastic channel models such as Rayleigh to Power Delay Profile (PDP)-based channel models such as typical urban [9], that is used in this thesis, and interface-based models such as QUAsi Deterministic RadIo channel GenerAtor (QuaDRiGa) [21].

4.8 Feedback

The scheduler needs various information about the channel state reported by the user feedback, such as Rank Indicator (RI), CQI and Precoding Matrix Indicator (PMI) [11] in order to determine the optimal transmission parameters for the following TSs. Furthermore, ACK and NACK information about the success of user transmissions is transferred by the feedback to the HARQ to perform retransmissions if necessary. In this work, HARQ feature is disabled.

4.9 Link Performance Model

Several steps in the receiver chain are abstracted in Link Performance Model (LPM), e.g., layer demapping and decoding. In this part of the simulator, the time-frequency selective post-equalization SINR values, calculated by the LQM, of a given user and transmission layer are mapped to an average SINR with equivalent Additive White Gaussian Noise (AWGN) performance. The averaged SINR is called effective SINR. Mutual Information Effective Signal to Interference and Noise Ratio Mapping (MIESM) is the method used for the averaging process, which uses calibrated capacity curves. Then, the effective SINR values are mapped with the CQI value used by the scheduler to a BLER value. For this, Signal to Noise Ratio (SNR)-to-BLER mappings for a AWGN channel are used. Each CQI in these mappings corresponds to specific MCS. After that, the number of transmitted bits is calculated as user throughput. Based on the resulting throughput, the success of the current transmission is determined. For a successful transmission, the number of bits transmitted is equal to the maximum size of the transport block for the scheduled RBs and otherwise it is equal to zero.

4.10 Results Post Processing

Right after the main simulation loop, post processing is performed. In this phase, the acquired results of the independent chunks are combined to an eventual result. The simulator offers types of postprocessors that differ in the number of functionalities performed and the results collected at the end of the simulation. Results are selectively stored based on the chosen settings and combined in average values. A set of graphs is offered from which a conclusion can be drawn from the simulation results, such as user throughput, SINR, transmission latency, and BLER.

5 Optimization

An overview of mathematical optimization is presented in this chapter. First, the formulation of a general optimization problem and some widely known and used subclasses of convex optimization, namely least-squares and linear programming, are introduced. Then, convex sets and functions and convex programming are presented. Next, quadratic programming is summarized. Eventually, non-linear programming is overviewed. This chapter is based, if not stated otherwise, on [5]. It can be used as a reference upon reading about the RBs optimization problem in the following chapter.

5.1 Mathematical Optimization

A mathematical programming or mathematical optimization problem is formed as follows

$$\begin{aligned} &\arg\min_{\boldsymbol{x}} \ f_0(\boldsymbol{x}) \\ &\text{subject to:} \\ &f_i(\boldsymbol{x}) \leq 0, \ i \in \{1, \dots, m\} \\ &h_i(\boldsymbol{x}) = 0, \ i \in \{1, \dots, p\} \end{aligned} \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{x} = \{x_1, \dots, x_n\}$ is the vector of the optimization variables. f_0 is the objective function, which is a $\mathbb{R}$-valued function on some subset of $\mathbb{R}^n$. The inequality constraint functions $f_i : \mathbb{R}^n \to \mathbb{R}$ are $\mathbb{R}$-valued functions in some subset of $\mathbb{R}^n$. Similarly, the equality constraint functions $h_i : \mathbb{R}^n \to \mathbb{R}$. The domain of the optimization problem $\mathcal{D}$ is the set of points for which the objective function f_0 and all constraint functions f_i, h_i are defined

$$\mathcal{D} = \bigcap_{i=0}^{m} \mathcal{D}(f_i) \cap \bigcap_{i=1}^{p} \mathcal{D}(h_i). \tag{2}$$

If a point $\boldsymbol{x} \in \mathcal{D}$ satisfies the constraints functions $f_i(\boldsymbol{x})$ in (1), then it is feasible. The optimization problem is feasible if there exists at least one feasible point and otherwise is infeasible. The feasible set is the set of all feasible points. $\boldsymbol{x}^*$ which has the smallest objective value among all vectors that satisfy the constraints in the problem (1) is called the solution of the problem. Equivalently, the maximization problem is solved by minimizing the objective function $-f_0$ subject to the constraints.

Optimization problems have become an essential tool in all quantitative disciplines, from engineering and network design to economics. Our ability to solve the problem

(1) is limited by many factors, such as the forms of the objective and constraint functions and the number of functions considered in the optimization problem. Therefore, compromises are often considered. Each class of optimization problems is characterized by particular objective and constraint functions. In the following, we will investigate a few forms.

5.2 Least-Squares Programming

The special thing about least-squares optimization problem is that it is an optimization problem with no constraints and has an objective function as follows

$$\arg\min_{\boldsymbol{x}} \parallel \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \parallel_2^2 = \sum_{i=1}^{k} (\boldsymbol{a}_i{}^\intercal \boldsymbol{x} - b_i)^2 \tag{3}$$

where $\boldsymbol{A} \in \mathbb{R}^{k \times n}$ with $k \geq n$. The vector $\boldsymbol{a}_i{}^\intercal$ indicates the rows of the matrix $\boldsymbol{A}$. The optimization variables are $\boldsymbol{x} \in \mathbb{R}^n$. The analytical solution of the least-squares problem is $\boldsymbol{x}^* = (\boldsymbol{A}^\intercal \boldsymbol{A})^{-1} \boldsymbol{A}^\intercal \boldsymbol{b}$. There are many good algorithms for solving least-squares problems reliably and precisely.

5.3 Linear Programming

The optimization problem is called linear when the objective function $f_0(\boldsymbol{x})$ and the constraints $f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})$ in (1) are linear, i.e., satisfy

$$f_i(\alpha \boldsymbol{x} + \beta \boldsymbol{y}) = \alpha f_i(\boldsymbol{x}) + \beta f_i(\boldsymbol{y}) \tag{4}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ and all $\alpha, \beta \in \mathbb{R}$. The linear programming is expressed in a standard form as

$$\begin{aligned} &\arg\min_{\boldsymbol{x}} \boldsymbol{c}^\intercal \boldsymbol{x} \\ &\text{subject to:} \\ &\boldsymbol{a}_i{}^\intercal \boldsymbol{x} \leq b_i, \ i \in \{1, \ldots, m\} \end{aligned} \tag{5}$$

where $b_i \in \mathbb{R}$ are scalars and $\boldsymbol{c}, \boldsymbol{a}_i \in \mathbb{R}^n$ are vectors. Unlike least-squares problems, there is no simple analytical formula for solving linear programming. However, effective methods exist to solve them, including the Dantzig simplex method [2] and the interior point methods [5].

5.4 Convex Programming

A special subfield of mathematical optimization problems is the convex optimization. It studies the problem of minimizing convex functions in convex sets. Simi-

larly, maximizing concave functions in convex sets. Thus, the objective and constraint functions are convex. This kind of functions is elaborated below. There are many classes that are convex optimization problems, or, via some transformations, they are reduced to convex optimization problems such as least-squares, linear programming, conic optimization and second-order cone programming. Similar to linear programming, there is no analytical solution of convex optimization problems; however, interior point methods are an effective method for this sake.

5.4.1 Sets

Since the convex optimization is an optimization problem to minimize convex functions on convex sets, it is necessary to get a glimpse of the convex sets. In order to do this, a line and line segment should be defined beforehand. Suppose that $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ are two points in $\mathbb{R}^n$ and the coefficient $\theta \in \mathbb{R}$. Then

$$\boldsymbol{y} = \theta \boldsymbol{x}_1 + (1 - \theta)\boldsymbol{x}_2 \tag{6}$$

forms the line that passes through $\boldsymbol{x}_1, \boldsymbol{x}_2$. The closed line segment corresponds to $\theta \in [0, 1]$.

Affine Set

In the light of prior information, a set $C \subseteq \mathbb{R}^n$ is affine if the line passing through two different points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$ lies in C, i.e., if $\theta \in \mathbb{R}$ then $\theta \boldsymbol{x}_1 + (1 - \theta)\boldsymbol{x}_2 \in C$. Thus, linear combination of any two points in C inclosed in C given that the coefficients sum up to one in the linear combination.

Convex Set

A set C is convex if the line segment between any two points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$ lies in C, i.e., if $\theta \in [0, 1]$ then $\theta \boldsymbol{x}_1 + (1 - \theta)\boldsymbol{x}_2 \in C$ as shown in Figure 10. Obviously, every affine set is a convex set as it contains the whole line between any two different points. Moreover, a convex set contains every convex combination of its points provided the coefficients sum to one and each coefficient $\theta \geq 0$. Some of the most important and relevant examples of convex sets are:

- The empty set $\varnothing$, any single point $\boldsymbol{x}_0$ and the space $\mathbb{R}^n$ are affine subsets of $\mathbb{R}^n$; hence convex.
- Any line is an affine set; hence convex .
- A line segment.
- A hyperplane is written as $\{\boldsymbol{x} \mid \boldsymbol{a}^\intercal \boldsymbol{x} = b\}$ where $\boldsymbol{a} \neq \boldsymbol{0}, \boldsymbol{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. It is an affine set; hence convex.

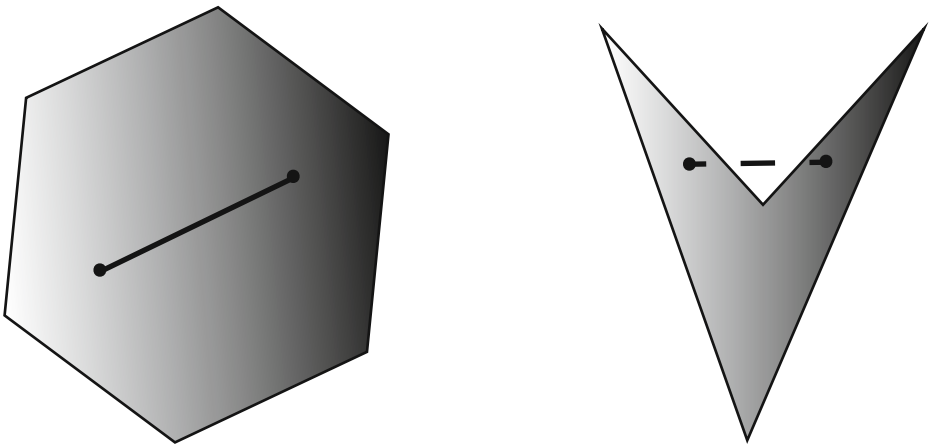


Figure 10: Convex and non-convex sets. On the Left, the hexagon is convex. Whereas the set on the right is non-convex since the line segment between the two points depicted as dots is not contained in the set.

- A halfspace formulated as $\{\boldsymbol{x} \mid \boldsymbol{a}^\intercal \boldsymbol{x} \leq b\}$ where $\boldsymbol{a} \neq \boldsymbol{0}$.

Cones

C is a cone if we have $\theta \boldsymbol{x} \in C$ for every $\boldsymbol{x} \in C$ and $\theta \geq 0$. Also, a set C can be called a convex cone if for any $\theta_1, \theta_2 \geq 0$ and $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$, we have

$$\theta_1 \boldsymbol{x}_1 + \theta_2 \boldsymbol{x}_2 \in C. \tag{7}$$

That is, C is a convex and a cone.

5.4.2 Convex Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ can be called convex if its domain $\mathcal{D}(f)$ is a convex set and

$$f(\theta \boldsymbol{x}_1 + (1-\theta)\boldsymbol{x}_2) \leq \theta f(\boldsymbol{x}_1) + (1-\theta) f(\boldsymbol{x}_2) \tag{8}$$

for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{D}(f)$ and $\theta \in [0, 1]$. Figure 11 illustrates an example of a convex function. One of the most popular examples of the latter is an affine function. It has an equality in (8). It is worth mentioning that all affine functions are convex and vice versa.

There are some transformations that preserve the convexity or concavity of functions; some are stated below:

- Scaling: If f is a convex function and $\alpha \geq 0$, then αf is convex.

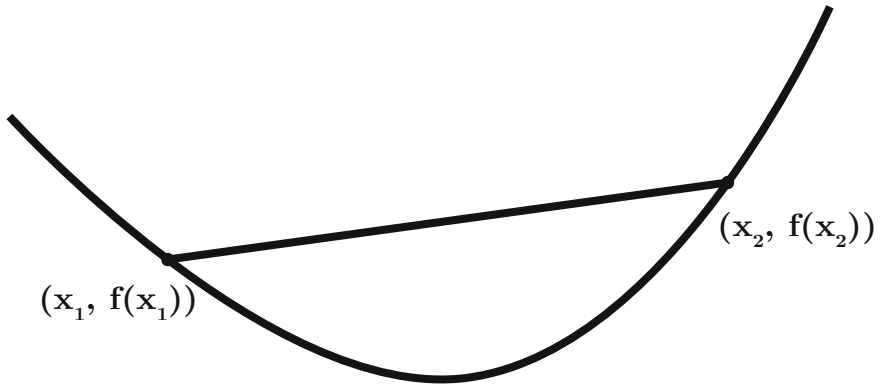


Figure 11: Example of a convex function.

- Addition: If f_1 and f_2 are convex functions, then $f_1 + f_2$ is convex.
- Non-negative weighted sum of convex functions $f = \sum_{k=1}^{K} w_k f_k$. Equivalently, a non-negative weighted sum of concave functions is concave.

5.4.3 Standard Form

The standard form of convex optimization problems is written as

$$\begin{aligned} &\arg\min_{\boldsymbol{x}} \ f_0(\boldsymbol{x}) \\ &\text{subject to:} \\ &f_i(\boldsymbol{x}) \leq 0, \ i \in \{1, \ldots, m\} \\ &\boldsymbol{a}_i{}^{\intercal}\boldsymbol{x} = b_i, \ i \in \{1, \ldots, p\} \end{aligned} \tag{9}$$

where $f_0(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})$ are convex, that is, satisfy (8). The feasible set of a convex optimization problem is convex because $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $\mathcal{D}(f_0)$ is convex, the m sublevel sets of convex functions $\{\boldsymbol{x} \mid f_i(\boldsymbol{x}) \leq 0\}$ are convex, and the p hyperlanes $\{\boldsymbol{x} \mid \boldsymbol{a}_i{}^{\intercal}\boldsymbol{x} = b_i\}$ are convex, and hence the intersection of convex sets is convex. Therefore, three conditions must be met according to (9):

- The objective function $f_0(\boldsymbol{x})$ is convex.
- The inequality constraints $f_i(\boldsymbol{x})$ is convex.
- The equality constraints $h_i(\boldsymbol{x}) = \boldsymbol{a}_i{}^{\intercal}\boldsymbol{x} - b_i$ is affine.

Equivalently, the maximization problem is

$$\begin{aligned} &\arg\max_{\boldsymbol{x}} f_0(\boldsymbol{x}) \\ &\text{subject to:} \\ &f_i(\boldsymbol{x}) \leq 0,\ i \in \{1,\dots,m\} \\ &\boldsymbol{a}_i{}^{\intercal}\boldsymbol{x} = b_i,\ i \in \{1,\dots,p\} \end{aligned} \tag{10}$$

where $f_0(\boldsymbol{x})$ is a concave function, and the problem can be solved by minimizing the convex objective function $-f_0(\boldsymbol{x})$.

5.5 Second-Order Cone Programming

We can call the optimization problem second-order cone if it can be expressed as

$$\begin{aligned} &\arg\min_{\boldsymbol{x}} \boldsymbol{f}^{\intercal}\boldsymbol{x} \\ &\text{subject to:} \\ &\parallel \boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{b}_i \parallel_2 \leq \boldsymbol{c}_i^{\intercal}\boldsymbol{x} + d_i,\ i \in \{1,\dots,m\} \\ &\boldsymbol{F}\boldsymbol{x} = \boldsymbol{g} \end{aligned} \tag{11}$$

where $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{A}_i \in \mathbb{R}^{n_i \times n}$ and $\boldsymbol{F} \in \mathbb{R}^{p \times n}$. In this programming, the first constraint is

$$\parallel \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \parallel_2 \leq \boldsymbol{c}^{\intercal}\boldsymbol{x} + d \tag{12}$$

is a second-order cone constraint, hence the name. The constraint requires the affine functions $(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b},\ \boldsymbol{c}^{\intercal}\boldsymbol{x} + d)$, where $\boldsymbol{A} \in \mathbb{R}^{k \times n}$, to lie in the second order cone in $\mathbb{R}^{k+1}$. Interestingly, the problem in (11) reduces to linear programming if $\boldsymbol{A}_i = \boldsymbol{0}, i \in \{1,\dots,m\}$.

5.6 Non-linear Programming

If the general optimization problem in (1) is not linear, i.e., the objective or constraint functions are neither linear nor convex, then it is called non-linear programming. In general, there is no effective formula for solving such problems; therefore, compromises are considered to solve non-linear programming.

6 Quality of Service Aware Scheduler Abstraction

We address our proposed QAS throughout this chapter in full detail. In the first section 6.1, QoS classes are categorized according to the set of network characteristics. In the second 6.2, the system model of QAS is presented by formulating the RBs allocation problem and introducing a powerful modeling tool to solve problems that incorporate convex programming.

6.1 Quality of Service Classes

The wide-ranging term QoS refers to the experience a user has over a network [36]. In the context of wireless networks, QoS is quantitatively measured using multiple features of the network service, such as

- Throughput: The rate at which data packets are delivered successfully between the transmitter and the receiver.
- Latency: The time which a packet takes to be transferred from one communication end point to the other.
- Packet loss ratio: The ratio of lost data packets due to transmission failures to the total number of packets sent on the network.

