

Traffic Model for Cellular Network Analysis

MOHANAD AMER

SHISHIRA PUTTASWAMY

MASTER'S THESIS

DEPARTMENT OF ELECTRICAL AND INFORMATION TECHNOLOGY

FACULTY OF ENGINEERING | LTH | LUND UNIVERSITY





LUND UNIVERSITY



ERICSSON

Master Thesis

Traffic Model for Cellular Network Analysis

By

Mohanad Amer and Shishira Puttaswamy

Department of Electrical and Information Technology

Faculty of Engineering, LTH, Lund University

SE-221 00 Lund, Sweden

Supervisors: Emma Fitzgerald, LTH

Christer Östberg, Ericsson Lund

Henrik Ronkainen, Ericsson Lund

Examiner: Christian Nyberg

2019

Abstract

The development of new wireless cellular network technologies is always in progress. As 3G has been considered the foundation of mobile broadband, the latest generation of cellular mobile communications, that is, 5G New Radio, is expected to realize the networked society, where everyone and everything are seamlessly connected everywhere and every time.

To ensure the connectivity and provide the required services, the cellular network could first be simulated, and the performance evaluated. Simulations are made as a method to evaluate the performance of the connectivity and services. In addition, building new network simulators and efficient scheduler algorithms, could be crucial for dimensioning the networks to fulfil the new services/use cases demands.

This thesis work includes two main parts. First, a network model that reflects modern cellular architectures and user services, are defined and implemented. Second, a system level simulator that uses different scheduler algorithms is also developed. Several main factors and requirements that have been considered when building this simulator are, scalability, avoiding high computational complexity as much as possible so that the processing time can be reduced, and the simulator can be easily further developed.

Acknowledgments

It is our greatest gratitude to our supervisor Emma Fitzgerald and examiner Christian Nyberg, of the Department of Electrical and Information Technology, Lund University, for their valuable guidance and supervision through entire process. Special acknowledgement to our supervisors Christer Östberg and Henrik Ronkainen from System & Technology, Ericsson Lund for their immense support and continuous feedback throughout our thesis work. Lastly, we want to thank our families and friends in Syria, India and Sweden for all their best wishes.

Contents

List of figures	vi
List of tables	vii
List of acronyms	viii
Popular Science Summary	xi
1. Introduction.....	1
1.1. Project aims and contributions	1
1.2. Background and motivation	2
1.3. Approach and methodology	3
1.4. Main Challenges	4
1.5. Limitations	4
1.6. Thesis outline	5
2. Related work	7
2.1. Background Theory	7
2.1.1. Cellular Network	7
2.1.2. Poisson and Erlang calculations	8
2.1.3. Scheduling Algorithms	12
2.1.4. Link Adaptation	13
2.1.5. Frame structure	14
2.1.6. Physical resource block	15
2.1.7. 5G New radio	15
2.1.8. Downlink Throughput calculation	18

2.2. Previous related work	20
3. Design & Implementation	23
3.1. Overall system block diagram.....	23
3.2. Traffic simulation model	24
3.2.1. Heterogenous cellular network	24
3.2.2. Cellular Environment	26
3.2.3. General traffic model	27
3.3. Scheduling Algorithms	29
3.3.1. Round Robin	30
3.3.2. Maximum Throughput	34
3.3.3. Maximum PRB calculation	37
4. Results	39
4.2. Evaluation of Traffic Model	39
4.3. Evaluation of the Scheduler	43
4.3.1. Actual resource demand	43
4.3.2. Cell data rate	46
4.3.3. Processing time of simulator	49
5. Conclusions	53
6. Future work	55
References	57

List of figures

Figure 1.1: Thesis work methodology

Figure 2.1: Basic depiction of Cellular Network

Figure 2.2: Generalized downlink packet scheduling

Figure 2.3: LTE Frame Structure

Figure 2.4: Physical Resource Block

Figure 2.5: 5G use-case classification

Figure 2.6: 5G NR Subframe Structure (TDD)

Figure 2.7: LTE TDD frame structure

Figure 3.1: Overall system block diagram

Figure 3.2: Basic heterogenous cellular network architecture model

Figure 3.3: Flowchart for Round Robin Scheduling Algorithm.

Figure 3.4: Illustration of Round Robin queuing and scheduling process - time quantum ($TQ = 2$)

Figure 3.5: Illustration of Round Robin queuing and scheduling process - time quantum ($TQ = 4$)

Figure 3.6: Flowchart for Maximum Throughput Scheduling Algorithm

Figure 3.7: Link Budget

Figure 3.8: Link quality for 10 User Equipment

Figure 4.1: Illustration of mapping in the timeline for 10 UEs

Figure 4.2: Illustration of snapshot of the timeline for 10 UEs

Figure 4.3: Illustration of timeline for entire simulation window for 10 UEs

Figure 4.4: Visualization of the traffic for 3 UEs

Figure 4.5: Data rate for one UE with profile 4 for LTE

Figure 4.6: Data rate for one UE with profile 4 for 5G-NR

Figure 4.7: Data rate for 10 UEs for entire timeline (MT)

Figure 4.8: Data rate for 10 UEs for entire timeline (RR)

Figure 4.9: Variance and standard deviation for RR and MT schedulers.

Figure 4.10: Confidence Interval for test 1,2,3,4 and 5.

List of tables

Table 2.1: 5G NR Subcarrier spacing for different frequency ranges

Table 3.1: Service list and their characteristics

Table 3.2: User service profiles

Table 3.3: Number of bursts per average session duration

Table 3.4: Example of Round Robin scheduling with main design factors - time quantum ($TQ = 2$)

Table 3.5: Example of Round Robin scheduling with main design factors - time quantum ($TQ = 4$)

Table 3.6: 3GPP CQI Table [19]

List of acronyms

- 3GPP – The 3rd Generation Partnership Project 5G
- AT – Arrival time
- BS – Base station
- BT – Burst time
- BTS – Base Transceiver Station
- BSC – Base Station Controller
- CN – Core network
- CP – Cyclic Prefix
- CPRI – Common public radio interface
- C-RAN – Centralized Radio Access Network
- CT – Completion Time
- DL – Downlink
- DwPTS – Downlink Pilot Time Slot
- eCPRI – Enhanced Common Public Radio Interface
- eNB – eNodeB
- eMBB – Enhanced Mobile Broadband
- FCS- First Come First Serve
- FDPS – frequency domain packet scheduler
- FR – Frequency range.
- GBR – Guaranteed Bit Rate
- GSA – Global mobile Suppliers Association
- GP – Guard period
- HARQ – Hybrid automatic repeat request (hybrid ARQ)
- HetNet – Heterogeneous Cellular Network
- HO – Handover
- IMT-Advanced – International Mobile Telecommunications-Advanced (4G)
- IMT-2020 – International Mobile Telecommunications-2020 (5G)
- ISD – Inter site distance
- KPI – Key Performance Indicator
- LTE – Long-Term Evolution
- MBB – Mobile Broadband
- MCS – Modulation and Coding Scheme
- MIL – Maximum isotropic loss

- MIMO – Multiple Input Multiple Output
- mmWave – millimeter wave
- mMTC – massive machine-type communication
- MNOs - Mobile Network Operators
- MT – Maximum Throughput
- MTC – Machine type communication
- M2M – Machine-to-machine
- NB – NodeB
- Non-GBR – non-Guaranteed Bit Rate
- NSA– Non-standalone
- NR – Mobile Broadband
- OFDM – Orthogonal Frequency Division Multiplexing
- PCFICH – Physical Control Format Indicator Channel
- PDCCH – Physical Downlink Control Channel
- PF – Proportional fair
- PDN – Packet Data Network
- PRB – Physical Resource Block
- PSS – Primary Synchronization Signal
- QoS – Quality of service
- QCI – QoS Class Identifier
- R – Cell radius
- RACH – Random Access Channel
- RAN – Radio access network
- RAT – Radio access technology
- RNC – Radio Network Controller
- RB – resource block
- RR – Round robin
- RRC – Radio Resource Control
- RT – Remaining Time
- SA– Standalone
- SIP – Session Initiation Protocol
- SRS – Signaling Reference Signal
- SS – synchronization signal
- SSF – Special Subframe
- TAT – Turnaround time
- TD-BET – Time domain blind equal throughput

- TB – Transport Block
- TDD – Time Division Duplex
- TD-MT – Time domain maximum throughput
- TDD – Time Division Duplex
- TQ – Time quantum
- TRxP - Transmission reception point
- TTI –Transmission Time Interval
- UE – User Equipment
- UL – Uplink
- UMa – Urban Macro
- UMi – Urban Micro
- UpPTS – Uplink Pilot Time Slot GP Guard period
- URLLC– Ultra-Reliable and Low-Latency Communication
- UT – User Terminal
- WT – Waiting Time

Popular science summary

As there is a saying, “*All models are wrong; but some are useful*” by George E. P. Box, aiming to implement an efficient model that models user’s traffic in modern systems and provisioning them with the actual resource demands, is considered to be very useful.

This Master Thesis has created the fundamental basic parts of the simulation tool such as the flexible traffic model for each user, the definition of the area covered by heterogenous cells, the positioning of the users in the cell area, a basic radio model to reflect the path loss of the radio signals due to the distance between the transmitter and receiver, and finally the fundamental allocation of the radio resources for the created traffic performed in the scheduler. Furthermore, the execution time of the simulator have been measured.

The cellular area has been modelled with different number of users, each using different type of services like voice, video streaming, file downloading, web browsing, etc. The characteristics of the services is part of the modelling such as if data is transmitted intermittently (bursty) or in a continuous stream. Based on these services, the users can be categorized as e.g. a typical office user, a social networking user, or user who streams a large number of videos.

Over the years, in each generation of cellular networks, user service demands are growing exponentially. The existing simulators and traffic dimensioning methods are not flexible in terms of type of traffic. Therefore, multiple simulations for different user density are performed and reasonable results for traffic dimensioning are observed.

We are serving the users by implementing two scheduling algorithms. In general, scheduling means to arrange or plan an event that takes place at a particular time and here it is referred to providing the users with required resources. One of the main difficult tasks is choosing scheduling algorithms out of number of available schemes. As the chosen Round Robin and Maximum throughput algorithms provide a certain result in terms of throughput and latency but the simulator can be expanded with other scheduling algorithms targeting different user characteristics/demands.

When the traffic demands of the users are accurately evaluated, mobile network operators can provide ensured quality of service. In this way, the thesis project work can be used as a useful tool for evaluating the new user traffic demands in the upcoming 5G technology as well. Hence, this will lead to a positive impact on the society as the service demands can be guaranteed to be fulfilled.

1. Introduction

This chapter is divided into different sections namely, Project Contributions, Background and Motivation, Approach and Methodology, Main Challenges, Limitations and Thesis Outline.

1.1 Project Aims and Contributions

Aims

The ever-increasing bandwidth demands from the network users imply increased capacity requirements, not only for the RAN processing but also for the backhaul transport which becomes a significant cost for the sites. To address this issue, other ways to optimize the network deployments are explored, such as Centralized RAN (C-RAN), where the baseband resources are centralized, connecting to radio units at antenna site(s) over digital fronthaul interfaces (CPRI). In this way, trunking effects in the backhaul interface can be achieved, as the likelihood for peak rate in all antenna sites decrease when number of radio units connected to the same centralized baseband increase. Within the industry, standards are evolving to further optimize the CRAN deployment, where one improvement is the packet based fronthaul interface (eCPRI), enabling a more flexible usage of baseband resources compared to the legacy CPRI interface. A clear benefit of the packet based fronthaul is the possibility for pooling of baseband resources which in the end reduces the deployment cost.

The overall situation described above implies new challenges when it comes to dimensioning of the baseband and transport resources. General Erlang calculations cannot be used as the user services vary and the actual resource demand for one and the same service differ for different users depending on their individual radio channel conditions. The same is valid also for deployment scenario simulations (with the purpose to identify spectrum demands and capacity (bps/area)) as they will not answer the questions induced above and thus, other means are required to satisfy these needs. The purpose with this thesis project is to address these questions and create a MATLAB simulator framework which can analyze a defined network considering different user densities and different traffic models.

The MATLAB simulator can create different deployment scenarios covering node density and node spectrum allocation. Within the deployment, users are randomly populated and traffic to each user is created based on a selected traffic model. The actual resource demand for each user is adapted to the user's channel conditions and thus, two different users with exactly the same service profile typically have different resource demands. The goal of this master thesis as set by Ericsson is to create the first framework of the simulator where finalized version of the simulator is estimated to be completed in approximately two more future thesis works. The details about the targeted capabilities of the finalized version are provided in future work chapter.

Contributions

This Master Thesis has created the fundamental basic parts of the simulation tool such as the flexible traffic model for each UE, the definition of the area covered by heterogeneous cells, the positioning of the UE's in the cell area, a basic radio model to reflect the path loss of the radio signals due to the distance between the transmitter and receiver, and finally the fundamental allocation of the radio resources for the created traffic performed in the scheduler. Furthermore, the execution time of the simulator has been measured.

Here, we summarize the main contributions in this thesis:

- Creating MATLAB basic simulator framework that can analyze a defined network considering different user densities and different traffic models.
- The developed traffic model can visualize various types of services which required for traffic analysis.
- The simulator framework can create different basic deployment scenarios covering node density and node spectrum allocation.
- Implementing channel-aware scheduling algorithm that considers the actual resource demand for each user/service and efficiently allocates radio resources based on the user's channel conditions (efficient spectrum usage).
- The simulator framework is flexible and scalable.

1.2 Background and Motivation

Wireless cellular communication networks are in a continuous development to meet the conventional user requirements and the envisioned use cases of the modern networks. In general, we have voice and data, where data can have different characteristics such as best effort, GBR, streaming video or streaming audio etc. The forecasted volumes for mobile data traffic is very high, mainly driven by the increased demand for video services. Global data traffic is estimated to increase to 49 exabytes (49 EB) per month by 2021, a seven-fold increase over 2016 as stated in the Global Mobile Data Traffic Forecast Update by Cisco [1] (1EB = 10^{18} bytes, One exabyte is equivalent to one million terabytes).

The rollout of 5G 'New Radio (NR)' connectivity will result in new levels of fixed/mobile convergence, as cellular networks will provide 22 percent of global Internet traffic by 2022, (up from 12 percent in 2017) [2]. Unlike the conventional wireless use cases that essentially demand and focus on high mobile broadband (MBB), 5G NR must meet stringent latency and power requirements [3]. Considering all these rapid advances, it is essential for mobile network operators (MNOs) to analyze the traffic and investigate modern cellular network performance.

A cellular system is a group of cells covering an area which includes user equipment (UE), Radio access network (RAN) and interface to the core network

(CN). Each cell in a network can be served by one or multiple antennas operating at different frequencies with certain bandwidth. Each of these cells can be defined as Pico cells, micro cells, macro cells or even satellite cells based on coverage area. The operating frequencies of the antennas can be reused by another set of cells at specified distance that ensures no interference. UEs placed in the cellular area can handle different services like voice, video, and data traffic. These UEs are allocated shared radio resources based on their traffic requirements by one or potentially more schedulers depending on deployment and coverage situation.

Designing robust and reliable network is becoming increasingly difficult. One helpful tool is to develop a traffic model describing the characteristics of the network. How the network behaves differently for the different service parameters like data traffic, data distribution, latency, different protocols and with respect to the increased developments for the 5G system becomes an interesting research topic.

1.3 Approach and Methodology

This thesis project is performed using MATLAB as a primary software for implementation, analysis and evaluating the performance. Figure 1.1 illustrates the main steps used to carry out the project work.

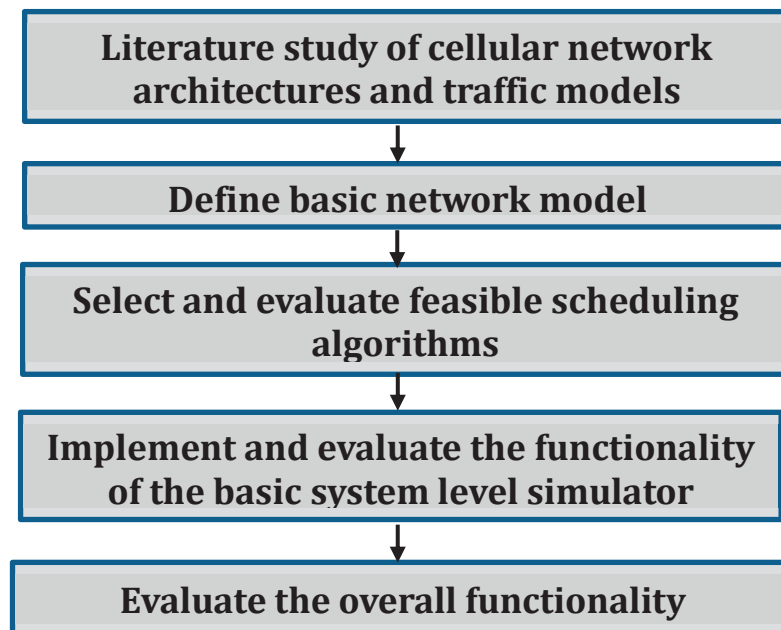


Figure 1.1: Thesis work Methodology

Step 1: This step involves studying about the suitable cellular architecture that includes different types of cells, their hexagonal geometry, ISD, frequency bands,

number of UEs, service categories and profiles with different service combination, UE distributions, etc.

Step 2: ISD is selected, based on this a cellular area is designed and traffic model is built including all the parameters as explained in step1. Pathloss and UE positions with respect to their base stations at the center of the cell are also determined.

Step 3: Studied various scheduling algorithms, and performance is compared to select the best suitable method for the overall architecture, keeping fairness and overall throughput as a selection criterion.

Step 4: Based on the first three steps, Round Robin and Maximum Throughput scheduling schemes are implemented and evaluated for the resource allocation on the traffic model designed.

Step 5: Overall simulator performance is evaluated for different channel conditions and scenarios. The results are plotted to provide complete understanding of our work.

1.4 Main Challenges

Some of the main challenges that we faced during the course of thesis work are listed below:

- Maintaining the average processing time as the number of users increases is very challenging.
- Deviation towards real time scenario while implementing measured traffic makes the process slow and results in unnecessary complications.
- Keeping track of all the parameters corresponding to a particular service of a specific user is complicated.
- Designing the model suitable for existing as well as for the coming 5G system will be a complex task because, the current existing models may not be compatible with the 3GPP requirements of 5G like Enhanced Mobile broadband (eMBB), Massive machine type communication (mMTC), and Ultra reliable and low latency communication (URLLC).

1.5 Limitations

The possible shortcomings of our thesis project are:

- Time – as the simulator can include many aspects and research possibilities to be a finalized version.
- Running all simulations ‘local’ rather than ‘On server’
-

1.6 Thesis Outline

This thesis work is organized into 6 chapters: Chapter 1 is an Introduction section. Chapter 2 is about background theory and previous related work. Chapter 3 gives detailed description of the channel model and environment, simulation setup, scheduler algorithms. Chapter 4 discusses the obtained results. Chapter 5 summarizes the conclusions based on results obtained. Chapter 6 includes future development related to the project work.

2. Related work

Several related work in this area have been done. It is worth mentioning that most of the study related work are dealing with either a specific aspect of network models and scheduling algorithms or evaluating and improving the performance based on provided system level simulators (by academia or industry). Before going deep into the previous related work, some of the basic concepts are described below.

2.1 Background theory

In this subsection, we mainly discuss the background knowledge of the concepts related to this thesis work namely, cellular network, types of scheduling algorithm, frame structure, physical resource block and 5G NR.

2.1.1 Cellular Network

A mobile network is a group of cells covering a particular area. Each cell in a network can be served by one or multiple antennas operating at different frequencies with certain bandwidth. Each of these cells can be defined as Pico cells, micro cells, macro cells or even satellite cells based on coverage area. The operating frequencies of the antennas can be reused by another set of cells at specified distance that ensures no interference. User terminals placed in the cellular areas can have different services like voice, video, live streaming, etc. These user terminals are then allocated with shared radio resources based on their traffic requirements and the area they are located by a set of schedulers.

Basic structure of a cellular network as shown in figure 2.1 includes the following units:

- **Core Network** - The main purpose of the core network (CN) is to facilitate the routes to exchange information among various users. It is also termed as Backbone network and is usually made up of switches or routers, with switches being used more frequently [7].
- **User Equipment** - User equipment is the type of device used by the end-users to communicate with the network. It can be a computer, mobile phone, broadband adapter or any device that has cellular antenna.
- **Radio access network** - As per [8], the radio functionality as well as providing connection to the core network is performed by the radio access network. In 2G, the RAN comprised of two main components: Base Transceiver Stations (BTS) and Base Station Controller (BSC). BTSs are the radio elements on the network side which are used to cover an area ensuring better coverage. Every BTS is connected to a cell site

and these cell sites consists of different types of cells based on the coverage requirements.

In the 3G system, BTS was replaced by NodeB (NB) and BSC was replaced by Radio Network Controller (RNC). Here, RNC controlled the actions of NB. In the 4G system, the RAN structure was changed by replacing the NB with eNodeB (eNB) and the RNC functionality was split between CN and the eNB. This resulted in a RAN consisting of interconnected eNBs that can handle the mobility directly and also some other additional features where, multiple eNBs can send data to the same UE to maximize the throughput.

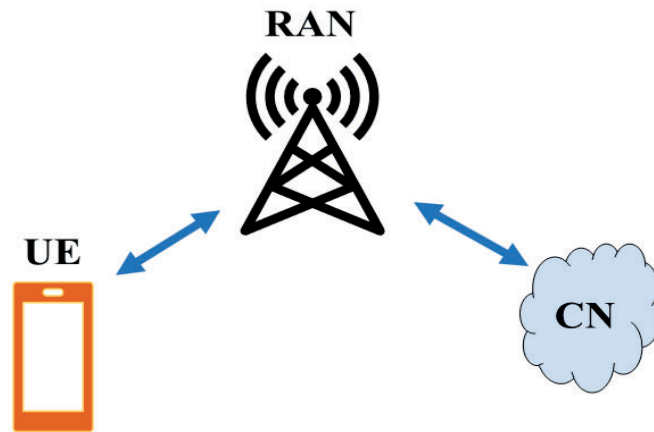


Figure 2.1: Basic depiction of Cellular Network [8]

2.1.2 Poisson and Erlang calculations

Poisson Process: The Poisson process [29] is one of the most important random processes in probability theory. It is widely used to model a series of discrete event where, the average time between events is known but their arrival time is independent, that is, waiting time between events is memoryless.

To illustrate it, consider as an example, a website goes down on average once per 60 days - as this availability provided by content delivery network (CDN)-but, once failure occurs that does not impact the probability of the next. The essential point is knowing the average time between events, but they are randomly spaced (stochastic).

In theory, Poisson Process should meet the following criteria:

- Events are independent of each other. The incidence of one event does not impact the incidence of another event.
- The average rate (events per time/space interval) is constant.
- Two events cannot occur at the same time

Last point means that each subinterval of a Poisson process can be deemed as Bernoulli Trial (or binomial trial), having either a success or a failure outcome. The provided example considers 60 days period as a measuring interval, the entire interval may be an interval of 600 days, but each sub interval, that is, one day the website either goes down or it does not.

Poisson Distribution: While Poisson process is a model for describing random occurrence of events, Poisson distribution [29] is necessary for doing statistics, for example, finding the probability of observing k events in a time interval when the events follow a Poisson process and length of the interval and the expected number of events per time are given.

Probability of events for a Poisson distribution [36] can be calculated using this formula:

$$P(k \text{ events in interval}) = e^{-\lambda} \times \frac{(\lambda)^k}{k!}$$

Where, λ is the average number of events per interval and k is non-negative integer represents the number of times an event occurs in an interval.

This equation can also be adjusted when the time rate of events (r) is given rather than the average number of events λ . Then $\lambda = rt$, where r is in units of 1/time. And hence the formula can be written as,

$$P(k \text{ events in interval } t) = e^{-rt} \times \frac{(rt)^k}{k!}$$

Erlang Calculations

What is an Erlang?

An Erlang [30] is dimensionless unit of measure for traffic density in a telecommunication system or network and it is extensively used for measuring load or efficiency. From Wikipedia [31], the erlang (symbol E) is a dimensionless unit that is used in telephony as a measure of offered load or carried load on service-providing elements such as telephone circuits or telephone switching equipment. A single cord circuit has the capacity to be used for entire 60 minutes in one hour. Full utilization of that capacity, 60 minutes of traffic, constitutes 1 erlang.

Erlang basics

Erlang [30] is mainly used as a statistical measure for the voice traffic density in a telecommunication system. It is essential for analyzing the traffic and understanding the required capacity in a network. Consequently, it helps in quantifying and performing the calculation of the traffic volume in a standard way. Telecommunication network designers use Erlang for understanding traffic patterns within a voice network and determining /evaluating the required capacity in any location of the network.

In a Cisco Technology White Paper for Traffic Analysis [32], Carried traffic is defined, the traffic that is actually serviced by telecommunication equipment. Offered traffic is the actual amount of traffic attempts on a system. Using the equation as shown below, offered load can be calculated from carried load. Note, this formula does not consider the retries that may occur when a caller is blocked.

$$\text{Offered load} = \text{carried load} / (1 - \text{blocking factor})$$

The above formula clearly shows that when the amount of blockage is small, the difference between the carried and offered load is also small. However, If the retry rate must be considered, the following formula can be used,

$$\text{Offered load} = \text{carried load} \times \text{Offered Load Adjustment Factors (OAF)}$$

$$\text{OAF} = [1.0 - (R \times \text{blocking factor})] / (1.0 - \text{blocking factor})$$

Where, R is a percentage of retry probability. For example, $R = 0.7$ for a 70 percent retry rate.)

Erlang function

Using the following simple function, the traffic can be calculated and expressed by the number of Erlangs that are required,

$$E = \lambda \times h$$

Where, λ is the mean arrival rate of new calls, h is the mean call length or holding time and E is the traffic in Erlangs.

Erlang development

Although basic concept of Erlang calculation, was a valuable tool for communication engineers to investigate load levels in different areas such as, call

centers and lines that connect different areas, this basic form did not consider some real-life aspects of loading for example, peak traffic density and number of blocked calls that result from short term overloading. Erlang B and Erlang C have been thus developed to address these aspects [30].

Erlang B

Erlang B concerns traffic loading in peak loading times and it can be calculated using the formula below:

$$B = \frac{\frac{A^N}{N!}}{\sum_{i=0}^N \left(\frac{A^i}{i!}\right)}$$

Where B is Erlang B loss (blocking) probability, N is the number of trunks (servers, telephone lines, etc.) in full availability group and A is traffic offered to group in Erlangs. This formula only applies in an Erlang system, that is when the arrivals follow a Poisson process, there is no queue, and the service time is exponentially distributed.

It is worthwhile to mention that Erlang B requires M/M/n/n system and that call arrivals can be modeled by a Poisson process, which is not always applicable in real life.

Erlang C

Erlang C formula addresses queuing aspects as it calculates the probability of queuing offered traffic, but it assumes that blocked calls stay in the system until they can be handled.

$$C = \frac{\frac{A^N}{N!} \frac{N}{N-E}}{(\sum_{i=0}^{N-1} \frac{A^i}{i!}) + \frac{A^N}{N!} \frac{N}{N-E}}$$

Where C is the probability that a customer has to wait for service, N is the number of servers and E is the total traffic offered in units of erlangs.

Moreover, it is worth highlighting that Erlang C assumes that call arrivals can be modeled by a Poisson process [31].

Erlang summary

Erlang calculations are still important part for communication network planning today. However, it worth remembering that these formulae are based on assumptions, that is, Erlang B assumes that a caller will not instantly try again after receiving a busy tone and Erlang C assumes that the caller will never hang up

while in queue. Furthermore, Erlang equations are also based on statistics, hence they require large number of resources to give accurate results.

2.1.3 Scheduling Algorithms

Scheduling is a mechanism at the eNB that is responsible for allocating shared time-frequency resources among the users as depicted in figure 2.2. It decides how many resources should be allocated to a user for data transmission and also the time at which it gets allocated. The decision of the scheduler is based on multiple principles for example fairness, throughput, delay, etc.

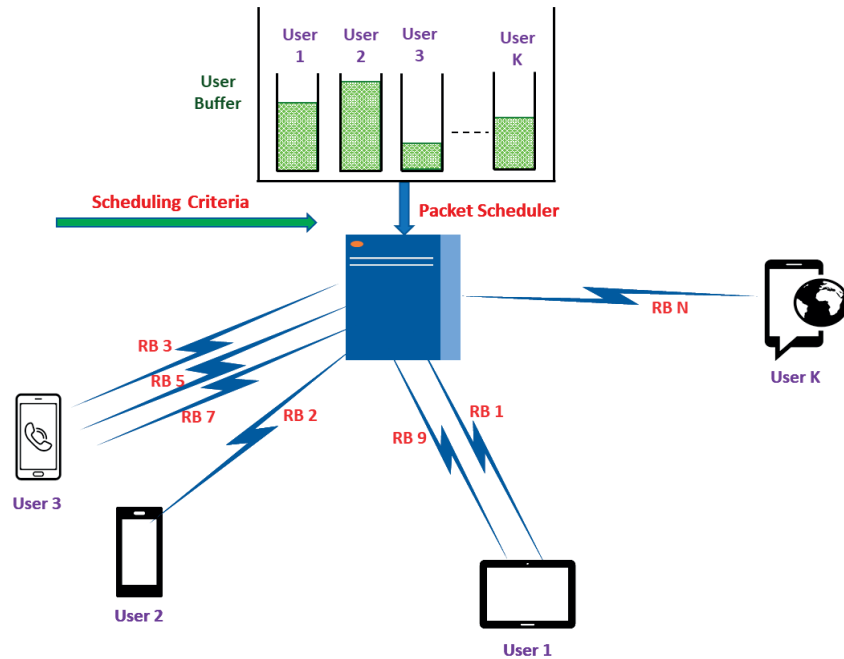


Figure 2.2: A generalized downlink packet scheduling [21]

Scheduler can consider number of functions for efficient resource allocation and they are [10]:

- **Link Adaptation** – performs the selection of transmission mode, modulation and coding scheme based on the radio link condition. More elaboration on Link adaptation is done in subsection 2.1.4.
- **Rate Control** – it controls the allocation of the resources among the radio bearers of same UE both at uplink and downlink.

- **Packet Scheduler** – for every TTI, access to the air interface is decided by the packet scheduler for all the active users.
- **Resource Assignment** – it assigns the air interface resources to the selected active users in every TTI.
- **Power Control** – it is mainly responsible for providing desired SINR that is, improving the power of the received signal and limiting the interference with the neighboring cells.
- **HARQ (ARQ+FEC)** – whenever a packet is received incorrectly, HARQ is implemented which is a combination of error detection and error correction. This mechanism therefore corrects the received erroneous data instead of discarding them.

Some of the scheduling schemes are summarized below.