There are several classes of QoS to which services belong according to their delay sensitivity, throughput expectancy, packet loss ratio, etc. These classes are described below according to [32]

- Conversational: Applications that include telephony speech, e.g., Global System for Mobile communications (GSM), VoIP and video conferencing. Therefore, RT conversations, in which the data is generated in real time between speakers oblige a stringent low latency transmission as provided by human perception.
- Streaming: Applications in which users look at RT video or listen to RT audio. In these cases, the data is usually recorded beforehand. Thus, transmitting adequate number of packets to the user is sufficient for enabling RT streaming of the file. Exploiting the times of good channel conditions for providing high throughput for the user is possible in this class. However, this is not feasible for the conversational class since the data is not generated yet. Less strict requirements on delay compared to the conversational class are required; however, the delay variation between the data packets within a flow must be preserved.
- Interactive: The traffic is characterized by request-response patterns between

end users. The most popular instance of such traffic is web-browsing. The payload must be transmitted reliably; hence, a low packet loss ratio is a fundamental characteristic of this scheme.

- Background: Applications transmit and receive data in the background, e.g., downloading of databases. Thus, this class is delay-insensitive. Nevertheless, the content must be delivered with a low packet loss ratio.

6.2 System Model

6.2.1 Resource Blocks Optimization Problem

Considering the downlink of 5G wireless networks, we assume a set of K users and a set of N available RBs during one subframe/TS. The number of bits that the user would transmit, in other words, the throughput of user k at a RB n is denoted by $t_{n,k}$. The weighted sum throughput maximization problem is written with the following QoS constraints

$$\begin{aligned}
&\underset{\{b_{1,1},\ldots,b_{n,k},c\}}{\arg\max} \left(\sum_{k=1}^{K} \alpha^{-\beta_k}\ \sigma^{-\max\{d_{c,k}-d_k,0\}} \sum_{n=1}^{N} t_{n,k} b_{n,k} \right) + c \\
&\text{subject to:} \\
&b_{n,k} \in \{0,1\},\ \forall\, n,k \\
&\sum_{k=1}^{K} b_{n,k} = 1,\ \forall n \in \{1,\ldots,N\} \\
&\sum_{n=1}^{N} t_{n,k} b_{n,k} \geq c\,\gamma_k,\ \forall k \in \{\text{non full buffer users}\} \subseteq \{1,\ldots,K\} \\
&0 \leq c \leq 1
\end{aligned} \tag{13}$$

where $b_{n,k}$ denotes the RB n allocated to user k. The RB allocation is restricted to be binary, i.e., can be either 1 or 0. The second constraint function implies that every RB is assigned to one user at a time. The predetermined value γ_k denotes the total number of bits in the buffer of user k, this inequality constraint ensures that the number of RBs that are allocated to a user is sufficient for its need. Achieving γ_k for all users may be infeasible due to insufficient available RBs. To account for this, the variable c, which proportionally reduces the assigned RBs of all users to achieve feasibility, is included. The power allocation is assumed to be equal for every RB from the transmit antenna. The throughput for user k on RB n is estimated as follows

$$t_{n,k} = \varsigma_n \times \eta_n \tag{14}$$

where ς_n denotes the number of data symbols sent on RB n and η_n refers to the CQI efficiency on the same RB. The CQI efficiency means the number of data bits transmitted per symbol according to CQI estimated by the feedback on RB n. As the problem in (13) is restricted to SISO transmissions, the number of codewords is equal to one. Therefore, the number of layers per codeword is not considered in the rough estimate of throughput in (14) as it is equal to one.

The positive base σ in (13) is linked to the latency parameter. The exponent $d_{c,k} - d_k$ represents the difference between the characteristic delay constraint of user k, $d_{c,k}$, and the current delay of that user. Characteristic delay constraints are predefined for every traffic model. The current delay determines the difference between the current TS and the generation time of the oldest untransmitted packet of that user. When $d_{c,k} - d_k < 0$, the exponent is set to zero, so that the packets are not prioritized. In the case of NRT applications, the delay is not a critical constraint; therefore, the exponent is set to zero. Thus, the latency parameter is

$$\begin{cases} \sigma^{-\max\{d_{c,k}-d_k,0\}}, & \text{RT user} \\ 1, & \text{NRT user} \end{cases} \tag{15}$$

which is exponentially increasing for RT users with a maximum value of one. The closer the current packet delay of a user is to the delay constraint, the more priority this user will be given.

The reliability parameter $\alpha^{-\beta_k}$ decreases exponentially with a base α. The exponent β_k represents the average BLER over codewords of user k. Therefore, the traffic of high reliability users is prioritized. This raises the question of how users with low reliability are treated. For the latter users, the reliability parameter $\alpha^{-\beta_k}$ is equal to one and a lower CQI is used for the transmissions of these users compared to the CQI measured by feedback from their links. This is done to guarantee robust transmissions for them. Reliability is determined based on an average BLER threshold, which is a typical operating point for mobile communication systems, as follows

$$\begin{cases} \alpha^{-\beta_k}, & \beta_k \leq 0.1 \\ 1, & \beta_k > 0.1. \end{cases} \tag{16}$$

All of the parameters $\gamma_k, t_{n,k}, \alpha, \beta_k, \sigma, d_{c,k}, d_k \in \mathbb{R}^+$ and must be checked on TS

basis for every user. Rewriting the problem (13) in a vector notation results in

$$\begin{aligned}
&\underset{\{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_k,c\}}{\arg\max} \left(\sum_{k=1}^{K} \zeta_k \boldsymbol{t}_k{}^{\intercal} \boldsymbol{b}_k \right) + c \\
&\text{subject to:} \\
&\boldsymbol{b}(n) \in \{0,1\},\ \forall n \in \{1,\ldots,N\} \\
&\boldsymbol{b}_j{}^{\intercal} \boldsymbol{b}_i = 0,\ \forall i \neq j \\
&\boldsymbol{t}_k{}^{\intercal} \boldsymbol{b}_k \geq c\, \gamma_k,\ \forall k \in \{\text{non full buffer users}\} \subseteq \{1,\ldots,K\} \\
&0 \leq c \leq 1
\end{aligned} \tag{17}$$

where $\zeta_k = \alpha^{-\beta_k}\ \sigma^{-\max\{d_{c,k}-d_k,0\}}$ and $\boldsymbol{b}_k = [b_{1,k},\ldots,b_{n,k}]^{\intercal}$. Next, the following compact notation is borrowed from [12]

$$\boldsymbol{b} = \text{vec}\left(\begin{bmatrix} \boldsymbol{b}_1^{\intercal} \\ \boldsymbol{b}_2^{\intercal} \\ \vdots \\ \boldsymbol{b}_K^{\intercal} \end{bmatrix} \right) \in \{0,1\}^{N \cdot K \times 1} \tag{18}$$

$$\boldsymbol{A} = \overbrace{\begin{bmatrix} 1\ldots1 & 0\ldots0 & \ldots & 0\ldots0 \\ 0\ldots0 & 1\ldots1 & 0\ldots & 0\ldots0 \\ \vdots & & \ddots & \vdots \\ 0\ldots0 & \ldots & \ldots & \underbrace{1\ldots1}_{K} \end{bmatrix}}^{N \cdot K} \tag{19}$$

where $\boldsymbol{b}$ contains the RBs allocation for all users. The first K rows correspond to the first RB assigned to all users, the next K rows to the second RB assigned to all users, etc. The matrix $\boldsymbol{A} \in \{0,1\}^{N \times N \cdot K}$ ensures that no RB is allocated to two users at a time. As a result, the optimization problem is described as

$$\begin{aligned}
&\underset{\{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_k,c\}}{\arg\max} \left(\sum_{k=1}^{K} \zeta_k \boldsymbol{t}_k{}^{\intercal} \boldsymbol{b}_k \right) + c \\
&\text{subject to:} \\
&\boldsymbol{b}(n) \in \{0,1\},\ \forall n \in \{1,\ldots,N\} \\
&\boldsymbol{A}\boldsymbol{b} \leq \boldsymbol{1}_N \\
&\boldsymbol{t}_k{}^{\intercal} \boldsymbol{b}_k \geq c\, \gamma_k,\ \forall k \in \{\text{non full buffer users}\} \subseteq \{1,\ldots,K\} \\
&0 \leq c \leq 1.
\end{aligned} \tag{20}$$

Letting the constraint function $\boldsymbol{Ab} \leq \mathbf{1}_N$ strictly equal to 1 may be infeasible due to excluding the possibility of no users assigned to a RB; hence the use of inequality.

Fairness is a key performance metric that allows equal sharing of resources between users. Jain's fairness index, the commonly used metric in literature,

$$\mathcal{J}(x_1, x_2, \ldots, x_K) = \frac{(\sum_{k=1}^{K} x_k)^2}{K \sum_{k=1}^{K} {x_k}^2} \tag{21}$$

is widely used to measure fairness [1]. As $\boldsymbol{x}$ is a throughput vector and x_k is the throughput of user k on its RBs, the index is equal to one if the same throughput is achieved by all users. This indicates the greatest fairness. With decreasing fairness, it gets closer to zero. When all RBs are allocated to only one user $\mathcal{J} = 1/K$; thus the worst fairness is attained. Writing Jain's fairness index in our optimization problem context

$$\mathcal{J} = \frac{(\sum_{k=1}^{K} {\boldsymbol{t}_k}^{\intercal} \boldsymbol{b}_k)^2}{K \sum_{k=1}^{K} ({\boldsymbol{t}_k}^{\intercal} \boldsymbol{b}_k)^2}. \tag{22}$$

Therefore, a constraint is imposed to ensure fairness among users of the full buffer such that $\mathcal{J} \geq \mathcal{J}_\circ$ where $\mathcal{J}_\circ = 0.7$ is the desired fairness index. This value is chosen such that the feasibility of the RBs allocation problem is guaranteed using the network loads that has been studied in this work. Thus, if infeasibility arises, this value must be changed accordingly. As stated in [16], the constraint can be reformulated as a second-order cone constraint

$$\frac{(\sum_{k=1}^{K} {\boldsymbol{t}_k}^{\intercal} \boldsymbol{b}_k)^2}{K \sum_{k=1}^{K} ({\boldsymbol{t}_k}^{\intercal} \boldsymbol{b}_k)^2} \geq \mathcal{J}_\circ \Leftrightarrow \sum_{k=1}^{K} {\boldsymbol{t}_k}^{\intercal} \boldsymbol{b}_k \geq \sqrt{\mathcal{J}_\circ K} \| \, {\boldsymbol{t}_k}^{\intercal} \boldsymbol{b}_k \, \|_2. \tag{23}$$

Adding the fairness constraint to the problem in (20)

$$\begin{aligned}
&\underset{\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k, c\}}{\arg\max} \left(\sum_{k=1}^{K} \zeta_k {\boldsymbol{t}_k}^{\intercal} \boldsymbol{b}_k \right) + c \\
&\text{subject to:} \\
&\boldsymbol{b}(n) \in \{0, 1\}, \ \forall n \in \{1, \ldots, N\} \\
&\boldsymbol{Ab} \leq \mathbf{1}_N \\
&{\boldsymbol{t}_k}^{\intercal} \boldsymbol{b}_k \geq c \, \gamma_k, \ \forall k \in \{\text{non full buffer users}\} \subseteq \{1, \ldots, K\} \\
&0 \leq c \leq 1 \\
&\sqrt{\mathcal{J}_\circ K} \| \, {\boldsymbol{t}_k}^{\intercal} \boldsymbol{b}_k \, \|_2 \leq \sum_{k=1}^{K} {\boldsymbol{t}_k}^{\intercal} \boldsymbol{b}_k, \ \forall k \in \{\text{full buffer users}\} \subseteq \{1, \ldots, K\}.
\end{aligned} \tag{24}$$

Analyzing the RBs optimization problem in (24) shows that

- Considering the objective function, each term within the non-negative weighted sum is a linear (affine) function of the RBs. According to Section 5.4.2, affine functions are convex. Also, non-negative weighted sums preserve convexity; hence the objective function is convex.
- This inequality constraint $\boldsymbol{A}\boldsymbol{b} \leq \boldsymbol{1}_N$ can be written as $\boldsymbol{a}_n{}^\intercal \boldsymbol{b} - 1 \leq 0$ for $n = \{1, \ldots, N\}$ where $\boldsymbol{a}_n{}^\intercal$ are the rows of $\boldsymbol{A}$. The prior functions of RBs are affine; hence convex.
- Similarly, the inequality constraint functions $\boldsymbol{t}_k{}^\intercal \boldsymbol{b}_k \geq c\ \gamma_k$ are linear (affine) functions of the RBs; therefore convex.
- $\sqrt{\mathcal{J}_\circ K} \| \ \boldsymbol{t}_k{}^\intercal \boldsymbol{b}_k \ \|_2 \leq \sum_{k=1}^{K} \boldsymbol{t}_k{}^\intercal \boldsymbol{b}_k$ is a second-order cone constraint that is convex according to Section 5.5.
- The optimization problem involves continuous and binary variables $b_{n,k}$. Such a problem is called mixed binary integer programming.

If we drop the assumption that the optimization variables are restricted to binary and assume $\boldsymbol{b}_k \in \mathbb{R}^n$, then we can call our optimization problem convex.

6.2.2 Problem Relaxation

In order not to use a binary programming solving method, relaxation can be applied. Relaxation implies substituting each non-convex constraint with a looser convex one, that is, replacing the integer constraints with convex ones, solving the convex programming, and rounding the results [5]. Solving the relaxed optimization problem is more convenient than solving the original one. With reference to the problem in (24), the integer constraint $b_{n,k} \in \{0, 1\}$ can be relaxed to a linear one, i.e., $0 \leq b_{n,k} \leq 1$. Furthermore, if the matrix $\boldsymbol{A}$ is totally unimodular and the right hand side of its constraint is an integer valued vector, then every feasible solution is an integer [2]. Totally unimodular means a matrix in which every square invertible submatrix is unimodular, i.e., a square integer matrix that has a determinant $0, 1$, or -1. Consequently, the solution returned by the relaxed problem is guaranteed to be an integer. This applies to the matrix $\boldsymbol{A}$ in our optimization problem in (24).

6.2.3 CVX Tool

To solve the problem in (24), a MATLAB-based modeling system called CVX is used, which supports convex programming [20]. The tool allows optimization problems to be formulated using standard MATLAB expression syntax. It supports

Disciplined Convex Programming (DCP), which is a methodology to construct convex optimization problems in a proper format for CVX. Problems that abide by the ruleset imposed by DCP are quickly verified as convex and transformed to a solvable form. Otherwise, they would be inadmissible even when they are convex. CVX supports standard optimization problems such as linear programming and much more complicated ones. Solvers with distinct capabilities are included in CVX, e.g., SeDuMi, Gurobi, and Mosek.

In a nutshell, to initiate CVX, one must write a CVX specification into a MATLAB script, function, or command prompt. A specification encompasses MATLAB statements and particular CVX commands for announcing optimization variables and specifying constraints and objective functions. It is noteworthy that CVX supports Mixed Integer Disciplined Convex Programming (MIDCP) in which problems are non-convex. MIDCP and DCP comply the same convexity rules; however one or more variables are restricted to be integer in MIDCP. Some CVX solvers support MIDCP without a guarantee of low computational complexity for any moderately sized problem. To obey the DCP ruleset and formulate our optimization problem (24) in CVX, we introduce the following vectors

$$\boldsymbol{t} = \begin{bmatrix} \boldsymbol{t}_1^\mathsf{T} \\ \boldsymbol{t}_2^\mathsf{T} \\ \vdots \\ \boldsymbol{t}_K^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{K \times N} \tag{25}$$

$$\boldsymbol{\zeta} = \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_K \end{bmatrix} \in \mathbb{R}^{K} \tag{26}$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_K \end{bmatrix} \in \mathbb{R}^{K} \tag{27}$$

$$\boldsymbol{t}^* = \mathrm{vec}\left(\begin{bmatrix} \zeta_2 \boldsymbol{t}_1^\mathsf{T} \\ \zeta_2 \boldsymbol{t}_2^\mathsf{T} \\ \vdots \\ \zeta_K \boldsymbol{t}_K^\mathsf{T} \end{bmatrix} \right) \in \mathbb{R}^{N \cdot K \times 1} \tag{28}$$

where $\boldsymbol{t}$ includes the throughput of all users, $\boldsymbol{\zeta}$ holds the tuning parameters of all users' traffic, $\boldsymbol{\gamma}$ has the remaining data in all buffers, and $\boldsymbol{t}^*$ contains the tuned estimated throughput of all users. The first K rows correspond to the throughput that every user would transmit over its first RB and the next K rows to the throughput that every user would transmit over its second RB, etc. In addition, we introduce matrix $\boldsymbol{T} \in \mathbb{R}^{K \times N \cdot K}$

$$\boldsymbol{T} = \overbrace{\begin{bmatrix} \boldsymbol{t}_{11}\,0\,0\ldots0 & \boldsymbol{t}_{12}\,0\,0\ldots0 & \ldots & \boldsymbol{t}_{1N}\,0\,0\ldots0 \\ 0\,\boldsymbol{t}_{21}\,0\ldots0 & \ldots & 0\,\boldsymbol{t}_{2(N-1)}\,0\ldots & 0\,\boldsymbol{t}_{2N}\,0\ldots0 \\ \vdots & \ddots & & \vdots \\ \vdots & & & \vdots \\ 0\;0\;0\ldots\boldsymbol{t}_{K1} & 0\;0\;0\ldots\boldsymbol{t}_{K2} & \ldots & \underbrace{0\;0\;0\ldots\boldsymbol{t}_{KN}}_{K} \end{bmatrix}}^{N \cdot K} \tag{29}$$

which contains the throughput of all users on all RBs. Eventually, the problem in (24) is formed in CVX as

$$\begin{aligned} &\arg\max_{\boldsymbol{b},c} \boldsymbol{t}^{*\intercal}\boldsymbol{b} + c \\ &\text{subject to:} \\ &\boldsymbol{b}(n) \in \{0,1\},\ \forall n \in \{1,\ldots,N\} \\ &0 \le c \le 1 \\ &\boldsymbol{A}\boldsymbol{b} \le \boldsymbol{1}_N \\ &\boldsymbol{T}\boldsymbol{b} \ge c\boldsymbol{\gamma} \quad \forall\{\text{non full buffer users}\} \\ &\sqrt{\mathcal{J}_\circ K}\,\|\,\boldsymbol{t}_k{}^{\intercal}\boldsymbol{b}_k\,\|_2 \le \sum_{k=1}^{K} \boldsymbol{t}_k{}^{\intercal}\boldsymbol{b}_k,\ \forall\{\text{full buffer users}\}. \end{aligned} \tag{30}$$

6.2.4 Tuning Parameters

Figure 12 shows the behavior of the latency parameter $\sigma^{-\max\{d_{c,k}-d_k,0\}}$ introduced earlier in (13) using multiple delay constraints. The tuning base σ is equal to 1.05 precisely. The reason stands behind this choice is that values higher than 1.05 eliminate the smooth growth of the latency parameter curve. Instead, a flipped L-shaped curve would result. This means that a significant range of small current delays d_k of user k would correspond to low values of the parameter in the flat region of its curve. Thus, parameter values correspond to distinct small current delays would barely differ. On the other hand, if σ lies in the interval $(1, 1.05)$,

curves of the parameter would shrink. Therefore, the current delay of the users would be assigned to a smaller scope of parameter values.