- **Round Robin (RR) Scheduling Algorithm** - Round Robin is the most basic scheduling algorithm which does not take channel conditions into account for resource allocation. It provides fixed amount of resources to every user regardless of the user's requirement. Therefore, offering very good fairness but degrading the system throughput performance. [11]
- **Maximum Throughput (MT) Scheduling Algorithm** - In this scheduling algorithm, the metric for resource allocation is experienced channel quality. The eNB receives the information about the Channel quality from every user and allocates the resource to the user with best channel condition. This results in the maximum throughput of the system, but, unfair resource allocation as the users with bad channel condition may suffer from starvation. [4]
- **Proportional Fair (PF) Scheduling Algorithm** - Proportional fair is a scheduling algorithm which maintains a balance between the throughput of the system as well as allows all the users to be able to access minimal amount of resources providing fairness. This method is based on channel condition and also avoids the users that are experiencing bad channel quality to starve [12]. Hence, proportional fair acts as a tradeoff between round robin and maximum throughput scheduling algorithms.

2.1.4 Link Adaptation

Link adaptation, or adaptive coding and modulation (ACM), is the ability to adapt the modulation and the coding scheme (MCS) and other signal and protocol parameters according to the quality of the radio link. The conditions of the radio link are usually characterized by the path loss, the interference due to signals coming from other transmitters, the sensitivity of the receiver, the available transmitter power margin, etc.

If the channel quality is good, a high-level efficient modulation scheme and small amount of error correction is used (low redundancy). Hence, a high throughput can be provided over the radio channel. However, if the channel condition is poor, a low level, more robust modulation scheme is used, and the amount of error correction is increased (high redundancy). Hence lower data throughput can be provided over the radio channel. In very bad link conditions, retransmissions due to HARQ ensures that the sent information is received correctly and in this case the bit rate is further decreased. [20]

2.1.5 Frame Structure

The frame structure of the radio resource is demonstrated here where Time Division Duplex (TDD) mode is explained. However, the frame in Frequency Division Duplex (FDD) system has similar structure but uplink and downlink are separated by frequency. From [9], in this mode, uplink and downlink operates on same frequency but at different time periods. Each frame is of 10ms duration and is divided into 10 subframes of 1ms each. The subframes have two slots and each slot is of 0.5ms duration. This slot consists of 6 or 7 Orthogonal frequency division multiplexing (OFDM) symbols for extended or normal cyclic prefix (CP) respectively. The frame structure is illustrated in figure 2.3 below.

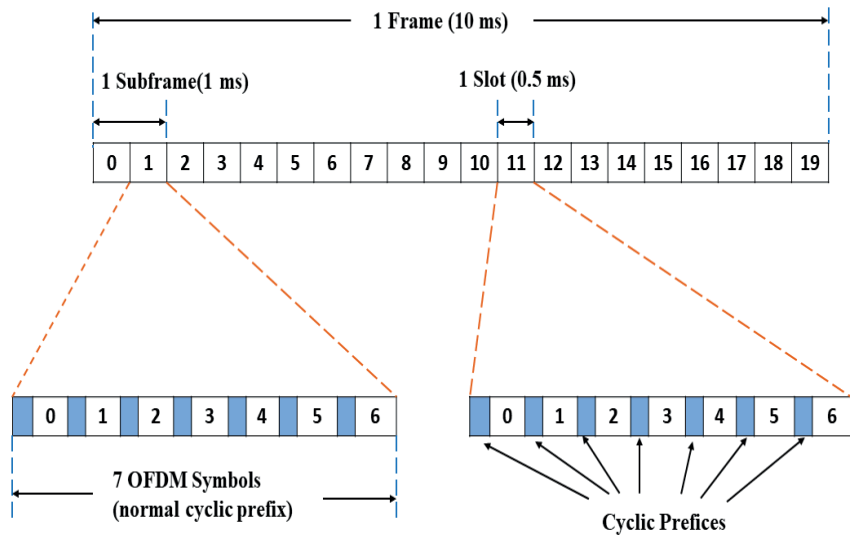


Figure 2.3: LTE Frame Structure

2.1.6 Physical Resource Block

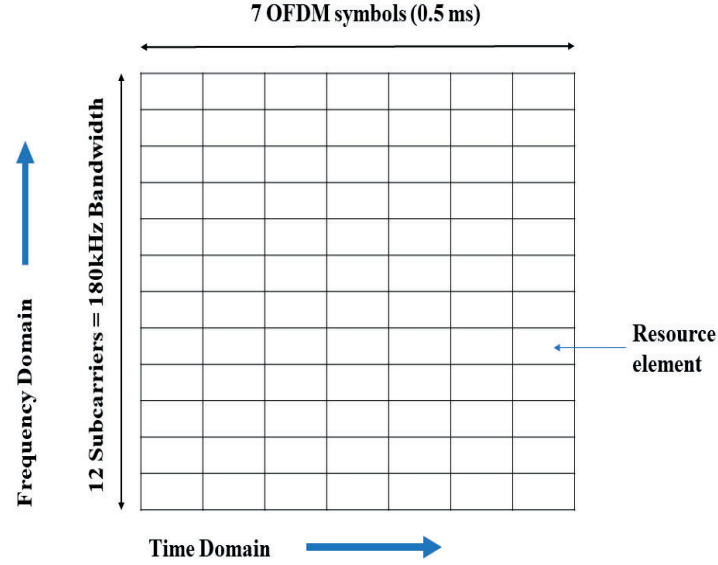


Figure 2.4: Physical Resource Block [9]

The physical resource block is the smallest resource unit that can be allocated to the user in both time and frequency domain. Each RB has a duration of 0.5ms that is, one time slot and consists of 12 subcarriers. We considered a subcarrier spacing of 15kHz and normal cyclic prefix with 7 OFDM symbols which results in bandwidth of 180kHz. Therefore, total resource elements of $12 \times 7 = 84$ in one RB is obtained as shown in figure 2.4.

2.1.7 5G New Radio (5G NR)

5G New Radio is a new radio access technology (RAT), its Non-standalone (NSA) and standalone (SA) specifications were approved by 3GPP in December 2017 and June 2018 respectively. Unlike the conventional wireless use cases that mostly demand and focus on high data mobile broadband (MBB), 5G New Radio further enhances the existing technologies and most importantly opens new business opportunities and strives to ensure forward compatibility. [14]

5G NR supports operation in wide and new(high) frequency ranges from below 1 GHz up to 52.6 GHz in both licensed and unlicensed spectra as shown in Table 2.1. The main advantage of this flexibility is that on one hand, it enables the

interworking and coexistence with LTE. And on the other hand, provides enhanced data rates, much lower latency, better coverage and higher reliability, which are crucial for the new use cases. Massive MIMO (minimum 16X16 array [38]) at mmWaves (up to 400 MHz bandwidth) is mainly used for beamforming to achieve coverage and increased received signal power(gain). Whereas, at lower frequency bands it enables full dimensional MIMO and avoids the interference with the help of spatial filtering, and hence significantly increases spectral efficiency. [3]

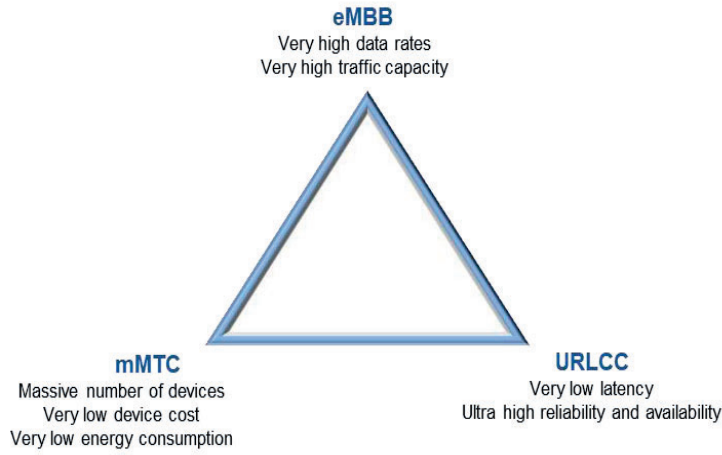


Figure 2.5: 5G use-case classification [15]

Some use cases require low energy consumption such as massive machine type communication (mMTC) and others are ultra-reliable low latency (URLLC) type cases. These requirements are met in 5G NR essentially by the flexibility in transmission schemes and frame structure. For example, the shorter slot duration resulting from the higher subcarrier spacing (up to 120 kHz) at higher frequency bands (up to 52.6 GHz) can support lower latency transmission. However, this increase of subcarrier spacing, shrinks the cyclic prefix, and hence this method is not practical in all deployments. Therefore, 5G NR can use a more feasible approach, that is, the possibility of transmission within a fraction of the slot, called, “mini slot” transmission, where latency-critical data transmission can start at any fraction of the slot (not necessarily at the slot boundary). Moreover, the NR device-side receiver-bandwidth adaptation enhances device energy efficiency. [3]

Frequency band	Subcarrier spacing	Maximum bandwidth
0.45 GHz – 6 GHz	15/30/60 kHz	50/100/200 MHz
24 GHz – 52.6 GHz	60/120 kHz	200/400 MHz

Table 2.1: 5G NR Subcarrier spacing for different frequency ranges. [3]

Several technical requirements of IMT-2020 and 3GPP are defined for NR. In [3], the achieved performance is illustrated that is, NR capabilities compared with the defined requirements. In addition to that it explains how they are achieved. The NR performance shows that MBB will reach double the rate of the ITU requirements of 20 Gbps downlink (DL) and 10 Gbps uplink (UL). This is achieved by utilizing wide uncongested bandwidths. Moreover, user experienced data rates 100 Mbps (DL) and 50 Mbps (UL) is to be achieved in 95% probability in loaded conditions. In addition, user plane (UP) latency is less than 1ms. Massive MIMO along with high and low bands interworking make the spectral efficiency three times higher than required in 4G (IMT-Advanced). mMTC performance is also greater than the requirement of 1,000,000 devices /km². URLLC has achieved 99.999% success probability with less than 1ms delay at the cell edge. This is achieved by means of low code rates, antenna diversity (→reliability) and fast and adaptive transmission (→ low latency). Ultra-lean transmission gives ten times improvement in energy performance; as it minimizes the always-on transmissions where the periodicity of the synchronization signal (SS) block is increased to 20ms and hence providing longer sleep time for more devices.

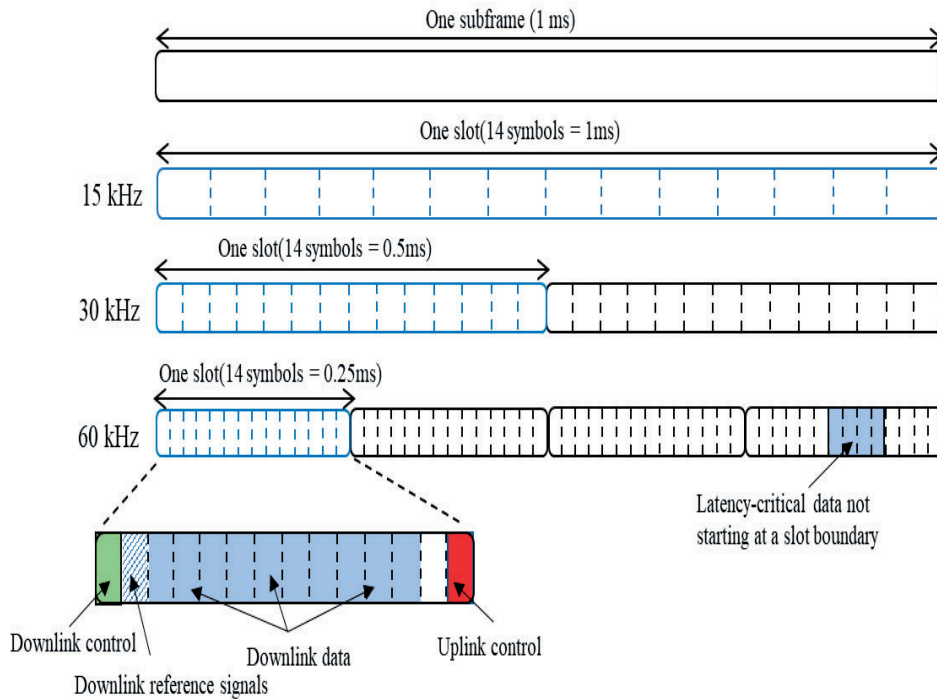


Figure 2.6: 5G NR Subframe Structure (TDD)

2.1.8 Downlink Throughput Calculation

In this thesis work both TDD and FDD systems are considered. TDD is unpaired spectrum (the same bandwidth is shared by Uplink and Downlink on time sharing basis) and extra parameters must be considered for throughput calculations.

The total (DL and UL) throughput in TDD is shared according to the time-sharing strategy. Figure 2.7 is an example that illustrates a possible TDD frame structure configuration, it has a duration of 10 milliseconds, consists of Downlink subframe, Uplink subframe and Special frame. TDD frame can have up to 7 different configurations. However, the frame should always start with a Downlink subframe to advertise the frame descriptor information, that is, PCFICH and PDCCH. Hence the UE learns the frame structure by the help of this subframe. The third subframe is always used for Uplink transmission. Special subframe is only needed when switching from Downlink to Uplink as, this guard period is needed at the UE side to avoid the time advanced UL to collide with the delayed DL. Whereas, it is not needed for switching from Uplink to Downlink as eNodeB has timing advance feature additionally, in UL to DL case the chance of collision is lower.

Since special subframe (SSF) impacts the cell size, it plays an important role when designing and deploying a cellular network, it is worth to define the functions of its main parts as in [23]:

- **DwPTS** – Downlink pilot time slot is considered as a “normal” DL subframe and carries reference signals, control information as well as data, for those cases when sufficient duration is configured. It also carries PSS (Physical synchronization signal).
- **GP** – Guard period is used to control the switching between the UL and DL transmission. Switching between transmission directions has a small hardware delay for both UE and eNodeB and needs to be compensated by GP. GP must be large enough to cover the propagation delay of DL interference. Its length determines the maximum supportable cell size.
- **UpPTS** – Uplink pilot time slot is primarily intended for sounding reference signals (SRS) transmission from UE. Mainly used for RACH (Random access channel) transmission. [23]

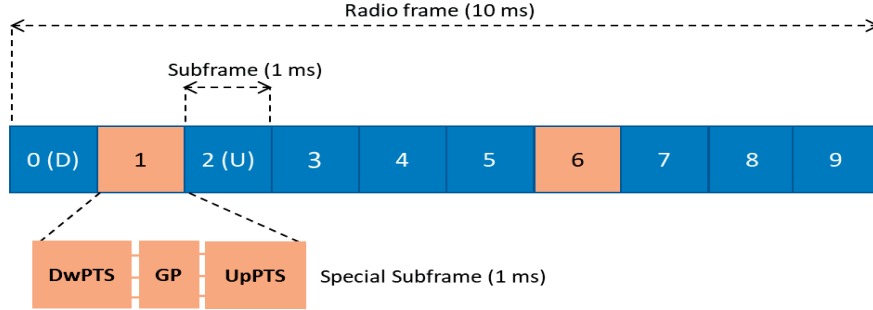


Figure 2.7: LTE TDD frame structure (one possible configuration)

DL and UL Throughput in TDD can be calculated using following equations 2.1 and 2.2:

$$DL\ Throughput = Number\ of\ Chains \times TB\ size \times (Contribution\ by\ DL\ Subframe + Contribution\ by\ DwPTS\ in\ SSF). \quad (2.1)$$

$$UL\ Throughput = Number\ of\ Chains \times TB\ size \times (Contribution\ by\ UL\ Subframe + Contribution\ by\ UpPTS\ in\ SSF). \quad (2.2)$$

Where,

- Number of Chains: refers to the number of spatial layers that allows the transmission of multiple layers when using MIMO systems.
- TB size: Transport block size represents the total number of symbols per Transmission Time Interval (TTI = 1ms in LTE) multiplied by number of bits per symbol (modulation order) where, the controlling and signaling bits are excluded. Hence, the TB size is proportional to the used system bandwidth and to the modulation and coding scheme (MCS). For example, in LTE for 20 MHz, there are 100 Resource Blocks, each resource block has $12 \times 7 \times 2 = 168$ symbols per millisecond in case of normal CP. So, there are 16800 Symbols per millisecond. If the modulation used is 64 QAM (6 bits per symbol) then, we will have $16800 \times 6 = 100800$ bits per millisecond which is equivalent to $\frac{100800}{10^{-3}} = 100.8$ Mbps. However, in LTE about 25 % of these bits are used for controlling and signaling, this should be excluded when calculating the effective throughput. After considering 25 % as overhead, the transport block size is about 75600

bits per 1ms which is equivalent to $\frac{75600}{10^{-3}} = 75.6$ Mbps (coding rate =1 is assumed here). In this thesis work, we assume the controlling and signaling overhead for 5G NR to be 24% (as suggested from Ericsson).