The behavior of the reliability parameter $\alpha^{-\beta_k}$ using the tuning base α of 2 is represented in Figure 13. If the base is larger, the range of the reliability parameter would approach values smaller than 0.5 and its curve tail would be flatter. Therefore, very low values of the parameter would be assigned to a large number of users with a high BLER. Therefore, the difference in the impact of tuning achieved by the parameter on scheduling decision would hardly be observed among these users. On the flip side, choosing a parameter in the interval $(1, 2)$ would narrow the parameter values zone.

6.2.5 Verification

To verify QAS abstraction introduced in Section 6.2.1, the output of each function, either used while developing QAS or influenced by it, is checked using the unit test framework [38]. Thus, we ensure that all newly implemented features are developed properly and coincide with the work of other parts of the simulator.

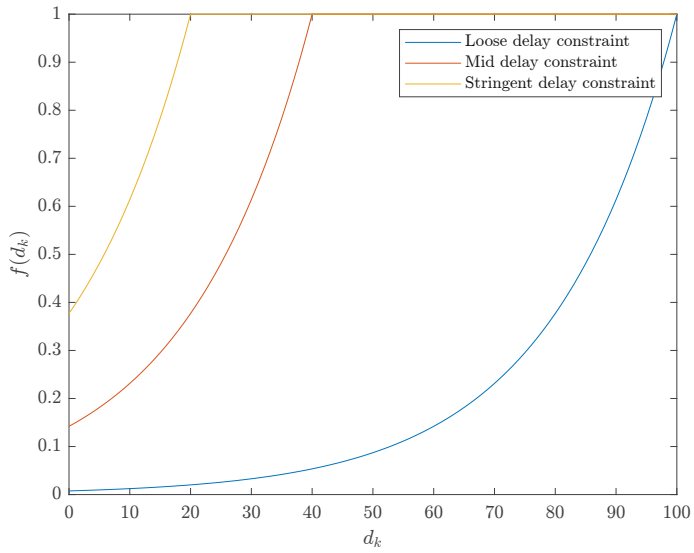


Figure 12: Latency parameter curves using various delay constraints. The function $f(d_k) = \sigma^{-\max\{d_{c,k}-d_k,0\}}$. The yellow curve corresponds to a delay constraint that is equal to 20. The red and blue curves correspond to delay constraints that are equal to 40 and 100 respectively.

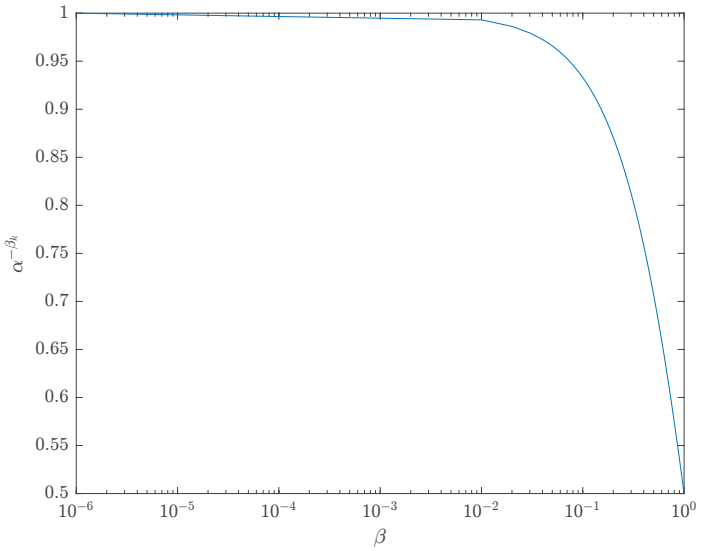


Figure 13: Reliability parameter curve over BLER values lie in the interval $(1e-6, 1)$.

7 Performance Evaluation

SL simulations of QAS are carried out using the 5G SLS [28]. In this chapter, two different scenarios are considered for QAS performance validation. In Sections 7.1 and 7.2, simulation-specific parameters are presented and simulation results of QAS are compared to those of classic scheduling strategies namely RR and Best CQI. Afterwards, we discuss insights into scheduling with QoS constraints in urban environments, problems of implementation of QAS, and some open issues in Section 7.3.

7.1 Outdoor Simulation Scenario

This section addresses the first simulation scenario. A hexagonal grid of macro BSs with 500 m distance between neighboring BSs is considered for network deployment. All users are located outdoors according to PPP where vehicular users move at a speed of 30 km/h in random trajectories. and other users are static, as shown in Figure 14. In this scenario, we use a typical urban environment [9]. To reveal the effectiveness of QAS, we consider users with different RT and NRT traffic, and each user is assigned a traffic model. Five types of traffic are imposed on the network, namely *vehicular*, *VoIP*, *video streaming*, *HTTP*, and *full buffer*. Each of these models is described in Section 4.4 and implies some constraints that we elaborate below. Figure 15 depicts the system model of this work.

HTTP users are NRT interactive users that need sufficient data rates to browse web-pages, and the content of their packets must be reliably transmitted [36]. A minimum data rate that achieves constant streaming with a loose delay constraint is desirable for video streaming users. Thus, we set a delay constraint of 100 ms for this RT application. Furthermore, the model is modified according to [7] to achieve a higher data rate of 90 kbps. As the VoIP user is considered in an outage if 98% of the packets of this user are delivered with a delay greater than 50 ms [6], we define a strict delay constraint of 40 ms for this RT model. The last RT application considered in our traffic mix is vehicle-to-network. The latter means the communication between a User Equipment (UE) and an application server, both supporting vehicle-to-network applications, communicating with each other through a cellular network (e.g., LTE or 5G) [8]. Such a safety-critical model imposes some latency, packet arrival rate, and packet size requirements to ensure a reliable communication; however, it does not have stringent data rate requirements. Therefore, we consider a maximum delay of 20 ms for vehicular communications. Finally, some users are modeled in full buffer mode [22] as they have sufficient data available to fully utilize the network. They are best-effort users that are not as reliable as other traffic models and there is no guarantee of delay-free delivery

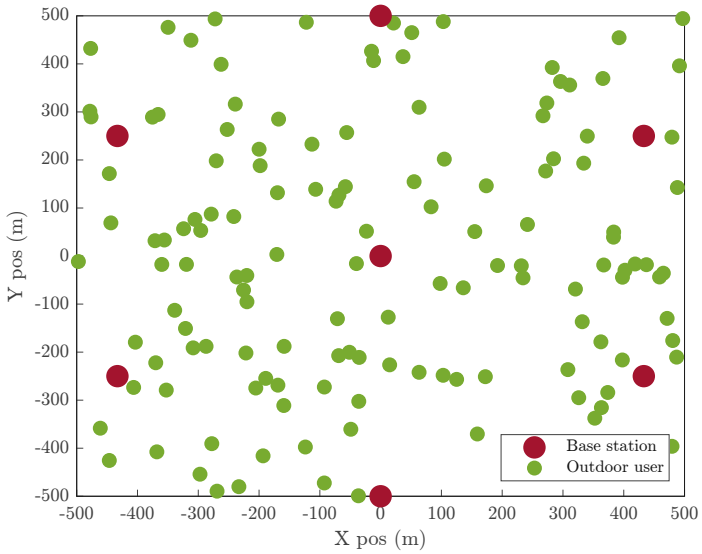


Figure 14: Network deployment of the outdoor scenario.

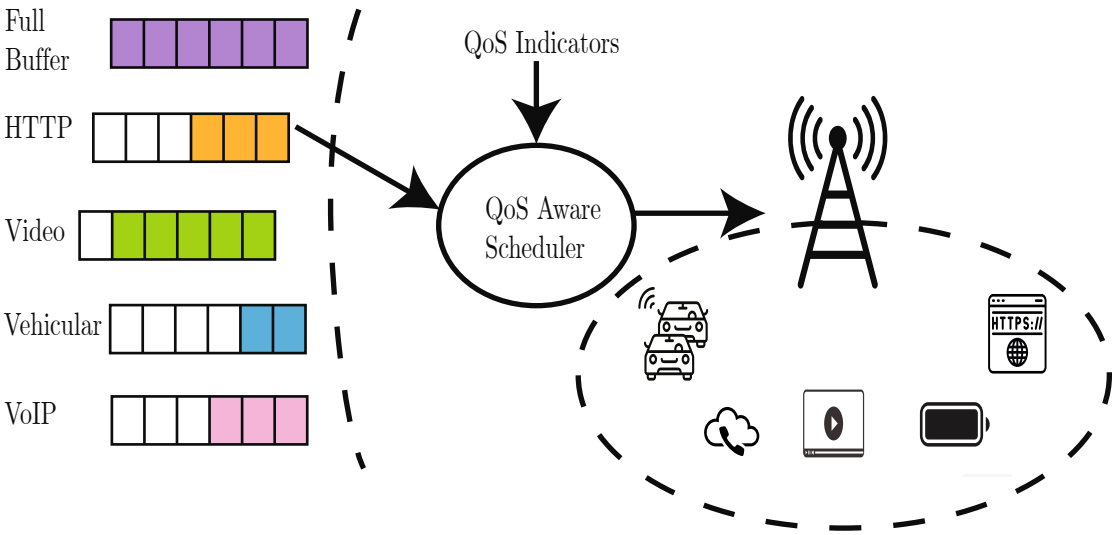


Figure 15: Proposed system model.

Table 2: Main simulation parameters of the outdoor scenario

Transmission parameters	
TS duration	1 ms
Center frequency	2 GHz
Bandwidth	20 MHz
Simulation duration	2000 TS
Network elements properties	
Number of users	140, 210, 420 users
Number of BSs	7 BSs in a hexagonal grid
BS antenna	1 omnidirectional antenna
User antenna	1 omnidirectional antenna
Traffic models	Vehicular, VoIP, video streaming, HTTP, and full buffer
Link properties	
Channel model	Typical urban
Path loss model	Free space

of their data. The main simulation parameters are reported in Table 2.

7.1.1 Simulation Results

In the following, the behavior of schedulers under different network loads is investigated. In the first simulation, 140 users are considered as follows, 24% vehicular, 24% VoIP, 24% video streaming, 24% HTTP, and 4% full buffer users. In the second and third sets, 210 and 420 users are considered respectively with the same percentage of users belonging to each traffic category.

Throughput

One of the QoS indices that we try to maximize for the aforementioned services is the user throughput. The resultant average and sum throughput for all traffic categories under different cell loads are shown in Figure 16. The results of our proposed scheduler are compared to those from RR and Best CQI strategies. After the throughput of each user is calculated in bits per second for every TS in the simulation, the average throughput of user k is determined as

$$t_k{}^{\text{avg}} = \frac{\sum_{s=1}^{S} t_{s,k}}{S} \tag{31}$$

where $t_{s,k}$ is the throughput assigned to user k in TS s. The total number of TSs is denoted by S. Then, the average throughput of all users belong to a traffic

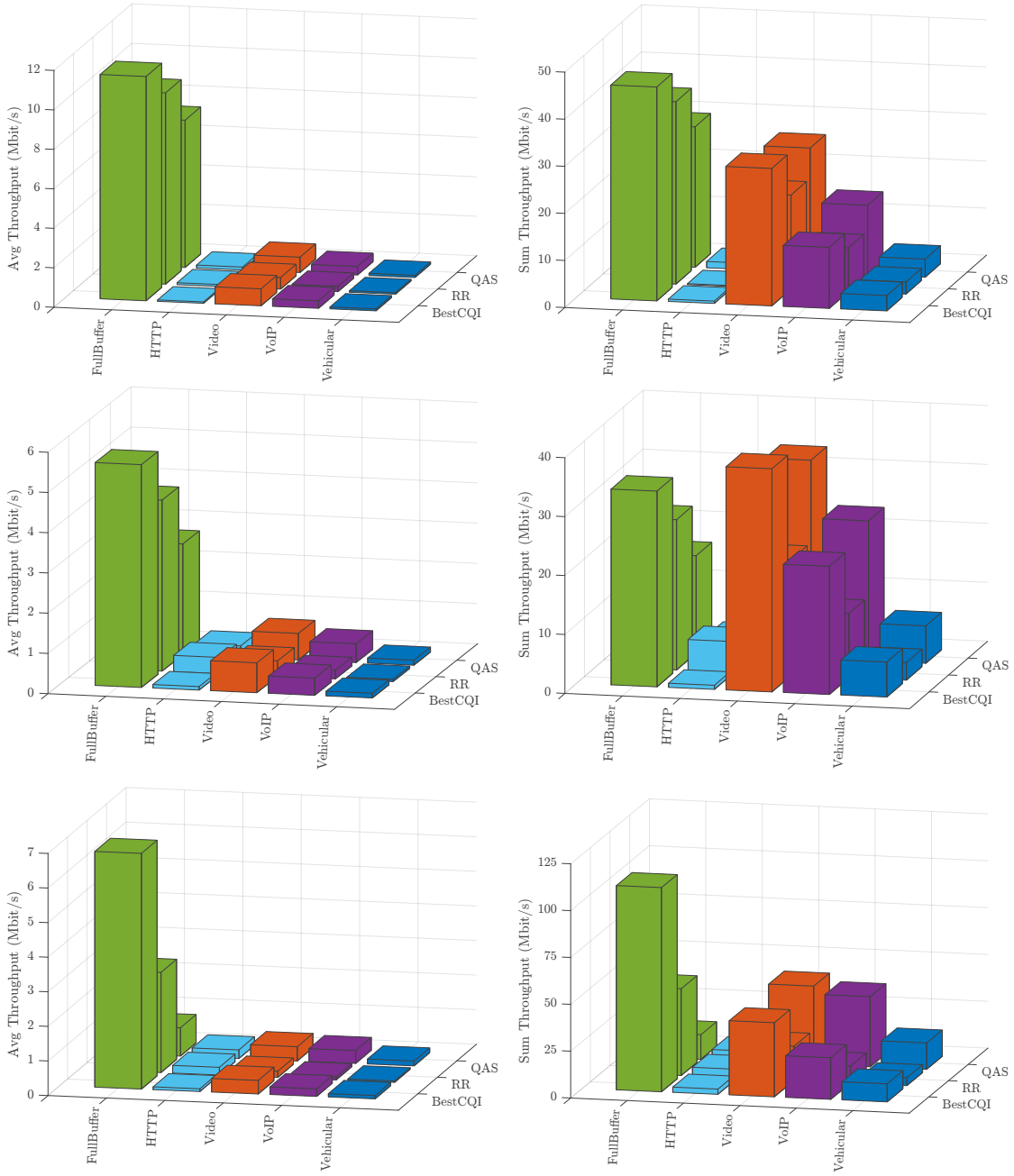
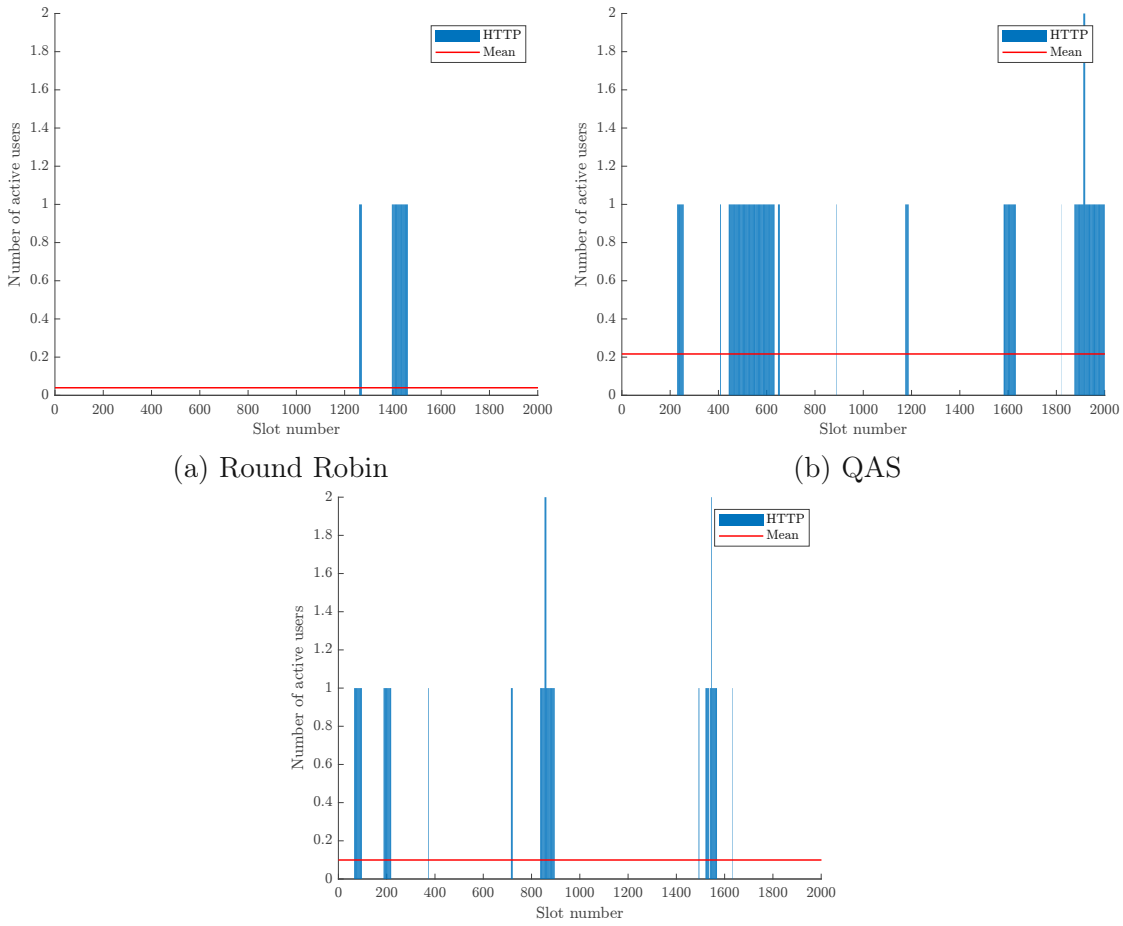


Figure 16: Average and sum throughput per traffic model and scheduling strategy for 140 (top), 210 (middle), and 420 (bottom) users consecutively.



(a) Round Robin

(b) QAS

(c) Best CQI

Figure 17: Number of active HTTP users per TS and scheduling strategy in a simulation of 140 users.

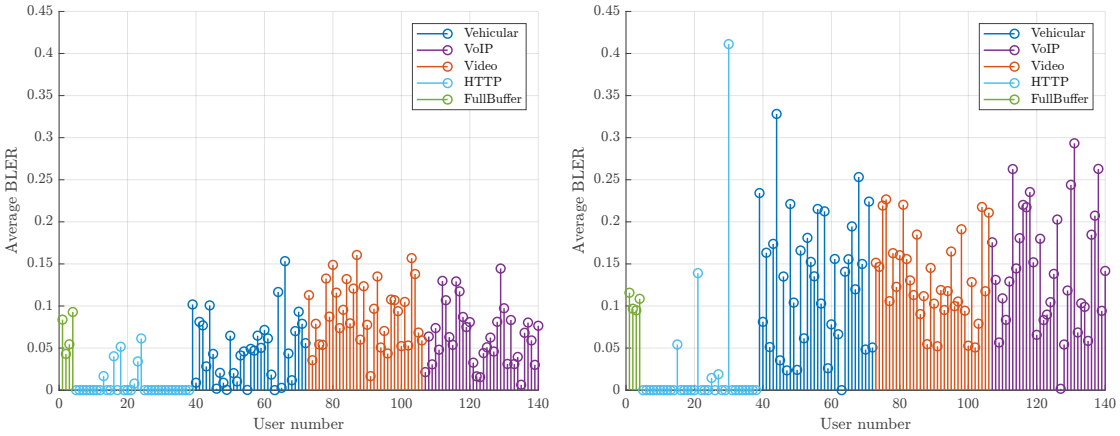
category is computed as follows

$$\mathcal{T}^{\text{avg}} = \frac{\sum_{k=1}^{K} t_k{}^{\text{avg}}}{K} \tag{32}$$

where K is the number of users. The sum throughput is calculated in bits per second by adding the throughput allocated to each user belong to the category over time as written below

$$\mathcal{T}^{\text{sum}} = \frac{\sum_{k=1}^{K} \sum_{s=1}^{S} t_{s,k}}{S}. \tag{33}$$

First, we consider the simulation of 140 users at the top of Figure 16. When studying the performance of the Best CQI scheduler, it can be seen in Figure 17c



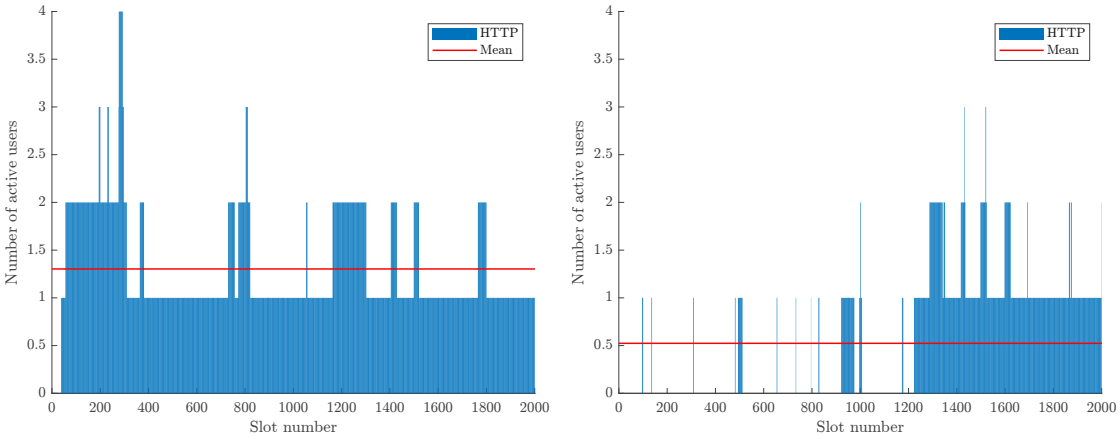
(a) QAS (b) Best CQI

Figure 18: Average BLER per traffic category for 140 users using QAS and the Best CQI scheduler.

that HTTP users have a low number of active users per TS due to the limited simulation duration. Looking at the average BLER of the users in Figure 18b, it is clear that many HTTP users have a zero average BLER because they are inactive; hence unserved. Some of the active HTTP users have a slightly high average BLER that results in the lowest sum and average throughput. Generally speaking, vehicular users have a high average BLER due to movement which let the CQI feedback about their links outdated; therefore, the assigned MCS does not fit to their actual channel conditions and their transmissions are unrobust. This leads to low average throughput. On the other hand, the highest sum and average throughput of full buffer users are due to their lowest average BLER. The reliability of video streaming users is slightly better than VoIP, so higher average and sum throughput.

Regarding the RR scheduler, allocating users fairly gives rise to less differences between the sum throughputs among different traffic categories. As Figure 17a shows, HTTP users are highly inactive, as a consequence, they get the lowest sum throughput. Vehicular users also have low user activity per TS, as they have a low packet arrival rate, resulting in the lowest average throughput. Throughput performance of full buffer users is similar to the Best CQI. In terms of user activity, video streaming users produce larger packets and much more frequently than VoIP; therefore, a higher sum and average throughput. The RR scheduler performs worst in the context of the sum throughput provided for all traffic categories, compared to other schedulers, except for the full buffer.

As expected, the performance of QAS for the strict delay constraint categories,



(a) Round Robin (b) QAS

Figure 19: Number of active HTTP users per TS using QAS and the Best CQI scheduler in a simulation of 210 users.

namely vehicular and VoIP, outperforms the other two strategies despite the slightly high average BLER of both categories, as shown in Figure 18a. This is due to the priority given to them by their delay constraints and the reliability enhancement technique of QAS. HTTP users obtain the lowest sum throughput since they are highly idle as shown in Figure 17b. However, throughput performance of HTTP outperforms other scheduling methods due to the considerably large packet size. Video streaming users also have relatively large packets and a high packet arrival rate as well; nevertheless, the average throughput provided by QAS is 10% less than what is allocated by the Best CQI scheduler. The reason behind this is that QAS is designed to serve users with sufficient throughput that matches their needs, while the Best CQI scheduler continuously serves them with high throughput as long as they experience good channel conditions. In view of the fact that video streaming users have a larger packet size and higher packet arrival rate than VoIP, they get a larger sum and average throughput. Finally, full buffer users acquire low sum throughput, as they do not provide any strict data delivery requirements that must be met; however, they get the highest average throughput due to their high reliability.

The performance of the second simulation of 210 users, can be observed in the middle of Figure 16. Generally speaking, the same behavior can be observed by scheduling strategy. With respect to the Best CQI scheduler, one can notice an increment of sum throughput for video streaming users compared to full buffer since the increment in number of users belonging to video streaming model is higher than in the latter. Also, the RR scheduler transmits more throughput to

Table 3: Improvement factors of average throughput achieved by introducing QAS.

Traffic model	140 UEs		210 UEs		420 UEs	
	RR	Best CQI	RR	Best CQI	RR	Best CQI
Vehicular	1.5	1.18	2.13	1.07	3.33	1.46
VoIP	1.59	1.14	2.18	1.09	4.10	1.69
Video streaming	1.32	0.90	1.71	0.89	2.27	1.06
HTTP	1.51	2.11	0.7	3.25	1.12	3.13

HTTP users as their user activity increases, as can be seen in Figure 19a. Unsurprisingly, QAS outperforms other strategies in the performance of VoIP and vehicular communications. HTTP users have a substantial packet size; however, the average throughput transmitted by QAS is 30% less than what is granted by the RR scheduler. This is because HTTP users are more active per TS while using the RR scheduler in this simulation as shown in Figure 19b.

As maintaining a high throughput at high loads is essential, we examine the last simulation of 420 users. The results can be observed at the bottom of Figure 16. A careful look at the figure reveals that QAS is successful in achieving its main objectives. Table 3 summarizes the results of average throughput offered by QAS divided by the average throughput transmitted by the two benchmarks for all simulations. Thus, the throughput gain or loss achieved for each traffic category can be seen numerically.

It should be noted that in this work, we consider a single random realization of users' positions for each simulation, i.e the throughput results are not averaged over multiple random realizations of users' positions. Therefore, one can observe that there is a throughput improvement using QAS for each simulation individually; however, there is no obvious trend of its behavior that matches the gradual increase in the network load. This non-monotonic behavior is quite clear when considering the performance of the Best CQI scheduler since it depends on the channel quality of the users that is related to their random positions and not only on the number of active users per TS as it is the case when using the RR scheduler.

Reliability

As reliability is essential for data delivery over wireless networks, we try to optimize the RBs allocation such that it is enhanced. The mean BLER per scheduling strategy for 140, 210 and 420 users is examined. The average BLER over the codewords of each user is calculated in every TS in the simulation and then the values are averaged among all users. Generally speaking, reliability degrades as the number of users increases, as shown in Table 4. The proposed scheduling strategy

Table 4: The mean BLER calculated among all users per scheduling strategy.

Scheduling strategy	140 UEs	210 UEs	420 UEs
RR	0.07	0.06	0.07
QAS	0.06	0.09	0.11
Best CQI	0.13	0.16	0.22

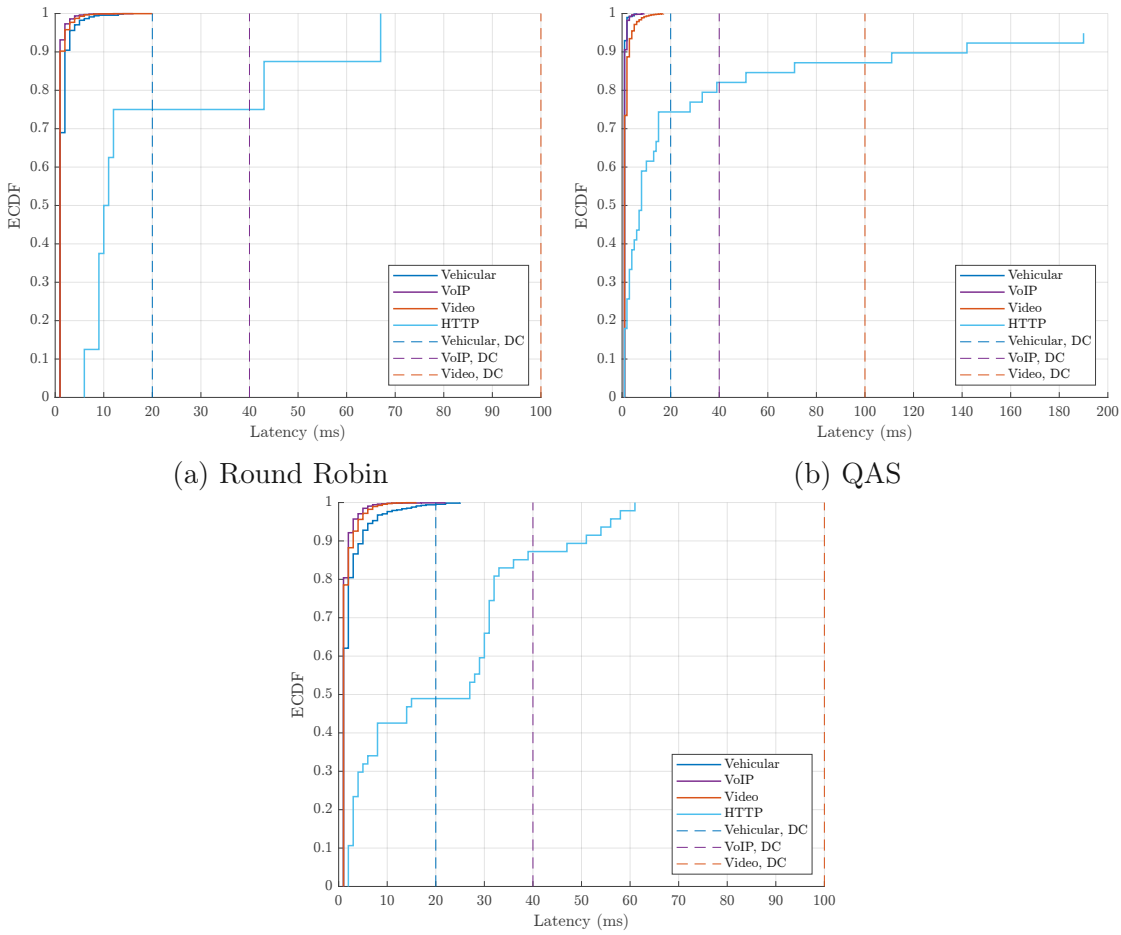
performs well in the first simulation. In the following simulations, QAS continues to boost the reliability of the network under high load; however, its performance is noticeable to be somewhere in the middle between the other two strategies since it prioritizes more low reliability users, e.g., vehicular users.

Latency

The last QoS indicator that we focus on in this work is the latency of RT applications. Delay performance is investigated for all traffic categories except full buffer, as their packet size and thus latency is assumed to be infinity. The packet latency is computed as the difference between its generation and successful transmission times. To obtain information on the performance of each traffic category, we examine the ECDF over latency measured in milliseconds.

First, we show the simulation results of 140 users in Figure 20. The vertical dashed lines indicate the predefined delay constraints, that are indicated by DC, of RT applications namely 20 ms for vehicular communications, 40 ms for VoIP and 100 ms for video streaming. One can see that neither QAS nor the RR scheduler violate the delay constraints. However, the Best CQI scheduler violates the delay constraint of vehicular communications as the maximum latency exceeds 20 ms. Looking at the HTTP performance in Figure 20b, one may wonder why the ECDF stops below one such that 95% of users have a latency less than 190 ms. The reason is that the rest of users have an infinite latency due to the finite simulation duration. HTTP is a best-effort interactive model; therefore, it does not have stringent latency requirements. It can be seen that QAS approaches higher latency values compared to the other strategies for HTTP users. This is due to giving more priority to RT applications as explained below.