- Contribution by DL Subframe: refers to the number of used downlink-subframes per frame based on the desired configuration. For example, if the Contribution by DL Subframe equals to 0.6 that means 6 subframes are used for downlink and the 4 remaining subframes are for UL and special subframes.
- Contribution by DwPTS in SSF: refers to the percentage of special subframes (out of 10) multiplied by percentage of downlink symbols (out of 14), where the percentages are determined based on the selected standardized configuration [24].

2.2 Previous related work

There are many papers and research projects based on traffic model analysis and resource allocation. This section will mainly give an idea about what exactly our project involves in comparison with related work.

In [26], a master thesis project is carried out for analyzing traffic models for Voice over LTE. Here, the student performed the research on the LTE network of Magyar Telecom Plc. The real SIP traffic was obtained and investigated for base of network traffic analytics and modelling. The stationarity of all the statistical properties like mean, variance, autocorrelation over time was examined using some MATLAB functions. The received traffic was verified by creating a Poisson model using a single queuing system for simulation test in MATLAB. Comparison between the behavior of the queue for both measured traffic and generated traffic showed that Poisson model described the network traffic well.

In [27], a detailed analysis of traffic model for machine type communication is performed at Vienna University of Technology. The different traffic patterns and simulation scenarios exhibited by M2M or MTC were dealt by proposing an approach that overcame the challenges faced by the source and aggregated models. It was found that traffic models based on MTC faced many challenges like, i) Source models were designed to capture behavior of each MTC device with good precision and feasibility but increased complexity with increased number of devices, ii) Whereas, Aggregated traffic that combined all devices into one stream resulted in lower complexity but less precision. Therefore, it was observed from the proposed approach namely, Coupled Markov Modulated Poisson Processes (CMMPP) framework provided an alternative for aggregated model by making massive parallel deployments feasible which reduced computational cost and increased achievable accuracy.

As we can see that these research projects are based on a particular type of traffic and they provide the analysis for either an existing model or for a new

approach, but our aim is practical implementation of a traffic model that makes it possible to make a snapshot in time where different traffic types, such as voice, video and data appear and the radio resources are allocated according to the selected scheduling algorithm. Rather than providing deep analysis for a specific traffic, we concentrate on designing a general model as a requirement for our simulator. The model will be able to adapt to LTE features as well as 5G NR features without any complications.

In [34], The common traffic model Erlang C was studied to evaluate its practicability in real call centers. Erlang C assumes that callers never hang up (ignores caller abandonment) while in queue and that calls arrive according to Poisson process. The prediction performance of Erlang C is compared with a call center simulation model. The results shows that Erlang C is prone to considerable errors in estimating system performance. After conducting experimental analysis, the authors present various important observations. When real systems experience higher levels of hang-up calls, measurement errors is high, and this error is substantially positively correlated with the realized abandonment rate. Hence, Increased caller patience leads to decrease in errors. When the number of agents is large, and utilization is low, Erlang C model accuracy becomes high. Erlang C is normally pessimistically biased, that is, the real system performs better than predicted by Erlang C. Erlang C continued popularity can be attributed to its pessimistic estimates. In order to achieve high levels of precision, Erlang C assumptions must be met – which cannot always be fulfilled in reality.

In [35], the authors analyze Erlang A model (a queuing model that allows for abandonment and used to analyze call center performance) and evaluate its accuracy predictions in high traffic environments as the Erlang C model is not applicable in this high traffic region. The results shows that in this region, Erlang A is pessimistically biased and is prone to a moderate to high level of errors. In addition, the model is to great extent sensitive to arrival rate uncertainty. Finally, the study confirms that under realistic conditions, even the advanced Erlang A model is liable to considerable error.

Overall, Erlang is equivalent to one voice call, that is, it can still be used in a 2G system where each voice call occupies a fixed set of time slots in the air interface. However, in today's cellular networks the resource demand for a certain service differs depending on the channel conditions and thus, there is no way to express the resource demand required for 1 Erlang. Erlang C is a simple and common model that ignores caller abandonment. Although Erlang A considers the abandonment, it introduces additional complexity as the calculation of the performance measures is more complicated. However, our traffic model is a simulation-based model and is designed to be a general and modern traffic model that visualizes different traffic types without introducing any additional complexity. In addition, dimensioning resource demands based on the channel conditions is facilitated.

In [4], key design issues of downlink packet scheduling in LTE cellular network has been thoroughly surveyed. The authors divided the scheduling strategies into three major categories, that is, channel-unaware strategies, channel-aware/ QoS-unaware strategies and channel-aware/ QoS-aware strategies. In each one of these categories, several scheduling methods were studied, and their performance evaluated according to several important design parameters mainly accuracy and the computational complexity. Proportional fair (PF) scheme (channel-aware/ QoS-unaware strategy) was used as a reference strategy to compare performance of different proposed solutions in the literature. After analyzing simulation results, authors found that there exist many interesting solutions but due to their implementation difficulty and high computational cost, they cannot be deployed in real systems. Furthermore, a dynamic and robust strategy should have flexible parameter settings so that, its ability to work in different scenarios can be guaranteed.

In [6], several time and frequency domain algorithms are investigated. For example, time domain blind equal throughput (TD-BET) scheduling method improves the fairness between users by maintaining equal user throughput, independent of the user's location (channel- unaware strategy). Its metric is based only on the past delivered user throughput, the lower the past throughput, the higher the priority metric. On the other hand, time domain maximum throughput (TD-MT) can maximize the overall cell throughput by assigning resource blocks (RBs) mainly to the users with high channel quality (channel- aware strategy). Thus, this method is unfair regarding resource sharing with respect to the users experiencing bad channel conditions. However, time domain proportional fair (TD-PF) scheduler can provide balance between throughput and fairness by utilizing the abovementioned strategies together, that is, TD-BET and TD-MT. Hence TD-PF metric is based on two parameters, past average delivered user throughput (as a denominator value), and the expected instantaneous data rate for the user (as a numerator value). Depending on the used metric, a time domain scheduler chooses N active users that have the highest scheduling priority which then can be an input to a frequency domain (FD) scheduler. At the frequency domain packet scheduler (FDPS), the resources are physically allocated. This joint method decreases the number of candidate users for resource allocation, resulting in a reduced computational complexity.

Before exploiting the joint method in the simulator that we are building, basic scheduling algorithms, one Round Robin which is mainly based on fairness and other Maximum throughput which is on other hand based on best throughput are implemented. The results are obtained for different scenarios and are compared to check for the quality of the simulator mainly based on the processing time and its adaptive behavior. When these basic scheduling algorithms are successfully executed in the new designed simulator, other scheduling algorithms can be implemented.

3. Design & Implementation

3.1 Overall system block diagram

The overall model consists of the following entities:

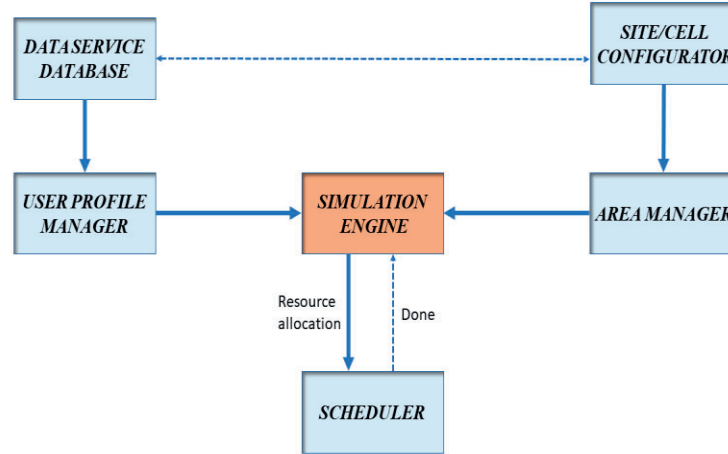


Figure 3.1: Overall system block diagram

As shown in Fig.3.1, to assess the performance of the simulation engine, we need inputs from:

- **User profile manager:** Services that are characterized based on average data rate, burst interval, burst size along with the frequency per hour and priority are listed in this section. Multiple services are combined to create a profile with certain distribution and is later assigned to the UEs.
- **Area manager:** This section consists of macro, micro cell sites and users for which simulation is done. By using the selected Inter-site distance (ISD), sites are deployed to obtain the desirable radio coverage. Each site is populated with randomly uniformly distributed user density. Few parameters like pathloss of the user from the BS, position of the user and the cell IDs are computed and listed together.
- **Scheduler:** Obtains the list of active users from simulation engine and implements scheduler algorithms to allocate the radio resources.

In the Simulation engine, every user from the area manager is assigned with a profile from the user profile manager for the simulation purpose. We also determine the number of active users in each cell, so that scheduling can be applied only on the active users. Finally, we aim to obtain a clear comparison of performance of the simulator for different number of users in terms of average

throughput, average processing time with calculated confidence interval and also accuracy of the results obtained for measured traffic and that of the real time traffic.

3.2 Traffic simulation model

In this section, we illustrate the design of the basic heterogenous cellular network model, traffic model and cellular environment.

3.2.1 Heterogeneous Cellular Network

Today, Mobile Network Operators (MNOs) must handle growing speed of data, voice and video traffic. The latest update of Ericsson Mobility Report [16] shows that, mobile data traffic grew close to 88 percent between Q4 2017 and Q4 2018. The main reason for this growth is the rising number of mobile subscriptions and average data volume per subscription where, viewing more video content plays a major role [16]. Although, 5G NR can provide more capacities, MNOs still need to use the limited spectrum resources more efficiently.

Heterogenous network is a cellular network architecture where a mix of macro and small sites can be deployed. Figure 3.2 demonstrates the designed network model for the simulator. It includes two types of sites, that is a macrocell site and a small cell site. In addition, the cellular area is populated with a uniformly and randomly distributed users.

Each site is divided into three hexagonal sectors(cells). Both macro and small sites have the same structure (120-degree sectorization), but they could differ in terms of operating frequency band, cell radius (R) and the location on the cellular area. The distance between base stations of two macro sites is known as inter site distance (ISD) and is normally equal to three times of the macro cell radius ($ISD = 3 \times R$). Small sites have smaller cell radius, they could be placed at the edge of the macro cells and operate on same and/or higher frequency bands. The deployment of small cell sites depends on the geographical area, user density, traffic load, traffic service type, etc. Deploying small sites at the macro cell edge has several advantages, such as increasing the cellular network capacity by providing more radio resources, enhancing the user experience for the users who would experience bad link quality at the edge of the macro cell. It could also improve the fairness among the users in the cellular area, so that a lower complexity scheduling algorithm can be implemented at the base stations.

In [17], Small Cell Network White Paper (GSA Small Cell Network White Paper with input from Ericsson and Huawei) emphasizes that the enhanced performance of a heterogenous networks is very dependent on the planning of small cell deployments. Hence, a good understanding of how well the small cells

interact with the macro network is necessary to determine the overall user experience and cost. Three functions are required when building a network with integrated small cells:

- Optimizes end user experience – including application coverage and mobility.
- Enhances network operations including KPI measurements and network operations.
- Deploys and delivers a seamless network.

As mentioned above, the users are randomly and uniformly distributed. The position of each user (x,y coordinates) is used for example to calculate the path loss to the TRxP (Transmission reception point) for different environment scenarios such as 3GPP urban macro (Uma) and urban micro (UMi) path loss models. In a fully developed simulation tool the position could also be used when modeling beam directions, interference, mobility and handover situations. The knowledge of the user position enables us to estimate the users channel conditions. This information, together with the user profile, will allow us to also estimate the resource demands per user.

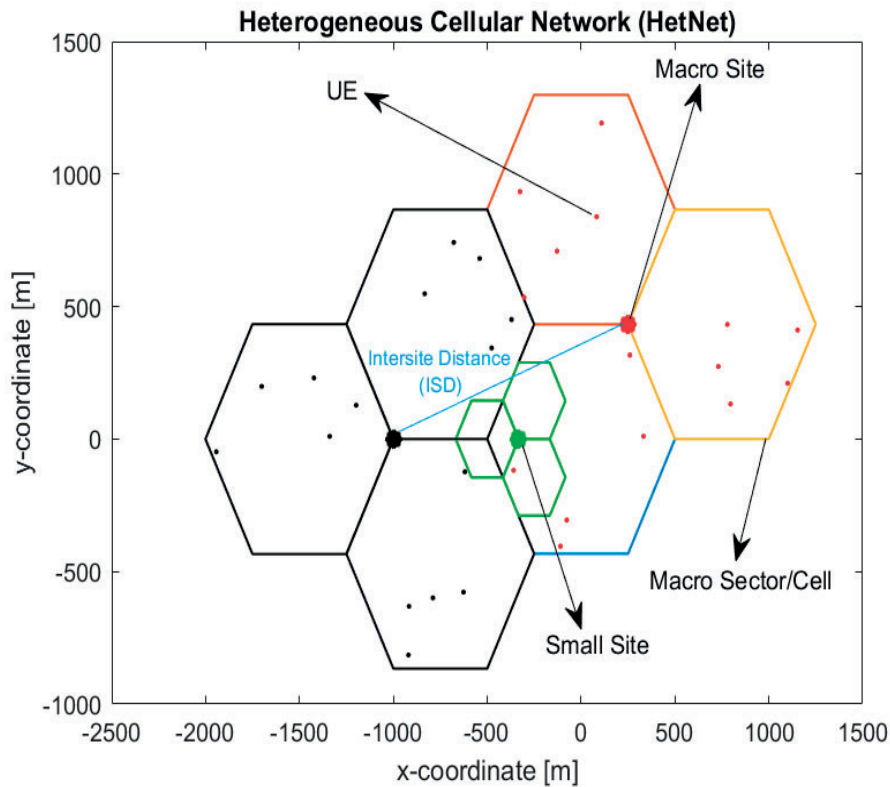


Figure 3.2: Basic heterogeneous cellular network architecture model

3.2.2 Cellular Environment

In this section, we give a brief overview of signal propagation mechanisms and environment scenarios, as this information could help in choosing the proper scenario and model for evaluating the performance of the simulator.

Radio signal characteristics can be affected differently depending on the environment where the cellular network is to be deployed. Wireless signals usually experience three propagation mechanisms, it can be reflected, diffracted or scattered. Due to these propagation mechanisms, multiple waves arrive at the receiver through multipath propagation experiencing large and rapid fluctuations, and sometimes also includes a direct Line-of-Sight (LOS) signal.

Reflection occurs when the electromagnetic wave hits an object with smooth surface and has large dimensions as compared to the wavelength, for example, surface of earth, buildings, walls, etc. However, Diffraction is also referred to bending or shadowing, it occurs when the radio signal path is obstructed by an object with large dimension relative to the wavelength and its surface has sharp (irregularities) edges. The propagated signals bend around the obstacles and reach their destinations; this mechanism allows radio signals to propagate in urban and rural environments without needing to a line of sight path condition. The third mechanism is scattering, it occurs when radio wave impinges upon medium having objects that are smaller or comparable to the wavelength, for example small objects such as street lights, sign boards and tree foliage.

There exist several models that are used to approximate the signal attenuation as a function of the distance between the transmitter and receiver, where other important factors such as environment (urban, rural, etc.), the height and location of antennas are usually considered. In our simulator, a standardized path loss model is used, it is specified by the 3GPP technical report (TR) TR 38.901 which includes 5G models for different scenarios and for frequencies from 0.5 to 100 GHz [22].

Urban macro (Uma) and Urban micro (UMi) environments have been chosen as main scenarios for the evaluation in the simulator, as they could be best suited for the two designed cells (macro and small cells) and 5G cellular network deployment. The models are defined in [24] as

$$PL_{UMa-NLOS} = 13.54 + 39.08 \log_{10}(d) + 20\log_{10}(f_c) - 0.6(h_{UT} - 1.5) \quad (3.1)$$

$$PL_{UMi-NLOS} = 22.4 + 35.3 \log_{10}(d) + 21.3\log_{10}(f_c) - 0.3(h_{UT} - 1.5) \quad (3.2)$$

Where, the carrier frequency f_c in GHz and the distance between the base station and the user terminal, d in m. h_{UT} refers to the height of the user terminal (UT) and is limited from 1.5 to 22.5 m from the ground. Equations

(3.1), (3.2) clearly show that increasing the distance or the carrier frequency, results in higher path losses.

3.2.3 General Traffic model

Users can have different service requirements and profiles. In our traffic model, a database for user services is created based on the provided information from Ericsson and is as shown in table 3.1. The service list is a combination of different types of traffic namely, voice, video streaming, browsing and so on. Services are assigned with unique service IDs and each service type is mainly characterized based on its burst size, burst interval, average rate, average session duration and session activity factor. Session activity factor is a variable that represents the session activity per hour (average session duration (s))/(3600 (s)). The burst size can be computed as follows,

$$\text{Burst size (bytes)} = \frac{\text{Average rate (bps)} * \text{Burst interval(s)}}{8} \quad (3.3)$$

This calculation is valid only for streaming and/or circuit switched services, that is, voice, video and audio streaming. Other services are assumed to be downloaded in big chunks and ends up in the buffer as a whole.

Service ID	Service type	Average rate (Mbps)	Burst size (bytes)	Burst interval (s)	Average session duration, in sec (for interactive services)	Session activity factor
1	Conversational voice (GBR)	0,064	160	0,02	120	3,3%
2	Video, buffered streaming (2min)	7	13 125 000	15	120	3,3%
3	Video, buffered streaming (10min)	7	13 125 000	15	600	16,7%
4	Video, buffered streaming (60min)	7	13 125 000	15	3600	100,0%
5	Music, buffered streaming (20min)	0,32	2 400 000	60	1200	33,3%
6	Music, buffered streaming (60min)	0,32	2 400 000	60	3600	100,0%
7	Web browsing	max	3 000 000	N/A	N/A	
8	Social, picture download	max	5 000 000	N/A	N/A	
9	File download (new software)	max	100 000 000	N/A	N/A	
10	File download (document/app SW)	max	10 000 000	N/A	N/A	

Table 3.1: Service list & their characteristics

Once the services are listed we assign them into 5 different profiles as shown in table 3.2. Each profile has a combination of services along with their frequency of occurrence per hour. We can observe from table 3.2 that the combination of services implies that profile 1 consists of services associated with office environments. Profile 4 reflects heavy file downloading users. Users are now assigned with the certain distribution of profiles randomly, for example, 20% of users are mapped to profile 1, 30% to profile 2, etc. Once the users have been assigned with the profiles, we can create their traffic pattern accordingly. While creating the traffic pattern, we have made few assumptions which are stated as follows: The start time of each session of a service is randomized and different sessions of same service type shall not overlap, for example a user cannot have a 20min session between 10-30min and another between 20-40min. Also, if a session occupies complete 100% of the simulation window then no other services of this type can be used simultaneously. Finally, if a service of 20min starts at 45min of a total simulation window of 60min real time, it is stopped by the end of simulation.

Profile ID	Services	Service ID	Number of times per hour (Service frequency)	Example	Data volume [MB]
1	Conversational voice (GBR)	1	4	Office person	159.84
	Music, buffered streaming (20min)	5	2		
	File download (document/app SW)	10	6		
2	Social, picture download	8	20	Social networking	1675
	Video, buffered streaming (2min)	2	15		
3	Video, buffered streaming (60min)	4	1	Lazy guy	3150
4	File download (document/app SW)	10	12	MBB classic (with upgrade)	220
	File download (new software)	9	1		
5	Music, buffered streaming (60min)	6	1	Social networking and music	249
	Social, picture download	8	15		
	Web browsing	7	10		

Table 3.2: User service profiles

The traffic pattern for each service is based on the number of frequencies per hour and this pattern is then expanded depending on bursts over entire session duration specifically for the interactive services as per table 3.3. Number of bursts for each session of interactive service is computed as per equation 3.4. This process of expanding the timeline based on bursts is performed only on the active users that have data to transmit at that particular time. The expanded timeline is later given as an input to the scheduling algorithms which is explained in detail in later sections.

$$\text{Number of Burst} = \frac{\text{Average Session duration (s)}}{\text{Burst interval (s)}} \quad (3.4)$$

Service ID	Interactive services	Bursts per session duration
1	Conversational voice (GBR) (2min)	6000
2	Video, buffered streaming (2min)	8
3	Video, buffered streaming (10min)	40
4	Video, buffered streaming (60min)	240
5	Music, buffered streaming (20min)	20
6	Music, buffered streaming (60min)	60

Table 3.3: Number of bursts per average session duration

3.3 Scheduling Algorithms

In this section, we considered two of the scheduling algorithms to obtain a comprehensive comparison of the performance of Round Robin and Maximum

Throughput with respect to fairness and best throughput respectively. Each of these scheduling algorithms are explained in detail below:

3.3.1 Round Robin Algorithm

Round robin algorithm is a channel unaware scheduling scheme which performs resource allocation based on three main parameters namely:

- Time quantum (TQ) – maximum available time for each service that is scheduled.
- Arrival time (AT) – time at which service arrives in ready queue.
- Burst time (BT) – time required by a service to complete its execution.

Graphical representation of the flowchart is shown below in figure 3.3.

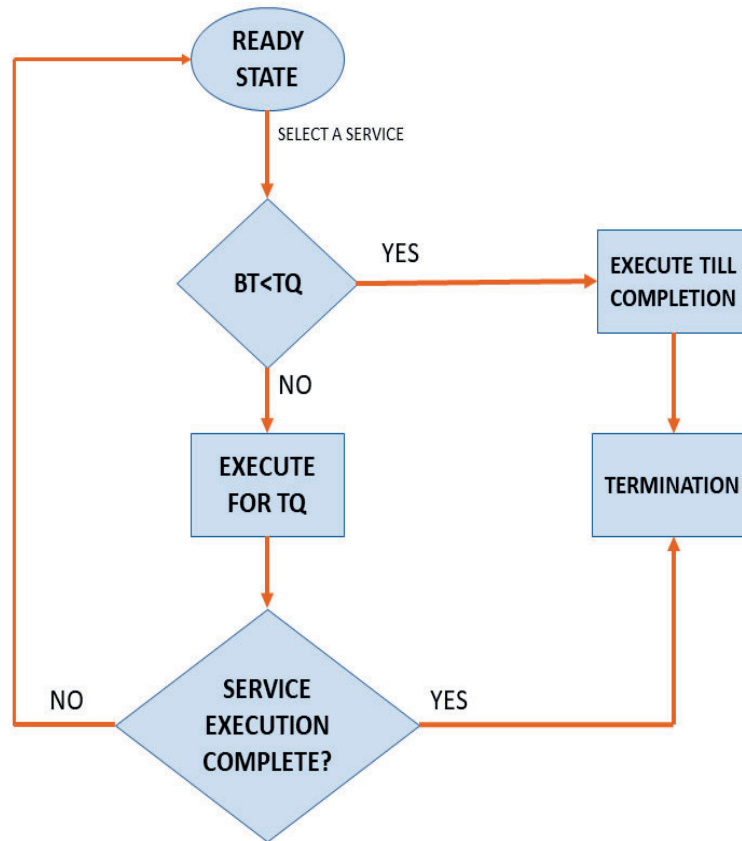


Figure 3.3: Flowchart for Round Robin Scheduling Algorithm

It can be observed from the flowchart that when a queue of expanded timeline for all the active users is obtained from the simulation engine, they are served based on their arrival time. Once the arrival time is sorted, the burst time of each

service is compared with the time quantum. If $BT < TQ$, then the service is executed till completion without utilizing entire TQ buffer. On the other hand, if $BT > TQ$, then the service is executed only till TQ and is placed on the queue again until it is served completely. The Burst sizes corresponding to the served queue is fetched and is further expanded based on number of physical resource blocks (PRBs) each session requires. This is calculated as per the equations 3.5 and 3.6 shown below,

$$PRB_{bits} = (1 - OH) * N_{sc} * N_{symbols} * k * rc * TDD_{factor} \quad (3.5)$$

$$Number\ of\ PRBs = \frac{Burst\ size\ (bits)}{PRB_{bits}} \quad (3.6)$$

Where, in equation 3.5, OH is the overhead for control information, reference channels signaling etc. in 5G NR, where 24% is assumed, N_{sc} is the number of subcarriers per subchannel which is equal to 12, $N_{symbols}$ is the number of OFDM symbols in a timeslot (subframe) and is equal to 14, k is the number bits per symbol based on the type of modulation scheme which is assumed to be 16QAM, rc is the corresponding coding rate for a CQI value of 6, TDD_{factor} is about 74.29% for the downlink share. This calculation is valid for FDD as well by substituting $TDD_{factor} = 1$.

The implemented Round Robin scheduling algorithm (time slicing scheduling) is based on two criteria, that is, time quanta (TQ) and arrival time (AT). To illustrate the functionality of the algorithm and the importance of selecting a proper time quantum value, two examples are given below. In the two examples turnaround time (TAT) and waiting time (WT) are calculated. Turnaround time for a specific service correspond to the time between the arrival time and the completion time (CT). Waiting time (WT) is how much time a service spends in the ready queue waiting for its turn to be scheduled.

Example 1:

This example assumes six services with corresponding arrival and burst times as shown in table 3.4. The bursts are scheduled with maximum limits equal to two, that is, $TQ = 2$. Service one (S1) has the lowest arrival time and hence it is the first service to be scheduled. Round Robin scheduler compares S1 burst time ($BT=1$) with the time quantum ($TQ=2$); Since the arrival time of S1 and the burst time is lower than the time quantum, it runs till 1 (completion). Then, at time 1 Round Robin algorithm looks up in the table if there are services which have arrived meanwhile. It can be seen that, service two (S2) has arrived at time 1,

therefore this service is put in the queue (ready state) in order to be scheduled. S2 has burst time unit of 3 which is greater than the TQ, so the scheduler will run until maximum limit of TQ and there will be one unit as a remaining time (RT) for S2. Before entering the RT for S2 in the queue, the scheduler will be at time 3 (see figure 3.3 below), so now it searches if there are any new services that have arrived at time 3 or earlier and put them in the queue. We notice that service three (S3) and service four (S4) have arrived, so they will be queued first before entering remaining time of S2. In the same manner the process continues until all the services are scheduled.

Service	AT	BT	RT	CT	TAT (CT – AT)	WT (TAT – BT)
1	0	1	0	1	1	0
2	1	3	1 0	8	7	4
3	2	5	3 1 0	19	17	12
4	3	2	0	7	4	2
5	4	4	2 0	16	12	8
6	5	6	4 2 0	21	16	10

Table 3.4: Example of Round Robin scheduling with main design factors - time quantum ($TQ = 2$)

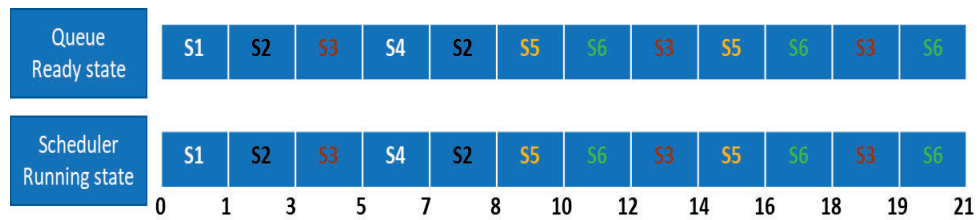


Figure 3.4: Illustration of Round Robin queuing and scheduling process - time quantum ($TQ = 2$)

Example 2:

This example follows the same procedure as in example 1 with only difference being another time quantum to investigate and show the changes in

performance parameters. In this example, a larger time quantum value is chosen (TQ=4).

Service	AT	BT	RT	CT	TAT (CT – AT)	WT (TAT – BT)
1	0	1	0	1	1	0
2	1	3	0	4	3	0
3	2	5	1 0	19	17	12
4	3	2	0	10	7	5
5	4	4	0	14	10	6
6	5	6	2 0	21	16	10

Table 3.5: Example of Round Robin scheduling with main design factors - time quantum (TQ = 4)

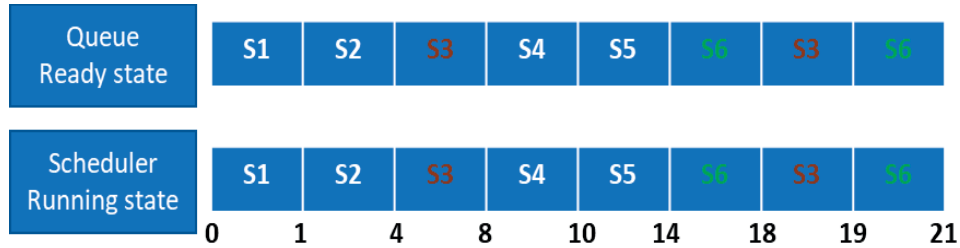


Figure 3.5: Illustration of Round Robin queuing and scheduling process - time quantum (TQ = 4)

In order to answer the question of how we choose a proper time quantum value we described two examples with different time quantum and their performance. We can see that, examples 1 and 2 clearly shows the impact of increasing or decreasing time quantum. Choosing a proper time quantum value is a key design parameter. Round Robin scheduling in example 1 (TQ=2) requires more context switching compared with example 2 (TQ=4) and this leads to undesired overhead for implementation in real systems. However, choosing a larger time quantum value, reduces the context switching overhead and minimizes the turnaround time (TAT) (average TAT in example 2 is 9 whereas it is 9.5 in example 1), but if the value is too large Round Robin algorithm will behave like First come first serve (FCFS) scheduling algorithm which simply schedules the

jobs according to only their arrival time. As a rule of thumb, 80% of bursts should be shorter than time quantum as per [25]. Therefore,

time quantum should be chosen in such a way that overhead is reasonable and also fairness is maintained by avoiding the scheduler from behaving like FCFS.

Hence, the principal advantage of Round Robin algorithm is that fairness is guaranteed with all the services being served equally regardless of the channel condition. Thus, this scheme results in a lower network throughput since it ignores the channel quality information [11].