Inspecting the zoomed-in latency in Figures 21a and 21c, we see that VoIP users have the least delay values followed by video streaming, vehicular and HTTP users consecutively. Regarding the RR scheduler, this is due to the high packet arrival rate and the large packet size of the video model compared to VoIP. Also, vehicular users generate packets infrequently. With respect to the Best CQI scheduler, vehicular users have low reliability, as stated earlier; therefore, they are barely served and their packets experience higher latency compared to video streaming



(a) Round Robin

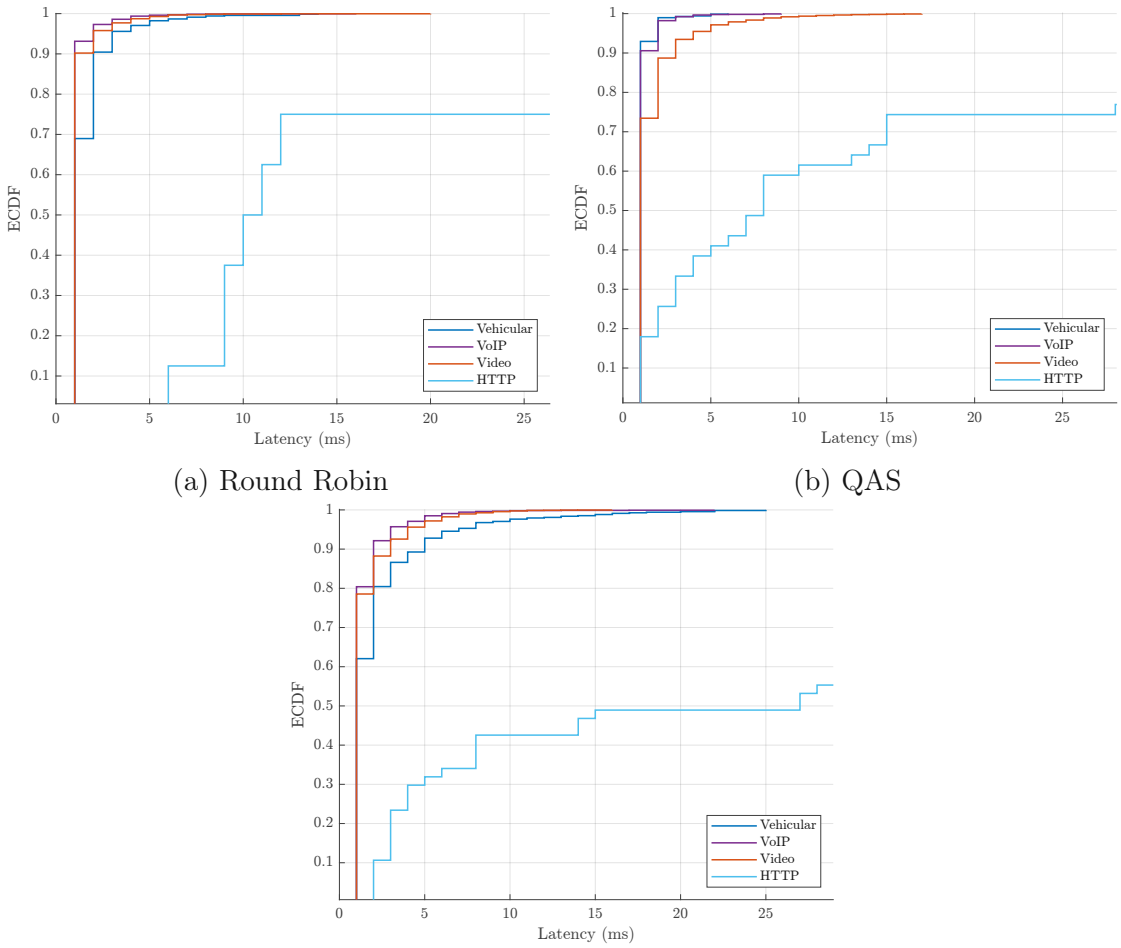
(b) QAS

(c) Best CQI

Figure 20: Latency empirical cumulative distribution function (ECDF) per traffic category and scheduling strategy for 140 users.

and VoIP users. This order of the latency curves is not what we wish for, since the curves must act according to the increase of delay constraints of the RT models followed by the NRT models. This is the case when using QAS, the latency curves in Figure 21b are in ascending order starting with vehicular and followed by VoIP, video streaming, and HTTP consecutively. One more thing to wonder about is the stairs shape of the ECDF of each traffic model. This non-smooth behavior can be observed for each scheduling strategy. It is a matter of the small number of users considered in this simulation.

Roughly speaking, as the number of users increase to 210, the latency performance degrades, i.e, higher latency values can be seen on the x-axis as shown in Figure 22. It can be seen that none of the strategies violate the delay constraint of video



(a) Round Robin

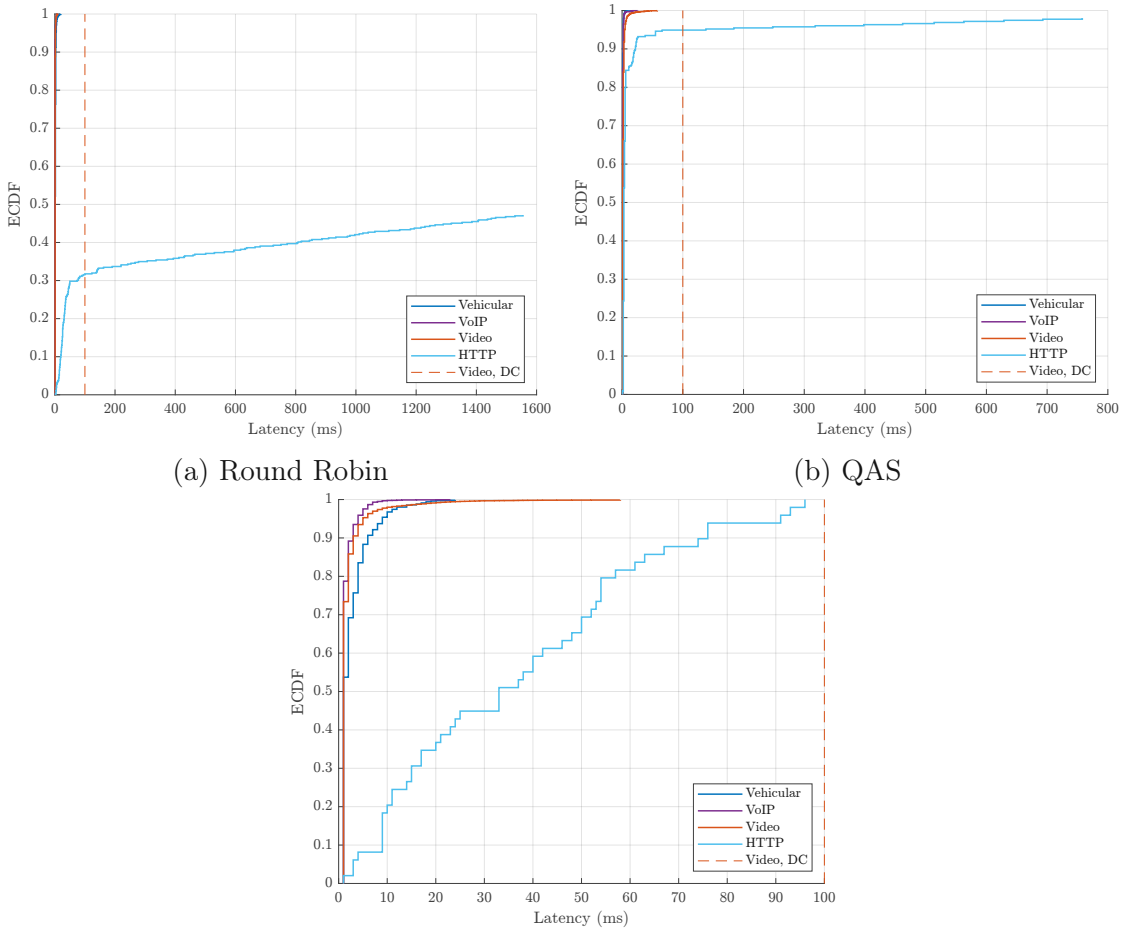
(b) QAS

(c) Best CQI

Figure 21: Zoomed-in latency ECDF per traffic category and scheduling strategy for 140 users.

streaming model. Regarding HTTP, high latency values are recorded using the RR scheduler, since 47% of the users have a delay less than 1555 ms, while others have an infinite delay. On the other hand, 98% of the users have a delay less than 758 ms using QAS, which means that fewer users experience infinite delays compared to the RR scheduler. The first reason of this is that, on average, higher number of HTTP users are active using the RR scheduler as shown earlier in Figure 19. The second is the priority given to these users using QAS due to their large packet size. The Best CQI scheduler grants low latency values for HTTP users; however, latency is not a strict performance indicator for this traffic model.

Looking at the zoomed-in latency in Figure 23, one can see that the RR and the



(a) Round Robin

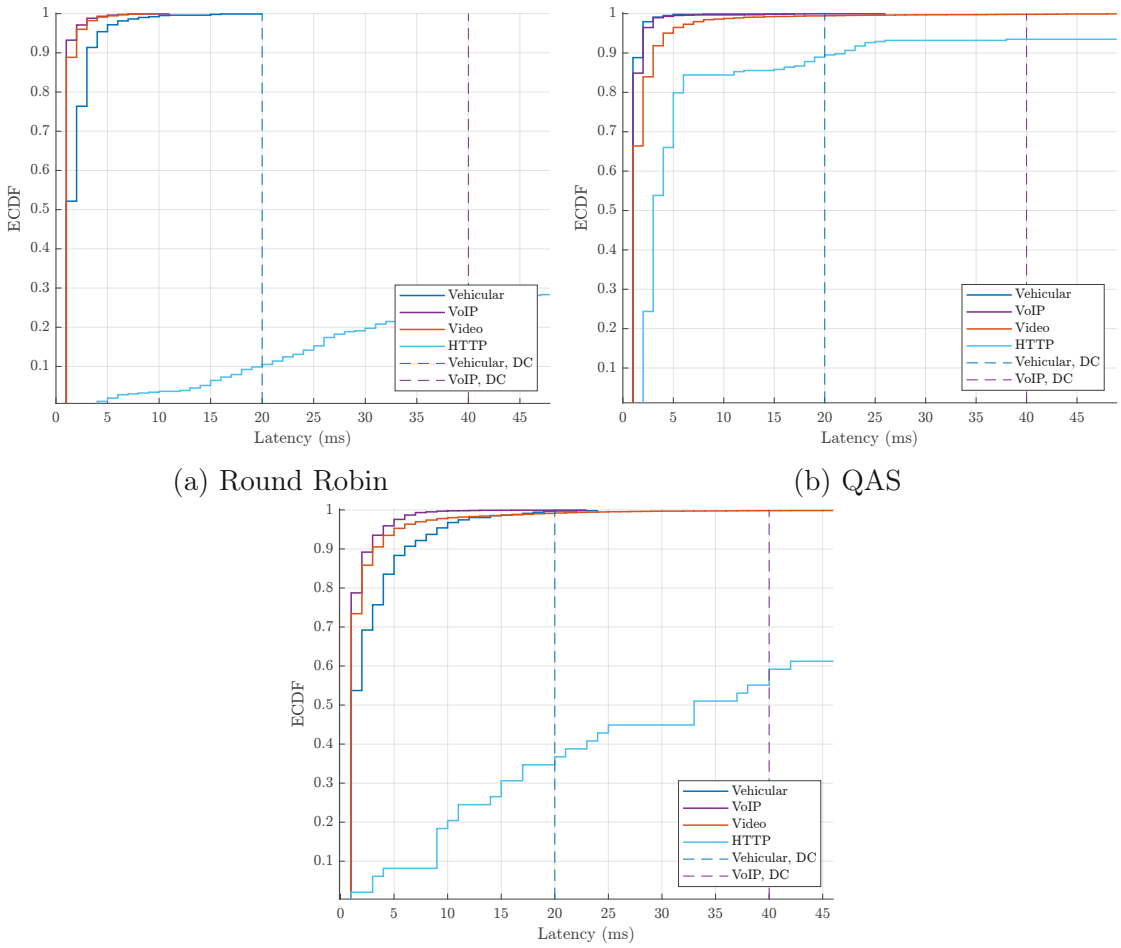
(b) QAS

(c) Best CQI

Figure 22: Latency ECDF per traffic category and scheduling strategy for 210 users.

Best CQI schedulers perform worst, since the order of the latency curves does not follow the increase in the delay constraints of the RT models. In other words, low delay values are assigned to VoIP users followed by video and vehicular users consecutively. When using QAS, we see in Figure 23b that low delay values are assigned to vehicular users, followed by VoIP and video users respectively. For example, considering a delay less than 4 ms, 99% of vehicular users have such a delay using QAS while 91% and 76% of them have the same delay using the RR and the Best CQI schedulers consecutively. Also, the Best CQI scheduler violates the delay constraint of vehicular communications for 0.3% of the users. None of the scheduling strategies violates the delay constraint of VoIP.

Finally, we consider the simulation of 420 users shown in Figure 24. There is a



(a) Round Robin

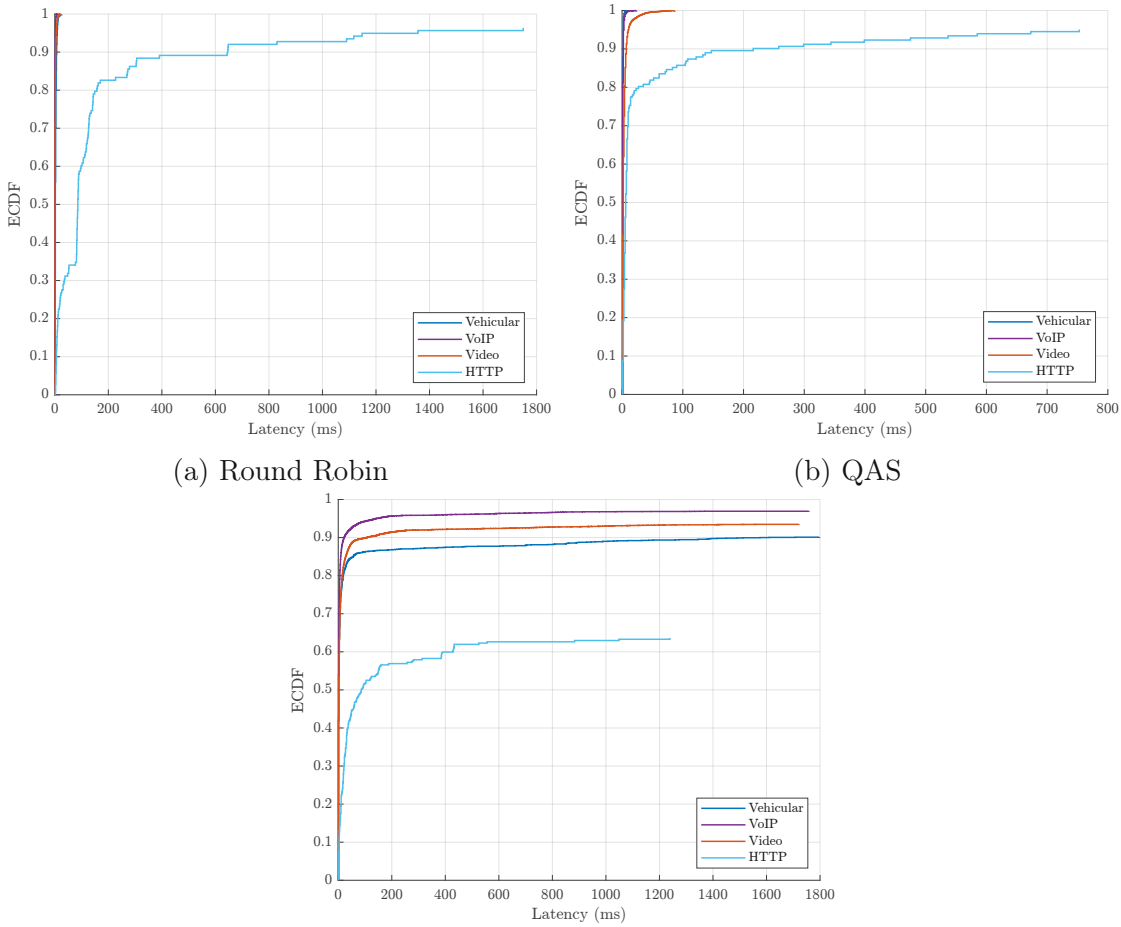
(b) QAS

(c) Best CQI

Figure 23: Zoomed-in latency ECDF per traffic category and scheduling strategy for 210 users.

noticeable rapid increase of latency values assigned to HTTP users using the RR and the Best CQI schedulers. This is due to the high number of HTTP users that are active per TS in this simulation. On average, less number of HTTP users are active while using QAS; hence, the incident is not observable in QAS performance. Also, HTTP users have a low reliability using the Best CQI scheduler as shown in Figure 25; therefore, their packets encounter high delays. The Best CQI scheduler performs the worst with 64% of the users having a delay less than 1236 ms, which translates to an abundant percentage of users experiencing an infinite delay.

Inspecting the zoomed-in latency in Figure 26, one can observe that the RR scheduler violates the delay constraint of vehicular communications for 0.2% of the users.



(a) Round Robin

(b) QAS

(c) Best CQI

Figure 24: Latency ECDF per traffic category and scheduling strategy for 420 users.

As stated earlier, the number of active HTTP users per TS increased during this simulation; therefore, one can notice the bend around 80 ms since the packets of the users stay in the buffer for quite a while until their next scheduling instance and encounter even more delay. Knowing that the delay constraint of all RT applications, Figure 26c shows that the Best CQI scheduler violates the delay constraint of all of them which is not the case for QAS as shown in Figure 26b. The last observation is that the latency curves of RT applications adhere to the desired order discussed earlier, that is vehicular, VoIP, video streaming, and HTTP consecutively, while using QAS only.