3.3.2 Maximum Throughput Algorithm

Maximum throughput algorithm is a channel aware scheduling scheme which performs resource allocation based on Channel quality indicator (CQI). A higher CQI value represents good channel condition and vice-versa. The flowchart for this scheduling scheme is as shown in the figure 3.6.

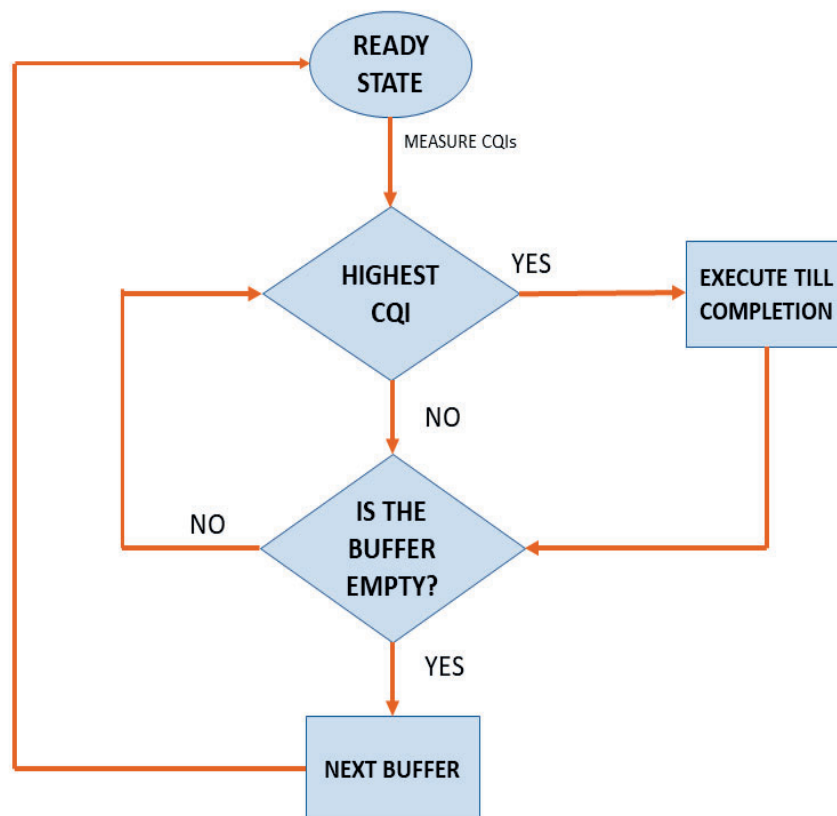


Figure 3.6: Flowchart for Maximum Throughput Scheduling Algorithm

The expanded timeline is obtained from the traffic model and each burst is now mapped to a CQI value based on the signal-to-noise ratio of the users. The SNR is computed using Link budget function and pathloss. With the following parameters, we computed the Maximum isotropic loss (MIL) from the link budget and found the difference between this obtained attenuation and the measured pathloss (equation 3.1) of every user to get the measured SNR values. Equation for calculating MIL using below mentioned parameters is as follows,

$$DL_{MIL} = P_{OUT} + G_{BS} - NF_{UE} + G_{UE} - T_{NP} \quad (3.7)$$

- Output power (P_{OUT}) = 27.5 dBm
- Base station antenna gain (G_{BS}) = 24.5 dBi
- UE antenna gain (G_{UE}) = 11 dBi
- UE noise factor (NF_{UE}) = 13.8 dB
- Thermal noise power (T_{NP}) = $-174 + (10 * \log_{10} BW)$ dBm

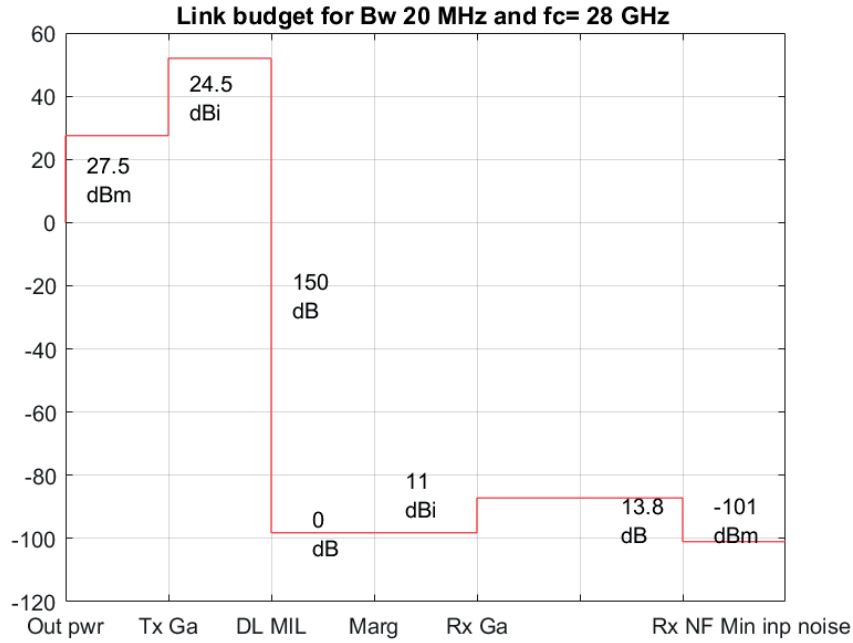


Figure 3.7: Link Budget

Based on the SNR, the coding gain is extracted according to the [37] and calibration estimates of the SNR/Coding rate relation from Ericsson. As we didn't add any interference calculations based on traffic from neighboring cells in the

simulator, an arbitrary offset of ~ 6 dB was subtracted from the SNR given by the difference between the MIL and the pathloss. This offset mimic interference just to get some more interesting variations of the SNR estimates when evaluating the functionality in the simulator.

We have assumed the range of expected SNR values between -19 dB to +25 dB and this is mapped to the CQI values as shown in the table 3.6. This table is extracted from 3GPP standard document [19], expect for the fifth column which shows SNR mapping.

CQI index	modulation	Code rate X 1024	efficiency	Assumed SNR in dB
0	out of range			
1	QPSK	78	0.1523	-19 to -17
2	QPSK	193	0.3770	-16 to -14
3	QPSK	449	0.8770	-13 to -11
4	16QAM	378	1.4766	-10 to -8
5	16QAM	490	1.9141	-7 to -5
6	16QAM	616	2.4063	-4 to -2
7	64QAM	466	2.7305	-1 to +1
8	64QAM	567	3.3223	+2 to +4
9	64QAM	666	3.9023	+5 to +7
10	64QAM	772	4.5234	+8 to +10
11	64QAM	873	5.1152	+11 to +13
12	256QAM	711	5.5547	+14 to +16
13	256QAM	797	6.2266	+17 to +19
14	256QAM	885	6.9141	+20 to +22
15	256QAM	948	7.4063	+23 to +25

Table 3.6: 3GPP CQI Table [19]

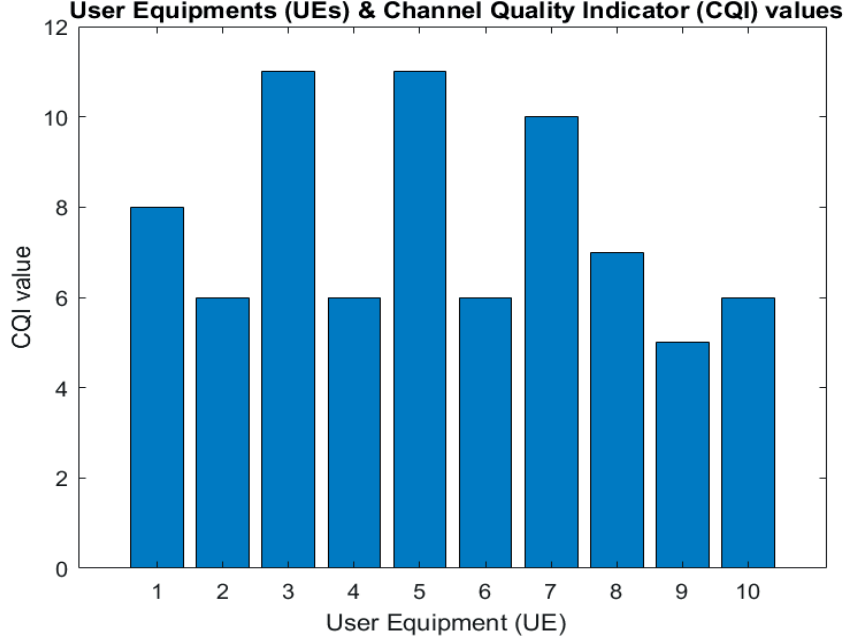


Figure 3.8: Link quality for 10 User Equipment

Figure 3.8 shows the results of simulating 10 users in a cell. The channel quality for every UE is obtained based on the mapping as explained above. UE 3 and 5 have the best channel condition with a CQI value of 11 which can be interpreted to be close to the base station or having good LOS conditions. Whereas, UE 9 with CQI value of 5 could be at the cell edge experiencing bad channel conditions.

The expanded time line in terms of bursts from the traffic model is now assigned with the CQI values and their corresponding UEs and service IDs are also mapped. Scheduler algorithm repeatedly checks for maximum CQI in specified interval of time until the timeline is empty. Once the maximum value is determined, the corresponding burst is placed in a buffer which is further expanded with the number of PRBs each service session requires using equations 3.5 and 3.6. One main difference in computing number of PRBs for maximum throughput is that the number of bits per symbol and code rate varies for every CQI accordingly.

3.3.3 Maximum PRB calculation

Maximum number of PRBs per TTI for both the scheduling algorithms is based on selected bandwidth and the type of radio access technology, that is LTE or 5G NR. If a burst requires PRBs more than Maximum available resource blocks, it will be served in the successive TTIs until it is completely executed.

$$BW_{utilised} = S_{utilisation} * BW \quad (3.8)$$

$$N_{sc-tot} = \frac{BW_{utilised}}{SCS} \quad (3.9)$$

$$PRB_{max} = \frac{N_{sc-tot}}{N_{sc}} \quad (3.10)$$

Where,

$S_{utilisation}$ = spectrum utilization factor (LTE = 90%, 5G-NR = 99% [28])

$BW_{utilised}$ = Bandwidth utilized

BW = system bandwidth chosen

N_{sc-tot} = total number of subcarriers

SCS = sub-carrier spacing (LTE =15 kHz, 5G-NR = 15 kHz, 30 kHz, 60 kHz, 120 kHz)

PRB_{max} = maximum number of physical resource blocks

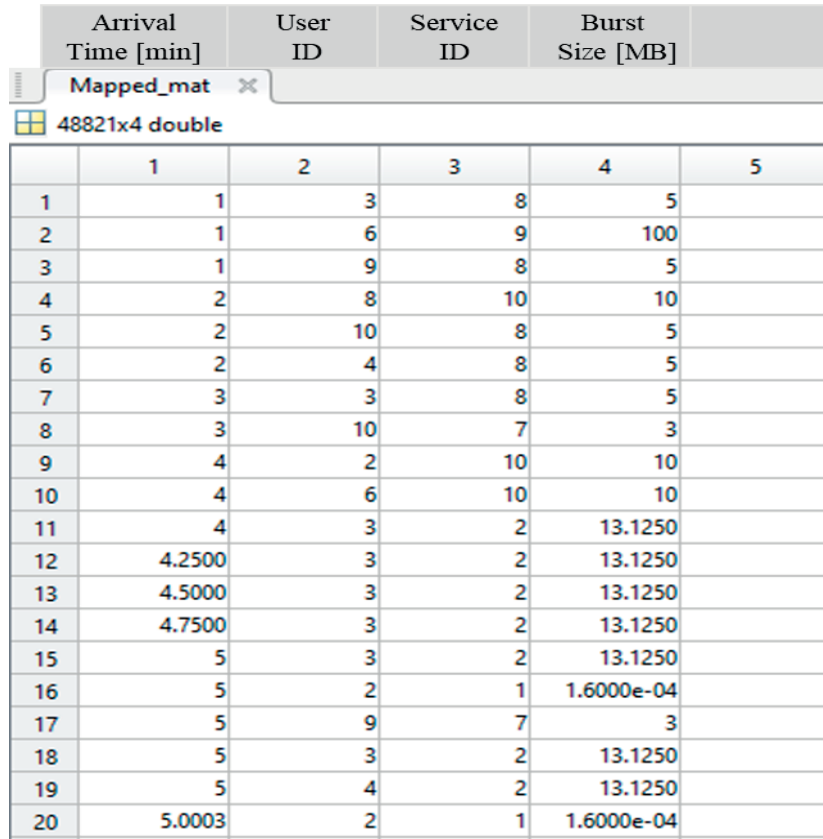
N_{sc} = number of subcarriers per subchannel equal to 12

4. Results

This section is mainly devoted for visualizing the results obtained from the designed simulator. The parameters and the models used in the simulator are based on 3GPP standards and the information provided from Ericsson.

4.1 Evaluation of Traffic Model

The traffic model is designed and implemented as explained in detail in section 3.1. We have generated arrival times for each user service randomly. To create the timeline for all the UEs, we arrange the arrival times in ascending order. Now, for each value of arrival time, based on corresponding User ID, Service ID, burst interval and session duration, the timeline is expanded as shown in the figure 4.1 below,



	Arrival Time [min]	User ID	Service ID	Burst Size [MB]	
	Mapped_mat				
	48821x4 double				
	1	2	3	4	5
1	1	3	8	5	
2	1	6	9	100	
3	1	9	8	5	
4	2	8	10	10	
5	2	10	8	5	
6	2	4	8	5	
7	3	3	8	5	
8	3	10	7	3	
9	4	2	10	10	
10	4	6	10	10	
11	4	3	2	13.1250	
12	4.2500	3	2	13.1250	
13	4.5000	3	2	13.1250	
14	4.7500	3	2	13.1250	
15	5	3	2	13.1250	
16	5	2	1	1.6000e-04	
17	5	9	7	3	
18	5	3	2	13.1250	
19	5	4	2	13.1250	
20	5.0003	2	1	1.6000e-04	

Figure 4.1: Illustration of mapping in the timeline for 10 UEs

In the above shown screenshot, first column refers to arrival times in minutes, second is user IDs, third is service IDs and the last column is their burst sizes in

MB. Each row represents an arranged arrival time value with its corresponding user ID, service ID and burst size. In order to explain how the burst intervals are generated, let us consider an example from the table. From row 11, UE 3 with service 2 has a start time of 4. We know from table 3.1, the burst interval for this service is 15 seconds with a total session duration of 2 minutes. Therefore, this service is repeated for every 15 second (0.25 minutes) interval which can be observed in the figure. This service consists of 8 bursts over one complete session (table 3.3).

[Note: not all the 8 bursts for this particular example is shown in the screenshot as there are arrival times of other services which occurs in-between.]

In order to show the behavior of traffic and active users, we take the snapshot from the entire timeline. The snapshot consists of traffic pattern of 10 UEs with different service profiles. Based on their arrival time, the burst is repeated over the entire session duration at the rate of respective burst intervals. This spread of the bursts over the timeline is plotted and is shown below in figure 4.2.

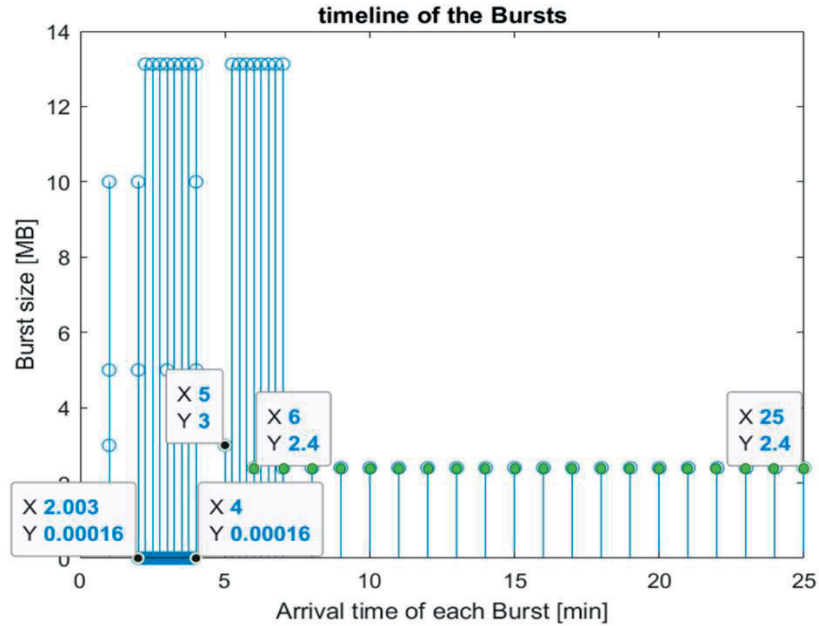


Figure 4.2: Illustration of snapshot of the timeline for 10 UEs

From the figure shown above, we can observe the distribution of bursts for different services of 10 UEs. For example, let us consider service ID 5 (S5 marked

in green) which is Music, buffered streaming with a session duration of 20 minutes, burst size of 2.4MB and burst interval of 60 seconds. The first burst for this service occurs at 6th minute and it repeats once every 60 seconds for 20 times till 25th minute as indicated in the figure above. Here, 20 times refers to the number of bursts over entire session duration of the service as illustrated in table 3.3.

Similarly, for conversational voice service with service ID 1 (S1), we can see 6000 repetitions from 2.003th minute till 4th minute. This is because S1 has very small burst interval of 20 milliseconds over a 2-minute session duration and the burst size is also very small (160 Bytes). In general, voice traffic is much more variable than data traffic.

Every interactive service is repeated in the same manner with their respective characteristics referred from tables 3.1 to 3.3. Whereas, the file downloading services that arrive in big chunks are considered as a whole without any repetitions. This can be seen at 5th minute which is a web browsing service with service ID 7 and burst size of 3MB.

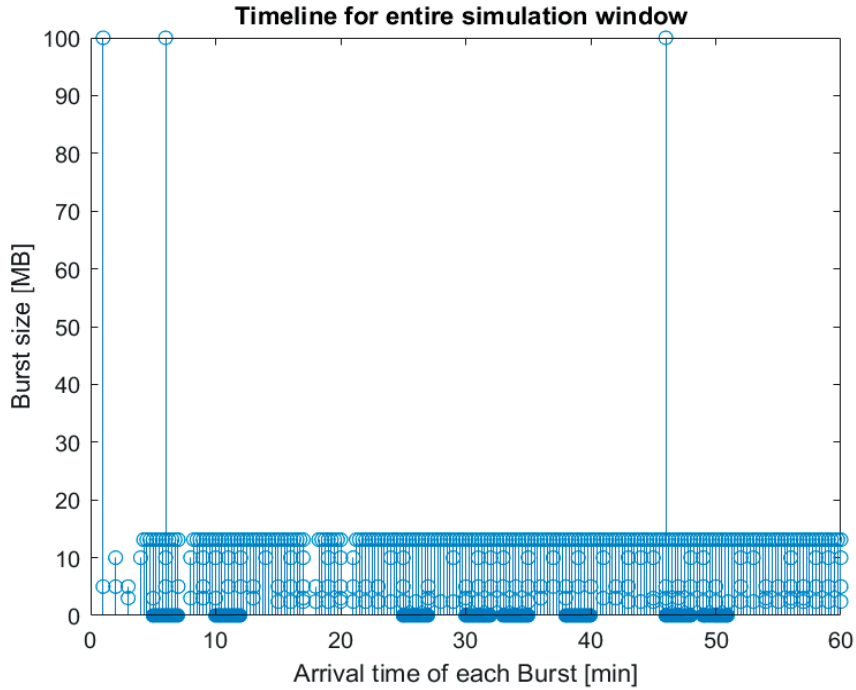


Figure 4.3: Illustration of timeline for entire simulation window for 10 UEs

Figure 4.3 shows the distribution of the start times of all the services over a simulation window of 60 minutes for 10 UEs. As we can see that this simulation window has heavy traffic load. Based on the number of bursts, sometimes, the

repetitions exceed 60 minutes. The arrival times occurring after simulation are blocked as per the assumptions made in section 3.1.3. We can also observe overlapping of different service types at same arrival time, but it is made sure that, no two same kind of services arrive at the same time for one UE.

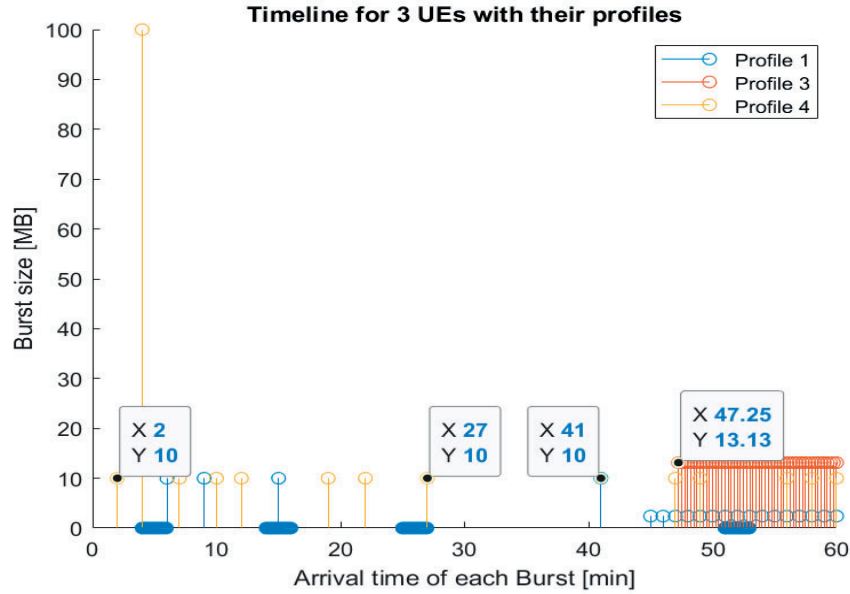


Figure 4.4: Visualization of the traffic for 3 UEs

The traffic timeline for 3 different UEs having different profiles 1,3 and 4 is as shown in the figure 4.4. The description of the profiles is as explained in table 3.2. From this figure, we can easily identify the intervals where the UEs are active or silent. For example, in the interval between 2nd minute and 27th minute two profiles, that is, 1 and 4 are active and we can observe that they overlap at certain instances. But between 27th minute till 41st minute, all the UEs are inactive. At the end of the simulation window, above 47th minute, all the 3 UEs are active with very small silent periods in between.

The observations made in this sub-section serves the purpose of analyzing a defined network for different user densities and traffic models. This analysis gives us information which can improve the system performance. By keeping track of all the parameters involved, we can expect the behavior of the simulator while allocating resources.

4.2 Evaluation of the Scheduler

The simulator for spectrum allocation explaining effects of different scheduler algorithms on the UEs is designed and executed. This sub-section is further divided into three main parts where the variation in actual resource demand among the UEs, average and peak cell bit rates and processing time of the simulator for varying number of UEs are discussed in detail.

4.2.1 Actual resource demand

Actual resource demand is the amount of resources that a user requires based on the type of traffic and its channel conditions. We can express this demand in terms of number of resource blocks irrespective of the bandwidth or both bandwidth and number of TTIs required.

In order to determine actual resource demand of a UE, we consider an example of one UE with Mobile broadband classic profile 4.

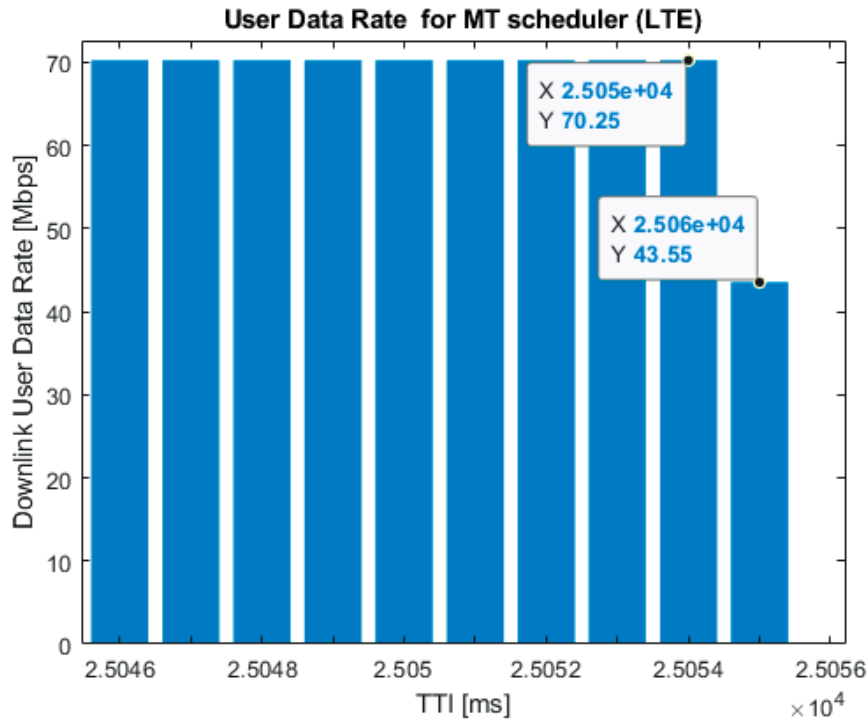


Figure 4.5: Data rate for one UE with profile 4 for LTE

This example considers LTE radio access technology, TDD system and bandwidth of 20 MHz with standard subcarrier spacing of 15 kHz. The UE has average data rate of 70.24 Mbps over a period of 25055 milliseconds as shown in the figure 4.5. We can observe that this UE has very high data rate as it has a very good channel condition (CQI=15) and hence it is served with lower number of resources comparably.

Theoretical calculations: From tables 3.1 to 3.3, this UE with profile 4 has a combination of services 10 and 9 (S10, S9). S10 is a file download service of 10 MB which is occurring 12 times per hour and S9 is also a file download service of 100 MB occurring only once per hour.

$$Total\ data = (10 \times 12) + 100 = 220\ MB$$

Practical calculations: From the simulation, the total data scheduled is computed as follows,

$$Scheduling\ duration = \frac{25055}{1000} = 25.055\ seconds$$

$$Data\ rate = \frac{70.24}{8} = 8.78\ MBps$$

$$Scheduled\ data = Scheduling\ duration \times Data\ rate = 219.98\ MB$$

From the above shown theoretical and practical calculations, it is clear that the scheduler matches the user's demand and schedules entire data as requested. Now, we compare the simulation explained above for same UE with same profile services but, for 5G-NR RAT.

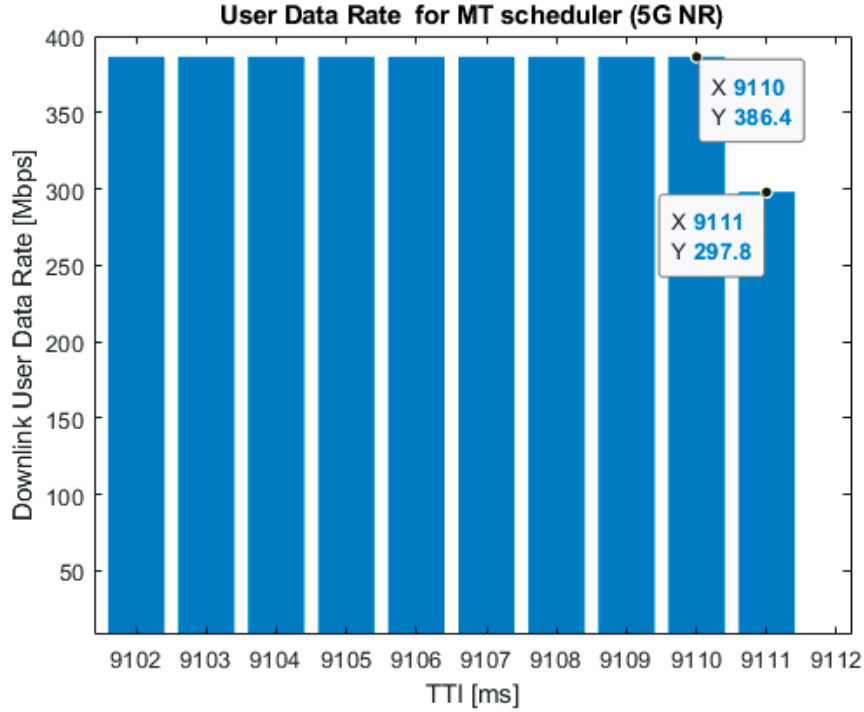


Figure 4.6: Data rate for one UE with profile 4 for 5G-NR

In this simulation example, the bandwidth of the system is chosen to be 100 MHz with subcarrier spacing of 30 kHz. Transmission time interval (TTI) will be equal to 0.5 millisecond and the maximum number of PRBs available per TTI is 275. Theoretical calculation is same as that for previous example (figure 4.5).

Practical calculations: From the simulation shown in figure 4.6, the total data scheduled is computed as follows,

$$\text{Scheduling duration} = \frac{9111 \times 0.5}{1000} = 4.555 \text{ seconds}$$

$$\text{Data rate} = \frac{386.39}{8} = 48.29 \text{ MBps}$$

$$\text{Scheduled data} = \text{Scheduling duration} \times \text{Data rate} = 219.96 \text{ MB}$$

From the above examples where the channel quality is the same, it can as expected be seen that the RAT with highest bandwidth, (in this case NR), manage

the download of the same file size largeBW/smallBW times faster. The average data rate of 5G-NR system is very high, and this is mainly attributed to the higher bandwidth. Therefore, we can state that this simulator computes the actual resource demand for different traffic scenarios, different type of RAT and adaptive channel qualities in an efficient way.

4.2.2 Cell data rate

We can calculate the average cell rate of the traffic model as the sum of data rate of all the scheduled user services divided by the total number of utilized TTIs. In addition, peak data rate which is the maximum experienced data rate over utilized TTIs, can be easily determined.