Table 5 summarizes the latency performance of RT traffic models by listing the ones that their delay constraints are exceeded using any of the aforementioned

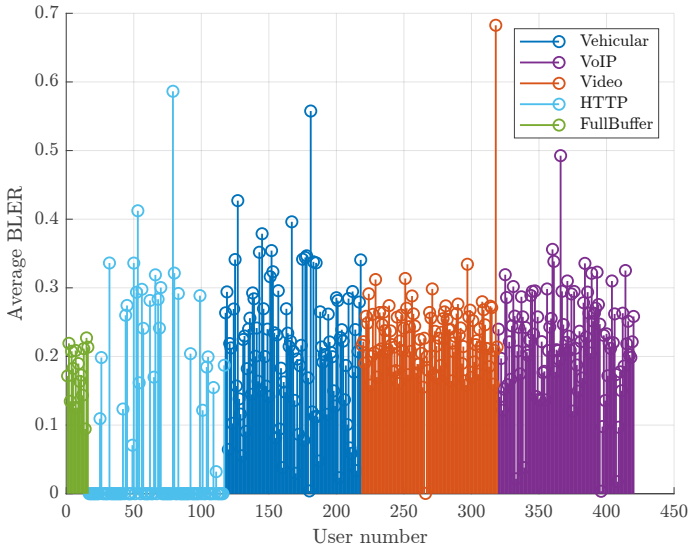
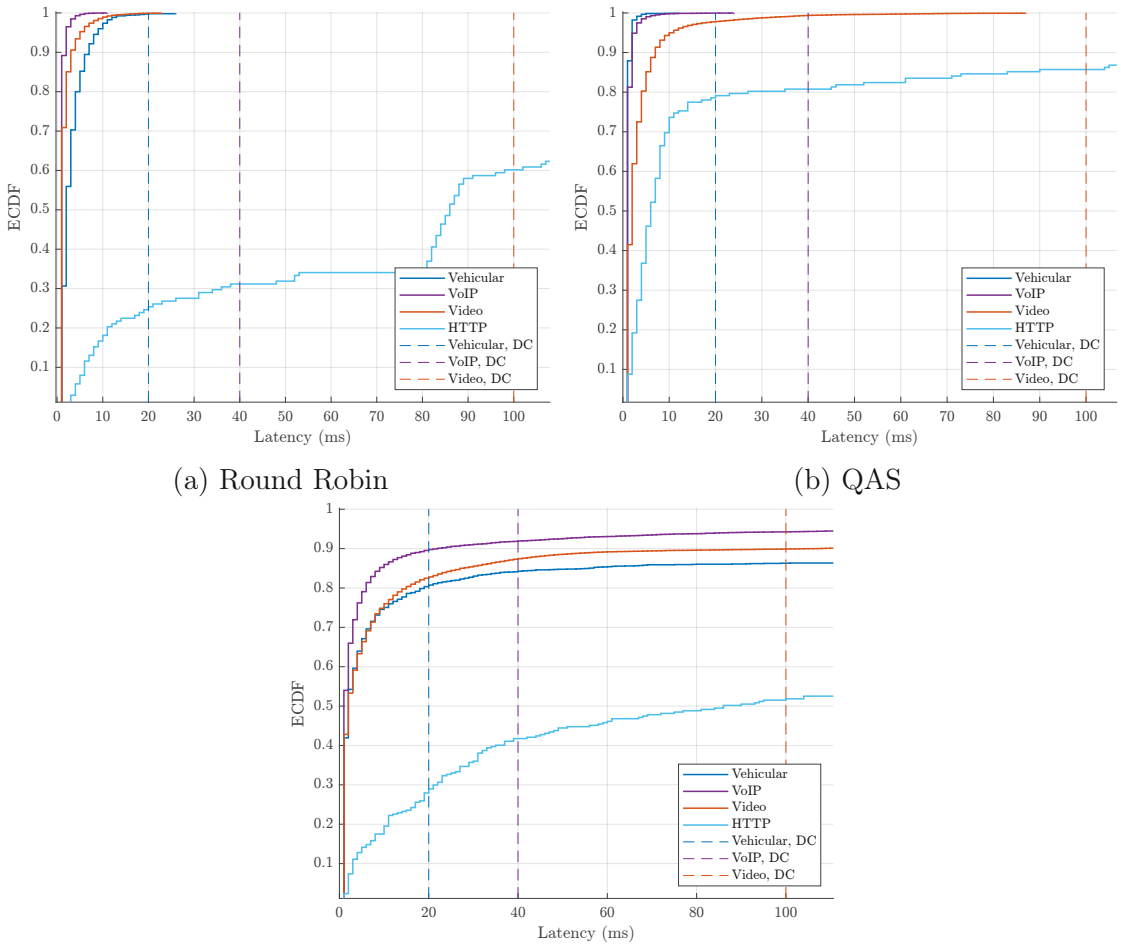


Figure 25: Average BLER per traffic category for 420 users using the Best CQI scheduler.

Table 5: The RT traffic models that their delay constraints are violated using each of the scheduling strategies.

Scheduling strategy	140 UEs	210 UEs	420 UEs
RR	✓	✓	Vehicular
QAS	✓	✓	✓
Best CQI	Vehicular	Vehicular	All models

scheduling strategies. The check mark means that none of the delay constraints is violated.



(a) Round Robin

(b) QAS

(c) Best CQI

Figure 26: Zoomed-in latency ECDF per traffic category and scheduling strategy for 420 users.

7.2 Indoor-Outdoor Simulation Scenario

This section elaborates the second scenario used for the performance investigation. Table 6 reports the main simulation parameters considered. In this scenario, we study the traffic mix of three applications, namely vehicular, VoIP and HTTP users. These applications adhere to the same constraints as elaborated in the previous section. The simulator allows for simulating buildings and streets arranged according to real world cities based on data gathered from the OpenStreetMap database. For this scenario, the location of Vienna University of Technology (TU Wien) Gusshausstraße Campus is chosen. Thus, HTTP and VoIP users are static and centered in the Campus experiencing wall loss inside the building and some vehicular users, moving at a speed of 30 km/h, are distributed on the streets around the building. Two BSs are placed in predefined positions in front of TU Wien Campus as shown in Figure 27.

7.2.1 Simulation Results

The performance of the schedulers is studied under two network loads of 50 and 76 users, respectively. The percentage of users belonging to each category in the first simulation is: 40% vehicular, 30% VoIP and 30% HTTP. The same percentages are considered for the second set.

Throughput

Similar to the outdoor scenario, we start by presenting the throughput performance and comparing our proposed scheduler against the other scheduling strategies. Figure 28 shows the performance measures of interest which are average and sum throughput. They are calculated according to (32) and (33). It is worth bearing in mind that, similar to the previous scenario, a single random realization of users' positions is considered.

Examining the first simulation at the top of the figure, it can be seen that the Best CQI scheduler allocates low average and sum throughput for vehicular users as they encounter a high average BLER as shown in Figure 29c and some of them experience an outage having an average BLER equal to one. Although HTTP users have a low average BLER, the lowest average and sum throughput is transmitted to them due to their low user activity per TS unlike VoIP users as shown in Figure 30. In the RR scheduler case, throughput is distributed fairly among users which induces less differences between the sum throughputs as well as the average ones among the traffic categories compared to other strategies; however, this fairness comes at the cost of drop in the average and sum throughput. Using QAS, HTTP users get the highest sum and average throughput compared to classic scheduling

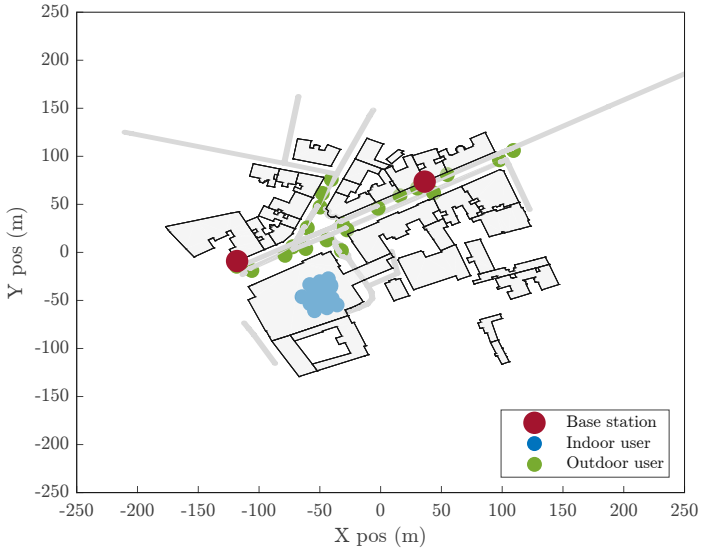


Figure 27: Network deployment of the indoor-outdoor scenario.

Table 6: Main simulation parameters of the indoor-outdoor scenario

Transmission parameters	
TS duration	1 ms
Center frequency	2 GHz
Bandwidth	20 MHz
Simulation duration	4000 TS
Network elements properties	
Number of users	50, 76 users
Number of BSs	2 BSs in predefined positions
BS antenna	1 omnidirectional antenna
User antenna	1 omnidirectional antenna
Wall loss	20 dB
Traffic models	vehicular, VoIP, HTTP
Link properties	
Channel model	Typical urban
Path loss model	Urban macro

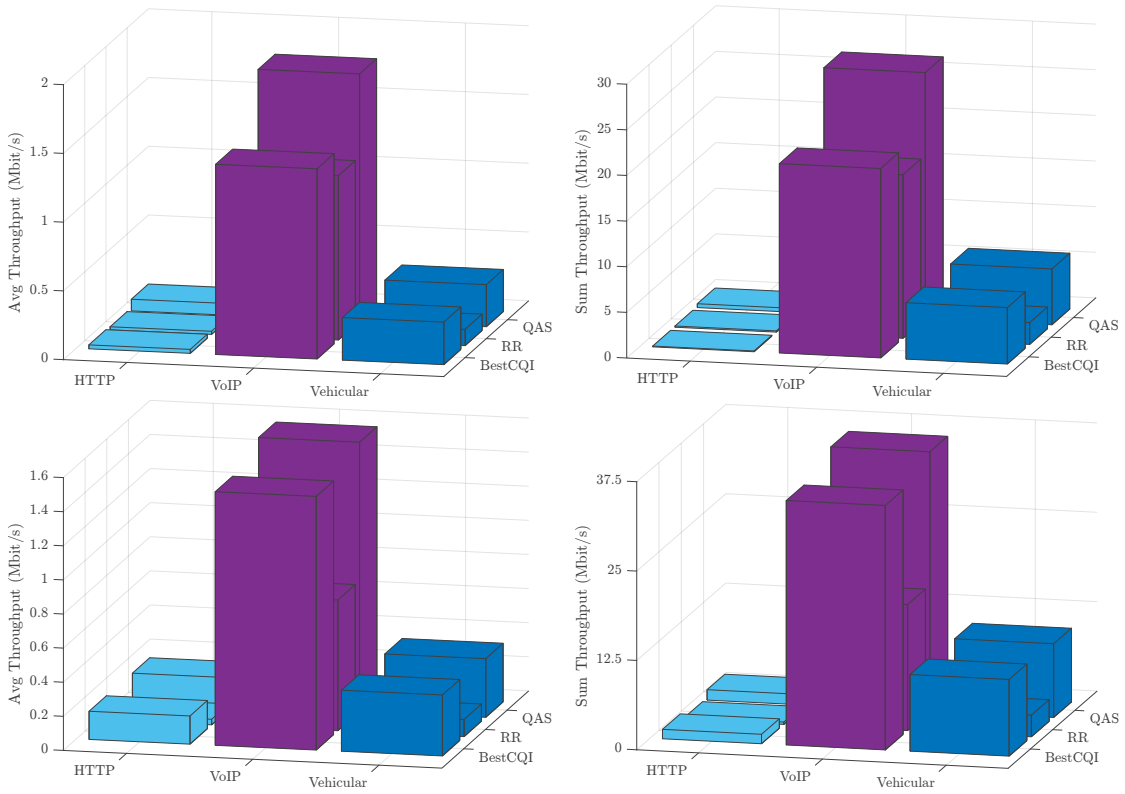
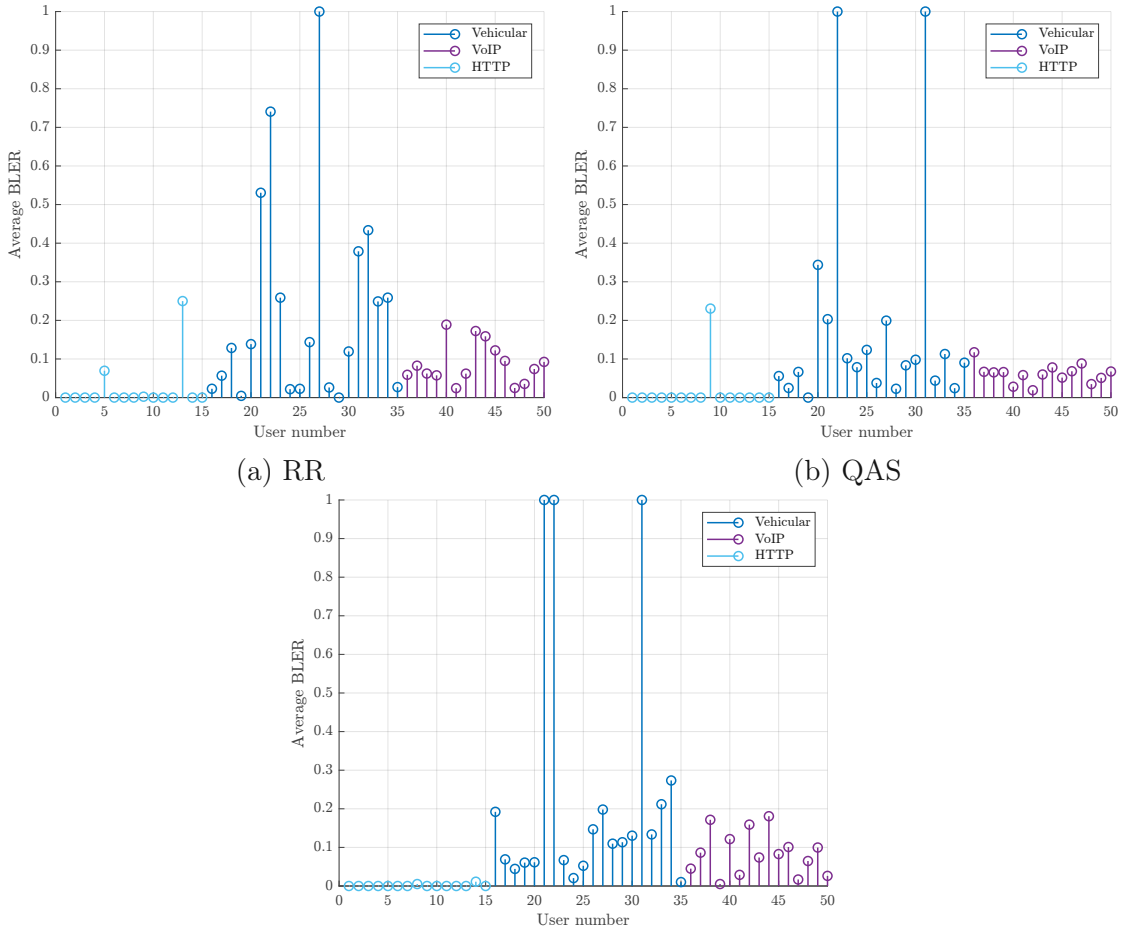


Figure 28: Average and sum throughput per traffic model and scheduling strategy for 50 (top) and 76 (bottom) users.

strategies, as they reliably produce large packets. Vehicular users still acquire a high average and sum throughput compared to other schedulers, despite the fact that they are highly unreliable as shown in Figure 29b. Their average throughput is a bit less than what is granted by the Best CQI scheduler. This is due to offering priority to all vehicular users and not transmitting to those with the best channel quality only, as is the case with the Best CQI scheduler. Finally, due to the priority given to VoIP as a RT application with strict delay constraint, they get the highest sum and average throughput compared to other schedulers.

In the following, we consider the results of the second simulation of 76 users at the bottom of Figure 28. In principle, the Best CQI scheduler achieves high throughputs for users with high reliability but users with low reliability are barely severed according to Figure 31c. As a matter of course, the RR scheduler performs worst as it does not take into account the channel conditions for resource allocation. The average throughput assigned by our proposed scheduler for HTTP users is 2% less than the throughput provided by the Best CQI scheduler since QAS serves these



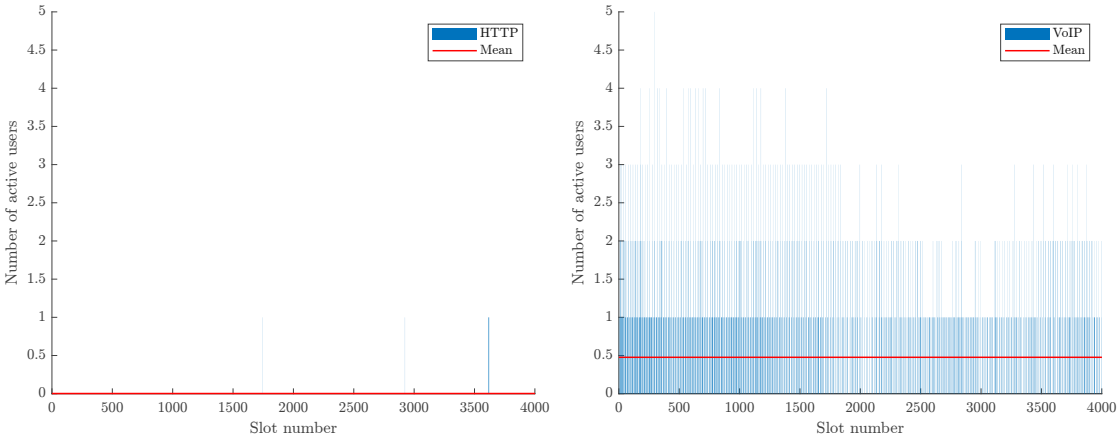
(a) RR

(b) QAS

(c) Best CQI

Figure 29: Average BLER per traffic category and scheduling strategy for 50 users.

users with throughput that matches the remaining bits in their buffers, whereas the Best CQI scheduler serves them continuously with high throughput due to their good channel conditions. For VoIP, QAS outperforms other strategies. In the case of vehicular communications, it performs well compared to the RR scheduler; however, the average throughput assigned by the Best CQI scheduler for this category is higher by 4% than what is granted by QAS for the same reason explained above for the simulation of 50 users. Table 7 presents the throughput upgrade or occasional downgrade that RT and NRT applications obtained using QAS.

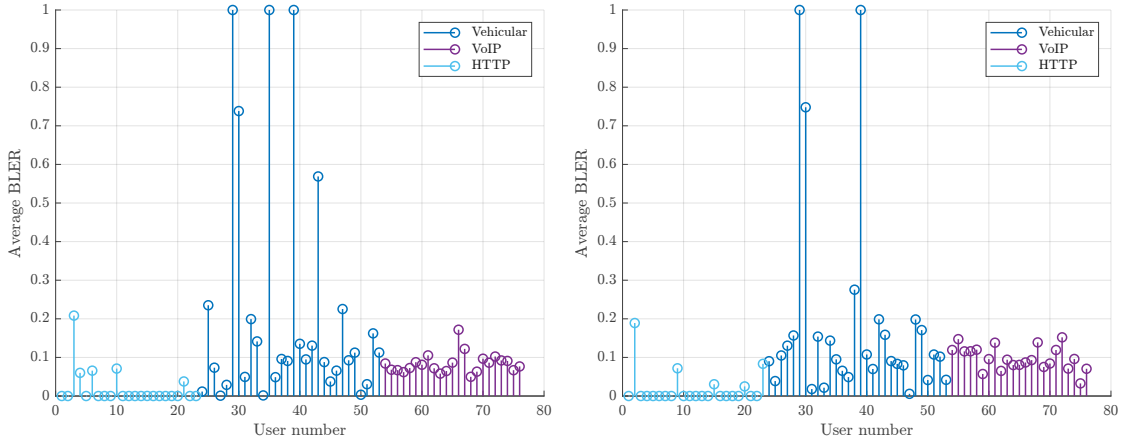


(a) HTTP (b) VoIP

Figure 30: Number of active HTTP and VoIP users per TS using the Best CQI scheduler.

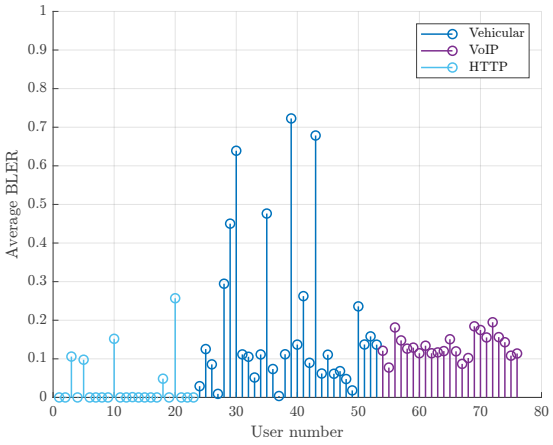
Table 7: Improvement factors of average throughput attained by introducing QAS.

Traffic model	50 UEs		76 UEs	
	RR	Best CQI	RR	Best CQI
Vehicular	2.57	0.99	3.42	0.96
VoIP	1.51	1.30	2.06	1.06
HTTP	3.77	2.96	4.97	0.98



(a) RR

(b) QAS



(c) Best CQI

Figure 31: Average BLER per traffic category and scheduling strategy for 76 users.

Table 8: The mean BLER calculated among all users per scheduling strategy.

Scheduling strategy	50 UEs	76 UEs
RR	0.15	0.14
QAS	0.12	0.13
Best CQI	0.16	0.15

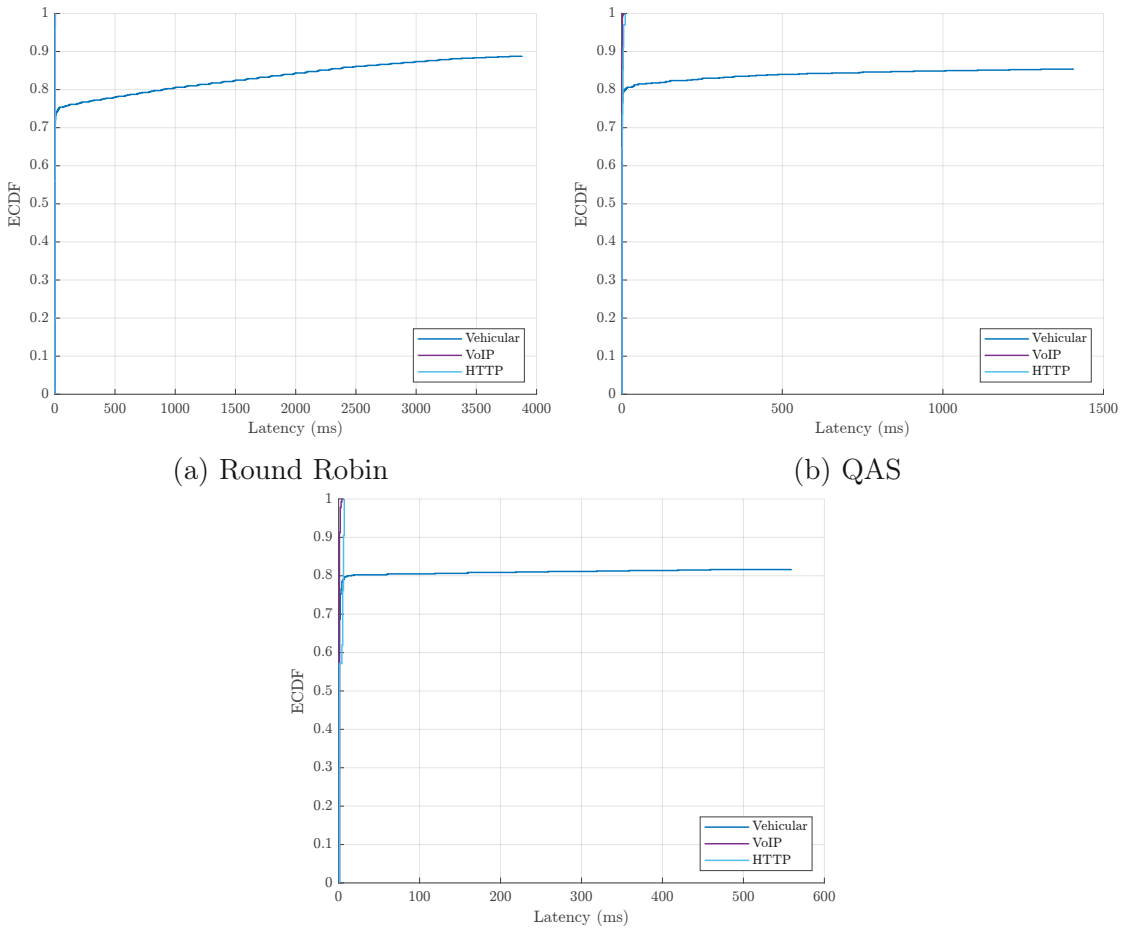
Reliability

Next, we examine the mean BLER per scheduling strategy for 50 and 76 users connected to BSs. Results in Table 8 show a slightly good trend of QAS on reliability among all users.

Latency

As stated earlier, there are few vehicular users in outage in this scenario, this results in an infinite delay for their packets as shown in Figure 32 for the three schedulers. Figure 32a shows that 11% of vehicular users have an infinite packet delay. Looking at QAS performance in Figure 32b, one can see that 86% of users experience a delay less than 1406 ms which leads to 14% with an infinite delay. Checking the performance of the RR scheduler for delays below 1406 ms, we find that 82% of the scheduled users experience a delay within the latter. Thus, QAS outperforms it with a higher percentage of users experiencing the same delays even though more users are in outage while using QAS as can be seen in Figure 29. Obviously, even a higher percentage of vehicular users encounter an infinite delay using the Best CQI scheduler which is 18%.

Inspecting the zoomed-in latency in Figure 33 shows that 74% of users scheduled by the RR scheduler attain the delay constraint of vehicular communications, while 81% of users scheduled by QAS do the same, which proves the validity of our proposed scheduler. Looking at Figure 33c, we see that 80% of the users attain the delay constraint of vehicular communications; therefore, one may think the Best CQI scheduler outperforms other strategies but this is not true. It allocates users with best channel conditions frequently; thus their packets do not experience high delays while QAS provide a priority for all vehicular users due to their strict delay constraint. None of the scheduling strategies violates the delay constraint of VoIP communications and all of them grant a low delay values for the HTTP model. Regarding the Best CQI scheduler, this is due to the high reliability of users belong to both categories. While the RR scheduler schedule users fairly which let their packets do not remain in the buffers for considerable amount of time. Finally, QAS offers a priority for VoIP users due to the strict delay constraint, for HTTP users



(a) Round Robin

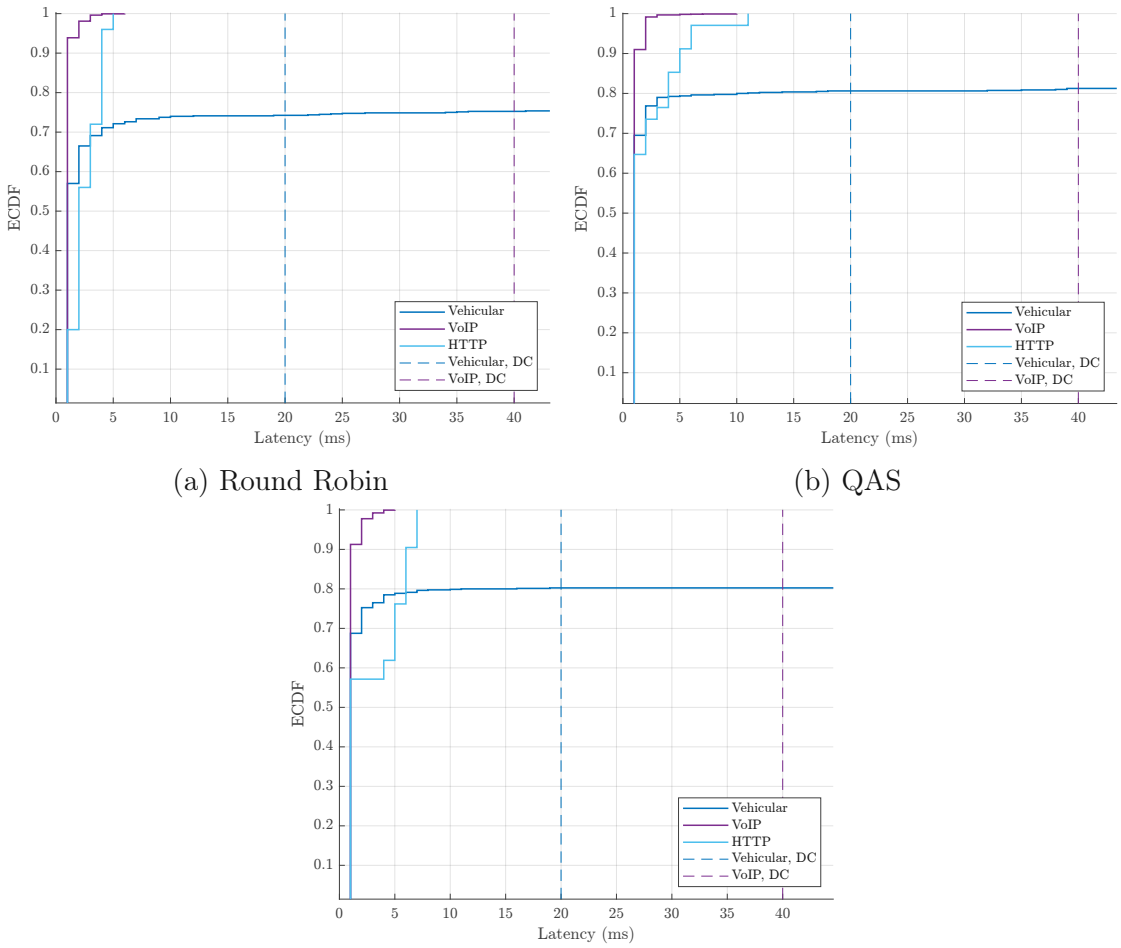
(b) QAS

(c) Best CQI

Figure 32: Latency ECDF per traffic category and scheduling strategy for 50 users.

due to the relatively large packets, and for both due to the high reliability.

When considering the increased network load as shown in Figure 34, one can observe that 87% of the vehicular users encounter a delay less than 3944 ms when scheduled by the RR scheduler, 94% of them experience a delay less than 3708 ms when allocated by the Best CQI scheduler, and 93% of them face a delay less than 3550 ms when assigned by QAS. This means 13%, 6% and 7% of users experience an infinite delay when using the RR scheduler, the Best CQI scheduler and QAS, respectively. One may think that Best CQI outperforms; however, no users experience an outage using the latter scheduler, unlike the RR scheduler and QAS as shown in Figure 31. Users suffer an outage have an average BLER of one. Also, the range of delays assigned by QAS to the packets of vehicular users is smaller than the one offered by the Best CQI scheduler.



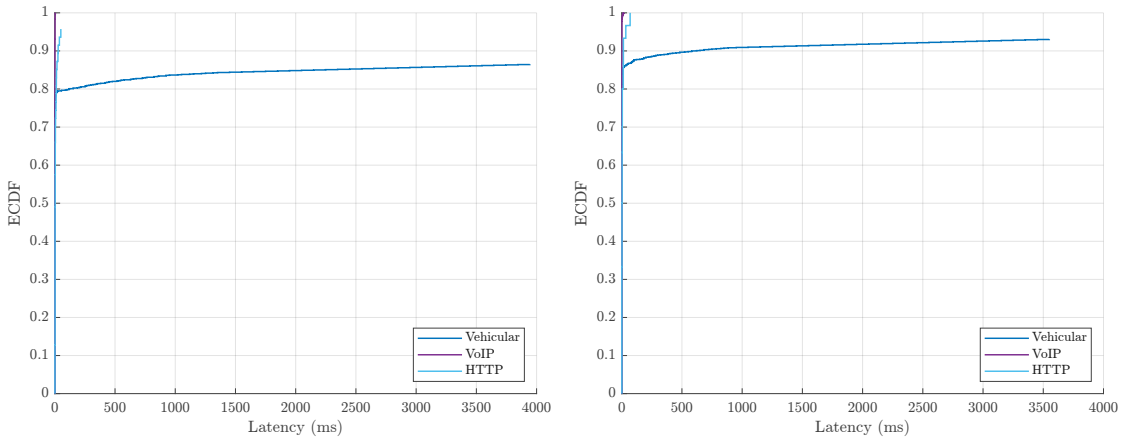
(a) Round Robin

(b) QAS

(c) Best CQI

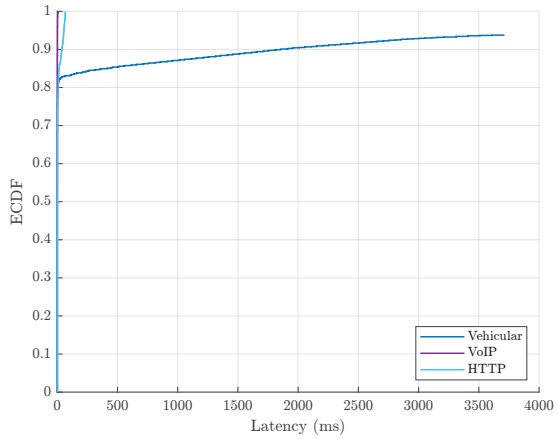
Figure 33: Zoomed-in latency ECDF per traffic category and scheduling strategy for 50 users.

Similar to the previous simulation, none of the scheduling strategies violates the delay constraint of VoIP communications as shown in Figure 35. Looking at the vehicular users performance, it can be seen that, 79% and 82% of the users attain the delay constraint using the RR and the Best CQI schedulers respectively. Whereas 86% of the users do the same using QAS which shows that our proposed scheduler performs better than other schedulers. Regarding the NRT model, the RR scheduler grants a delay less than 50 ms for 96% of HTTP users, hence the rest experience an infinite delay. The latter incident does not occur while using QAS or the Best CQI scheduler.



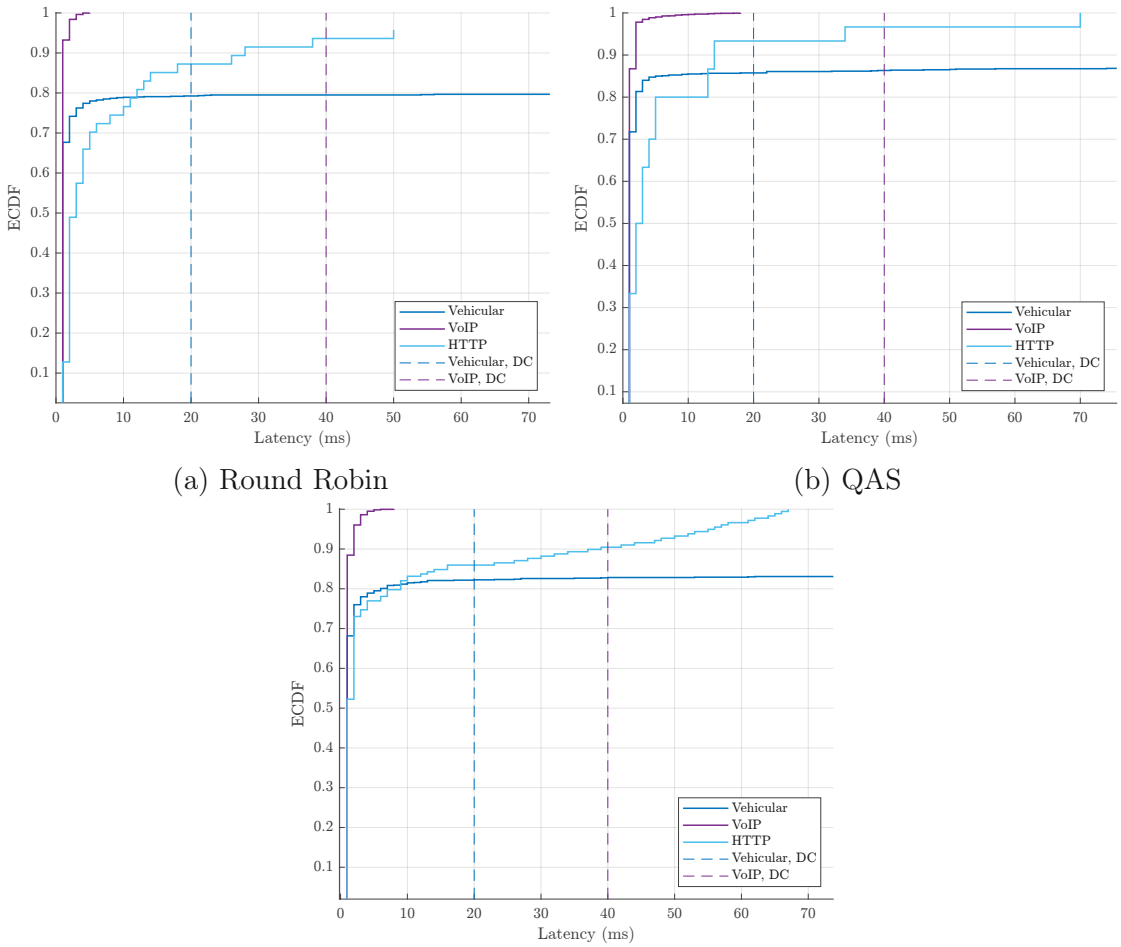
(a) Round Robin

(b) QAS



(c) Best CQI

Figure 34: Latency ECDF per traffic category and scheduling strategy for 76 users.



(a) Round Robin

(b) QAS

(c) Best CQI

Figure 35: Zoomed-in latency ECDF per traffic category and scheduling strategy for 76 users.

7.3 Discussion

The results aggregated in Sections 7.1 and 7.2 for two scenarios involving distinct traffic types prove that the formulation of the RBs allocation problem of QAS is valid since it achieves the desired aims. Values chosen for the priority bases ($\sigma = 1.02, \alpha = 2$) deliver a reasonable performance to handle mixed traffic scenarios. However, these values were chosen based on the theoretical assumptions in Section 6.2.4. The investigation of the optimal priority bases of this scheduler is left for future work. Comparing the performance of QAS to other classic scheduling strategies namely RR and Best CQI under different network loads shows that the promised gains in throughput, reliability, and latency are most likely feasible in a realistic environment.

Throughput improvement is achieved for all RT and NRT applications, since QAS grants all users as much resources as needed. This does not include full buffer users that are served with the remaining RBs. The gain achieved increases for highly loaded cells in some network configurations; however, the computational complexity also increases, which is a hindrance to extending the performance investigation. When considering the latency performance, QAS outperforms the two benchmark scheduling strategies in the RT applications. It increases the transmission rate for users based on their priority factor, leading to a smaller packet delay and, in turn, improved average throughput since more resources are allocated to RT users. Therefore, this should not be understood as small packet delay and improved throughput are generally related. However, it is not permissive in ensuring low delays for NRT services. Furthermore, prioritizing RT users with low reliability degrades the reliability performance in some scenarios. Due to insufficient random realizations of each traffic model, the randomness introduced by the statistics of the traffic models influences the number of active users during each simulation; hence occasionally biases the results. Considering lengthy simulations and repeating the same simulation sufficiently often would eliminate this effect; hence lead to an accurate performance. QAS is a practical solution for enhancing the reliability in wireless networks as it achieves an intermediate BLER performance between the other state of the art scheduling strategies and this is comprehensible because the RR scheduler allocates resources for all users and then the total reliability averages among them and the Best CQI scheduler assigns users with high channel conditions; thus their transmissions most likely would fail and the network reliability deteriorate consequently.

For fairness, an average level is offered among full buffer users. Higher levels have not been investigated in this work due to feasibility of the RBs optimization problem. Moreover, imposing fairness among other traffic categories is not in our interest as we aim for taking advantage of the QoS characteristic of each traffic. As

the impact of optimizing the QoS delivery to users on computational complexity is a key performance indicator, the processing time spent using each of the aforementioned scheduling strategies is measured. Roughly speaking, QAS spends $1.44 - 2.29$ more time than the classic schedulers. This computational effort is a severe limitation for handling the RBs optimization problem for high number of users. This result can not be generalized as it is obtained using our mixed traffic scenarios only. In this work, precise measurements for the computational complexity using each of the traffic models individually have not been examined. However, two methods are implemented to reduce the simulation time. The first is using the solver Mosek, which belongs to the solvers family of CVX [20] instead of Gurobi, for solving the optimization problem. The second is applying relaxation to the binary integer optimization problem as elaborated in Section 6.2.2. None of the aforesaid methods reduces the computational complexity considerably. These findings encourage further endeavors to possibly implement faster schemes of QAS.

Some questions are left open after this work, opening the door for extensive analysis and potential improvements. Two network designs were used for the performance evaluation of QAS. First, an outdoor scenario with a hexagonal grid of BSs and users outdoors. Second, an indoor-outdoor scenario with two BSs placed in predefined positions and some users experience penetration loss inside TU Wien Gusshausstraße Campus while others are outside. All of these BSs have omnidirectional antennas. Real wireless networks would involve directive antennas which should be investigated with QAS. The network setup only considers urban environments. Taking into account other channel models that reflect the practical networks would be favorable to reveal the effectiveness of QAS. To get better insights, various models for microscopic fading should be taken into consideration as well. Network interference was not adequately represented for cell-edge users since no interference region has been been considered in the SL simulations. Extending simulation areas to mitigate border effects may yield more accurate outcomes in terms of reliability and other QoS attributes. All users utilized in the networks framework are static except the vehicular. Introducing different types of user mobility can yield interesting results because this has an impact on the channel quality of the users.

8 Conclusion and Outlook

8.1 Conclusion

By leveraging recent results on scheduling, in this thesis, we formulated the weighted sum throughput maximization problem according to some QoS constraints in the framework of 5G and developed a practical downlink scheduler. Performance investigation of QAS against popular existing scheduling strategies in small- and large-scale multi-traffic scenarios was presented. Based on the results acquired, it can be concluded that QAS boosts the system throughput even under high network load and performs exceptionally well in terms of packet latency. Additionally, it slightly increases reliability compared to the common scheduling strategies in the literature. However, this improvement comes at the cost of computational complexity as the number of connections increase.

8.2 Outlook

The scope of this work is limited to SISO transmissions. An extension that looks appealing is developing QAS to involve MIMO as one of the leading transmission techniques of 5G which would increase the network capacity. This means to extend the RBs optimization problem for maximizing over the sum throughput of multiple user streams. The drawback is that the burden on the hardware would increase. Furthermore, the assumption of equal power allocation on the RB may not be handy in practical wireless networks. Therefore, the influence of an additional total transmit power constraint should be analyzed. The traffic mixes examined for the performance evaluation of QAS should address Internet of Things (IoT) scenarios considering massive connection densities with loose latency and data rates requirements. Another approach could be to combine QAS with HARQ. On the one hand, this might further increase the throughput and reliability of the network. On the other hand, the reliability enhancement mechanism of QAS would interfere with HARQ.

9 References

[1] R. Jain, D. M. Chiu, and H. WR, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," *Eastern Research Laboratory Digital Equipment Corporation*, Dec. 1984.

[2] W. J. Cook, W. H. Cunningham, W. R. Pulleyblank, and A. Schrijver, *Combinatorial Optimization*. John Wiley & Sons, Inc., 1998, ISBN: 047155894X.

[3] H. Chaskar and U. Madhow, "Fair scheduling with tunable latency: A round robin approach," in *Seamless Interconnection for Universal Services. Global Telecommunications Conference. GLOBECOM'99. (Cat. No.99CH37042)*, vol. 2, 1999, 1328–1333 vol.2. DOI: 10.1109/GLOCOM.1999.829988.

[4] 3GPP, "High Speed Downlink Packet Access: UE Radio Transmission and Reception," 3rd Generation Partnership Project (3GPP), Tech. Rep., 2002.

[5] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511804441.

[6] 3GPP, "Long Term Evolution (LTE) physical layer framework for performance verification," 3rd Generation Partnership Project (3GPP), TSG RAN1-070674, Feb. 2007.

[7] B. Sadiq, R. Madan, and A. Sampath, "Downlink scheduling for multiclass traffic in LTE," *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, Dec. 2009. DOI: 10.1155/2009/510617.

[8] 3GPP, "Service requirements for V2X services," 3rd Generation Partnership Project (3GPP), TS 22.185, Feb. 2010.

[9] 3GPP, "Universal Mobile Telecommunications System (UMTS) Deployment aspects," 3rd Generation Partnership Project (3GPP), TR 25.943, Feb. 2010.

[10] G. Riley and T. Henderson, "The ns-3 network simulator," in Jan. 2010, pp. 15–34, ISBN: 978-3-642-12330-6. DOI: 10.1007/978-3-642-12331-3_2.

[11] S. Schwarz, C. Mehlführer, and M. Rupp, "Calculation of the spatial preprocessing and link adaption feedback for 3GPP UMTS/LTE," in *2010 Wireless Advanced 2010*, IEEE, Jun. 2010. DOI: 10.1109/wiad.2010.5544947.

[12] S. Schwarz, C. Mehlführer, and M. Rupp, "Low complexity approximate maximum throughput scheduling for LTE," in *2010 Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers*, 2010, pp. 1563–1569. DOI: 10.1109/ACSSC.2010.5757800.

[13] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*, 1st. USA: Academic Press, Inc., 2011, ISBN: 012385489X.

[14] A. F. Molisch, *Wireless Communications*, 2nd. Wiley Publishing, 2011, ISBN: 0470741864.

[15] M. Rupp, S. Caban, C. Mehlführer, and M. Wrulich, "Evaluation of HSDPA and LTE: From testbed measurements to system level performance," 2011.

[16] S. Schwarz, C. Mehlführer, and M. Rupp, "Throughput maximizing multiuser scheduling with adjustable fairness," in *2011 IEEE International Conference on Communications (ICC)*, 2011, pp. 1–5. DOI: 10.1109/icc.2011.5963489.

[17] J. C. Ikuno, "System Level Modeling and Optimization of the LTE Downlink," Ph.D. dissertation, TU Wien, 2013.

[18] M. K. Müller, S. Schwarz, and M. Rupp, "QoS investigation of proportional fair scheduling in LTE networks," in *2013 IFIP Wireless Days (WD)*, 2013, pp. 1–4. DOI: 10.1109/WD.2013.6686478.

[19] 3GPP, "Study on 3D channel model for (LTE)," 3rd Generation Partnership Project (3GPP), TR 36.873, Sep. 2014.

[20] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming, version 2.1*, http://cvxr.com/cvx, Mar. 2014.

[21] S. Jaeckel, L. Raschkowski, K. Boerner, and L. Thiele, "QuaDRiGa: A 3D Multi-Cell Channel Model With Time Evolution For Enabling Virtual Field Trials," *IEEE Transactions On Antennas and Propagation*, vol. 62, pp. 3242–3256, 2014.

[22] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); further advancements for E-UTRA physical layer aspects," 3rd Generation Partnership Project (3GPP), TR 36.814, Mar. 2017.

[23] H. R. Chayon, K. B. Dimyati, H. Ramiah, and A. W. Reza, "Enhanced quality of service of cell-edge user by extending modified largest weighted delay first algorithm in LTE networks," *Symmetry*, vol. 9, no. 6, 2017, ISSN: 2073-8994. DOI: 10.3390/sym9060081.

[24] T. Dittrich, M. Rupp, and M. Taranetz, "An Efficient Method for Avoiding Shadow Fading Maps in System Level Simulations," in *WSA 2017; 21st International ITG Workshop on Smart Antennas*, Mar. 2017, pp. 1–8.

[25] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*, 1st. USA: Academic Press, Inc., 2018, ISBN: 0128143231.

[26] C.-K. Jao, C.-Y. Wang, T.-Y. Yeh, *et al.*, "WiSE: A system-level simulator for 5G mobile networks," *IEEE Wireless Communications*, vol. 25, no. 2, pp. 4–7, 2018. DOI: 10.1109/MWC.2018.8352614.

[27] N. Madi, M. H. Zurina, M. Othman, and S. Subramaniam, "Delay-based and QoS-aware packet scheduling for RT and NRT multimedia services in LTE downlink systems," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, Jul. 2018. DOI: 10.1186/s13638-018-1185-3.

[28] M. K. Müller, F. Ademaj, T. Dittrich, *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, no. 1, p. 17, Sep. 2018. DOI: 10.1186/s13638-018-1238-7.

[29] S. Pratschner, B. Tahir, L. Marijanovic, *et al.*, "Versatile mobile communications simulation: The Vienna 5G Link Level Simulator," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, p. 226, Sep. 2018. DOI: 10.1186/s13638-018-1239-6.

[30] J. Xu, C. Guo, H. Zhang, and J. Yang, "Resource allocation for real-time traffic in unreliable wireless cellular networks," vol. 24, no. 5, pp. 1405–1418, Jul. 2018. DOI: 10.1007/s11276-016-1413-x.

[31] D. Wang, R. R. Sattiraju, A. Weinand, and H. D. Schotten, "System-level simulator of LTE sidelink C-V2X communication for 5G," 2019, pp. 1–5.

[32] 3GPP, "Quality of Service (QoS) concept and architecture," 3rd Generation Partnership Project (3GPP), TS 23.107, Jul. 2020.

[33] C. W. Paper, "Cisco visual networking index: Global mobile data traffic forecast update, 2017–2022," 2020.

[34] 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, 2020. DOI: 10.1109/MNET.001.1900287.

[35] J. Lee, M. Han, M. Rim, and C. G. Kang, "5G K-SimSys for Open/Modular/Flexible System-Level Simulation: Overview and its Application to Evaluation of 5G Massive MIMO," *IEEE Access*, vol. 9, pp. 94 017–94 032, 2021. DOI: 10.1109/ACCESS.2021.3093460.

[36] 3GPP, "Services and service capabilities," 3rd Generation Partnership Project (3GPP), TS 22.105, Mar. 2022.

[37] A. Abdulazeez, M. M. Yahaya, I. Bala 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, pp. 1–5. DOI: 10.1109/NIGERCON54645.2022.9803123.

[38] MATLAB, *Class-based unit tests*, https://www.mathworks.com/help/matlab/class-based-unit-tests.html, R2022a, 2022.

[39] *Omnet++*, https://omnetpp.org.

[40] *OpenStreetMap*, https://www.openstreetmap.org/, Accessed: 2021-04-19.

A Appendix

Table of Abbreviations

3GPP	3rd Generation Partnership Project
5G	5th generation
ACK	Acknowledged
AMR	Adaptive Multi-Rate
AWGN	Additive White Gaussian Noise
BLER	Block Error Ratio
BS	Base Station
CQI	Channel Quality Indicator
DCP	Disciplined Convex Programming
ECDF	empirical cumulative distribution function
FTP	File Transfer Protocol
GSM	Global System for Mobile communications
HARQ	Hybrid Automatic Repeat Request
HTTP	Hypertext Transfer Protocol
IoT	Internet of Things
LLS	Link Level Simulator
LOS	Line Of Sight
LPM	Link Performance Model
LQM	Link Quality Model
LTE	Long Term Evolution
MAC	Multiple Access Channel
MCS	Modulation and Coding Scheme
MLWDF	Modified Largest Weighted Delay First
MDP	Markov Decision Process
MIDCP	Mixed Integer Disciplined Convex Programming
MIESM	Mutual Information Effective Signal to Interference and Noise Ratio Mapping

MIMO Multiple-Input Multiple-Output
NLOS Non Line of Sight
NACK Non-Acknowledged
OFDM Orthogonal Frequency Division Multiplexing
OMA Orthogonal Multiple Access
PDP Power Delay Profile
PedA Pedestrian A
PF Proportional Fair
PMI Precoding Matrix Indicator
PPP Poisson Point Process
QAS Quality of Service Aware Scheduler
QoS Quality of Service
QuaDRiGa QUAsi Deterministic RadIo channel GenerAtor
RB Resource Block
RI Rank Indicator
RT Real Time
NRT Non Real Time
ROI Region Of Interest
RR Round Robin
SINR Signal to Interference and Noise Ratio
SISO Single-Input Single-Output
SU-MIMO Single User Multiple-Input Multiple-Output
SID Silence Insertion Descriptor
SL System Level
SLS System Level Simulator
SNR Signal to Noise Ratio
TS Time Slot
TU Wien Vienna University of Technology
UDP User Datagram Protocol
UE User Equipment

VCCS Vienna Cellular Communications Simulators

VehA Vehicular A

VoIP Voice over IP

WRR Weighted Round Robin

List of Figures

List of Tables