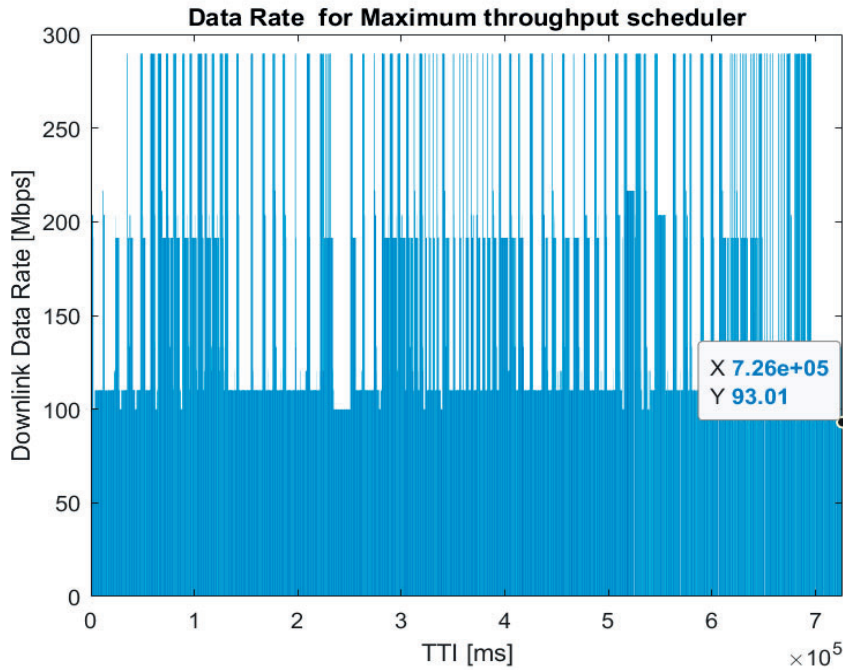


Figure 4.7: Data rate for 10 UEs for entire timeline (MT)

Figure 4.7 represents the obtained data rate for all the 10 UEs over one simulation. Using these values, average and peak data rate of a cell can be determined as follows,

$$\text{Average cell data rate} = \frac{\text{sum}(\text{data rates per TTI})}{\text{total number of TTIs}}$$

The system specifications for the above shown figure 4.7 are,

- RAT = 5G NR
- System = TDD
- Bandwidth = 100 MHz
- ISD = 1000 m
- Carrier frequency = 28 GHz
- Scheduler algorithm = Maximum throughput
- Pathloss model = Urban macro NLOS 3GPP model

Using these specifications, after the simulation, obtained CQI values and modulation schemes are:

- CQI values = 9, 10, 11, 13
- Modulation schemes = 16 QAM, 64 QAM

Now, based on the data rate values obtained for a total of 7.26×10^5 TTIs as indicated in the figure above, the average cell data rate is about ~175 Mbps. And the peak data rate is observed to be approximately 289 Mbps.

Similarly, the system specifications for Round robin algorithm is same as Maximum throughput except for the scheduling scheme, channel quality (CQI = 6) corresponding to modulation scheme of 16 QAM, coding rate of 616/1024 and time quantum TQ = 228.24.

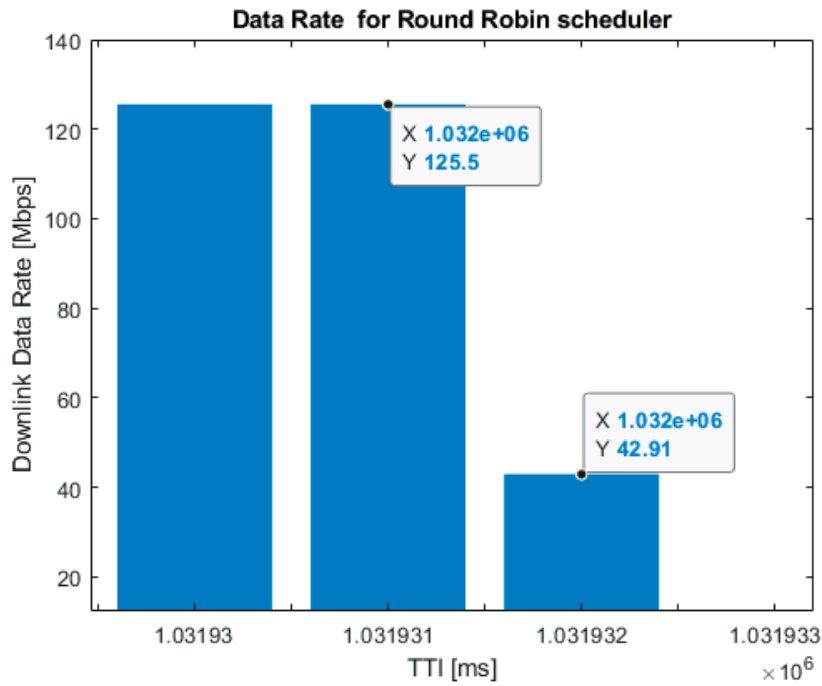


Figure 4.8: Data rate for 10 UEs for entire timeline (RR)

From the figure 4.8, we can observe that every TTI corresponds to equal data rate except for the last where, the TTI is not completely utilized as all the active services in the queue are completely served by this point in time. The average cell data rate is computed to be approximately 125 Mbps and since Round robin treats every user service with same channel quality, the peak data rate will be same as average data rate.

The simulation results discussed above can illustrate how the users are served in terms of either fairness or higher throughput. There is always a tradeoff between these factors, the suitable scheduling algorithm can therefore be selected based on the user's main service requirements, that is, some users may require not to wait for a longer time to be served and higher data rate is not their main service criteria. On the other hand, attaining maximum throughput can be the main demand for other users.

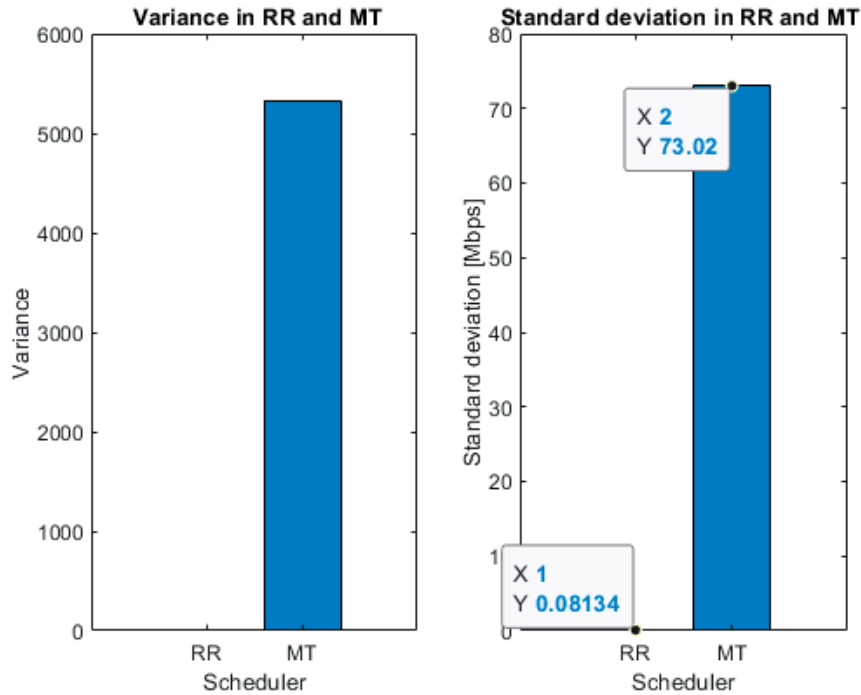


Figure 4.9: Variance and standard deviation for RR and MT schedulers.

Figure 4.9 demonstrates the variance and standard deviation for RR and MT schedulers that are shown in figure 4.7 and 4.8. It clearly shows that MT has much higher variance compared to RR scheduler as the data set in RR is much less-dispersed. The standard deviation for MT scheduler indicates that the data points are 73.02 Mbps far away from the mean (175.8 Mbps), on average. Whereas, in RR is about 0.08 Mbps indicating that it is very close to the mean (125.5 Mbps).

4.2.3 Processing time of simulator

To evaluate the performance of the simulator in terms of processing time, an adequate number of simulation runs are performed. The simulation is performed using laptop with the following main software and hardware specifications:

- MATLAB R2018b
- Windows 10 Enterprise 64-bit
- Intel CORE i5 vPro 8th Gen processor
- 16 GB memory

Equation 4.1 [33] is used to calculate the true average processing time with 95% confidence level.

Population mean = Sample mean \pm sample error

$$\mu = \bar{x} \pm \frac{t \cdot s}{\sqrt{n}} \quad (4.1)$$

Where,

μ : True mean (population mean)

\bar{x} : sample mean

t : T- statistics (t-table is dependent on degrees of freedom, ($df = n - 1$)

and the desired confidence value, that is, 90%, 95%, 98% ,99%, etc.)

s : standard deviation of the samples

T-statistics is considered (rather than Z-statistics) for all simulation tests as it is applicable, that is, following conditions are applied:

1. The population standard deviation (σ) is unknown.
2. Sample size is less than 30.

Simulation scenarios

Test 1

In this test, 10 simulations are run for 10 users with selected profile patterns. The total amount of simulated data is about 7.97 GB. The processing time of the simulation is measured in minutes and is recorded to calculate the confidence interval.

Test 1 simulation times:

[62.1000, 64.1324, 69.5167, 81.0333, 81.5207, 49.0569, 54.1700, 53.8425, 60.2500, 63.0423]

These values result in a confidence interval with a lower bound of 56.0698 and upper bound of 71.6630 minutes and a sample mean of 63.87 minutes.

Confidence Interval = [56.0698 71.6630]

Therefore, we can say that with 95% confidence the population (true) mean of the processing time is between 56.0698 and 71.6630. The Interval length is 15.5932 minutes.

Test 2

In this test, the same number of users and user profile settings in test 1 are used but the purpose here is to evaluate the performance when all users have a relatively good channel quality. To ensure High CQI values, a smaller ISD value is set and 10 simulations are performed. Simulation times are recorded and their values in minutes are,

[22.2617, 26.6033, 32.2008, 34.5333, 25.8658, 26.1390, 25.9256, 25.9805, 18.7634, 25.8859];

By calculating the true mean with 95% confidence level, we get

Confidence Interval = [23.2498 29.5820] and the sample mean is 26.42 minutes.

In this test, we can observe that the confidence interval is narrower (6.3322 min) compared with the pervious test, that is, the variability of the results relative to the sample size is small.

It was observed that variations of the processing time are dependent on the ISD value, higher the ISD higher the processing time as UEs will have lower CQI values which necessitates more resources to serve them and hence more processing time.

Test 3

In this test, 10 simulations are run for 5 users having distinct profile patterns. The total amount of simulated data is about 5.45 GB

The processing time of the simulation is measured in minutes and is recorded to calculate the confidence interval.

Test 3 simulation times:

[30.8368, 57.6602, 48.4543, 34.4728, 25.9516, 47.1958, 68.6037, 56.1945, 29.6989, 37.9833].

By calculating the true mean with 95% confidence level, we get

Confidence Interval = [33.5875 53.8228] and the sample mean is 43.71 minutes.

Test 4

A higher number of users having double data amount of test 1 is used in this test. 10 simulations are run for 20 users with the same profile distribution setting as in test 1. Therefore, the total amount of simulated data is about 15.95 GB (7.97×2)

The processing time of the simulation is measured in minutes and is recorded to calculate the confidence interval.

Test 4 simulation times:

[112.1305, 173.5509, 168.1416, 123.9414, 204.0189, 166.3845, 152.0104, 161.3598, 162.9790, 171.4614].

By calculating the true mean with 95% confidence level, we get

Confidence Interval = [141.0899 178.1057] and the sample mean is 159.6 minutes.

Test 5

Test 4 is repeated but when all the users have relatively high channel qualities.

Test 5 simulation times,

[45.1536, 47.5363, 50.2462, 54.6129, 56.5520, 57.2698, 57.6487, 59.4313, 64.0501, 66.5397]

Confidence Interval = [51.0375 60.7705] and the sample mean is 55.9 minutes.

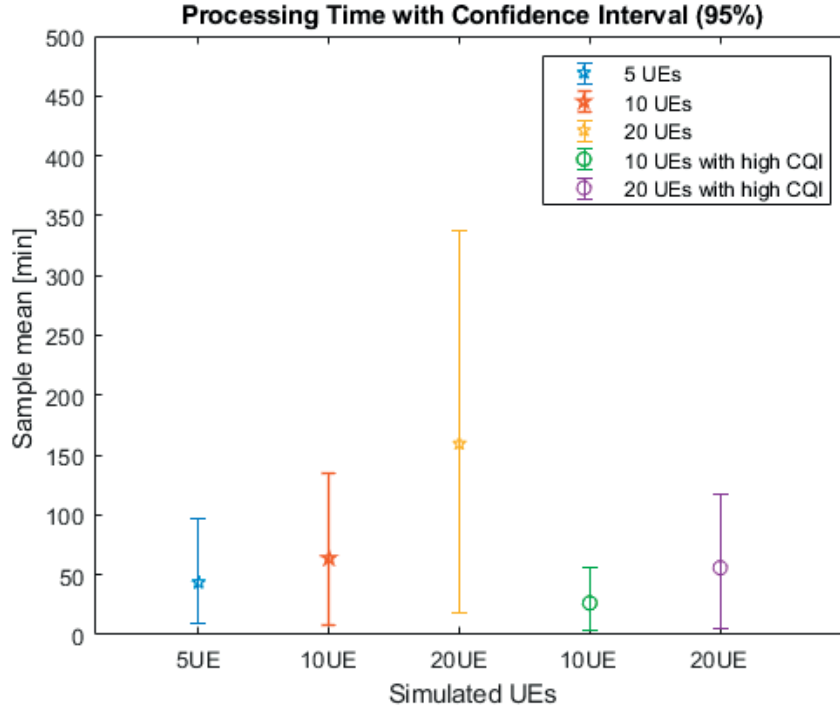


Figure 4.10: Confidence Interval for test 1,2,3,4 and 5.

The results obtained from all the simulation scenarios are visualized in figure 4.10. As we can see that, the average processing time for 10 users with relatively high CQI values is about 26 minutes (green line on graph). And when we consider twice number of users for the same profile settings and ISD value, the average processing time is around 55 minutes (purple line on graph). This shows that the performance of the simulator tends to be linear when the simulation parameters are carefully set and can be clearly observed with the obtained results.

The overall conclusion of the results is explained in detail in chapter 5.

5. Conclusions

Creating an efficient MATLAB simulator framework for analyzing the behavior of the traffic and considering different traffic models in the upcoming 5G technology is a challenging task of vital importance. The ability to accurately dimensioning the traffic and obtaining the actual user demands, can save many resources and hence, leading to a positive impact both on the mobile network operators (reducing costs) and on the users by having better service experience.

We conclude our thesis work by comparing the project aims and contributions with the achieved results. The primary intent of this thesis is to design a simulator that adapts to different channel conditions, different types of cellular structure, largely varying number of users and different categories of services with wide range of characteristics. Furthermore, the implementational and computational complexity of the simulator has to be as low as possible resulting in a useful simulation tool.

The results of this thesis work is presented in 3 main sections, that is, *Evaluation of Traffic Model*, *Evaluation of the Scheduler*, and *Processing time of simulator*. To clarify the obtained results, simulation and performance evaluation results are visualized. Previous studies in this field are mostly theoretical analysis or just considers a particular type of traffic. In this thesis, actual implementation of the traffic model and scheduling schemes are performed.

By reviewing these results, we can show that the simulator meets the main goals. The simulation results are accurate when compared to the input values in terms of amount of data and data rate. Resource allocation can be performed using many algorithms, each of them having their own pros and cons. For simplicity, we have implemented two of the algorithms, Round Robin and Maximum throughput scheduling schemes. From these algorithms, we can state that, Round Robin performs scheduling giving fair time sharing to all the user services. Whereas, maximum throughput gives importance to channel quality indicator, better the channel condition of the user service, faster and more efficient the resource allocation to that user service. Processing time depends on several factors, the result shows that when the users with high CQI and specified selected traffic patterns are doubled, the simulation time is also doubled. Consequently, the processing time can be predicted as it tends to have more linear behavior when users have equal link qualities and the simulator is properly set.

Finally, it is worth emphasizing that the main goal as set by Ericsson for this master thesis of creating the first framework of the simulator is achieved. However, the finalized version of the simulator is estimated to be completed in approximately two more future thesis works. The details about the targeted capabilities of the finalized version are provided in future work chapter.

6. Future work

In this section, several important aspects are highlighted to be considered for future work and development.

This Master Thesis has created the fundamental basic parts of the simulation tool such as the flexible traffic model for each user, the definition of the area covered by heterogeneous cells, the positioning of the users in the cell area, a basic radio model to reflect the path loss of the radio signals due to the distance between the transmitter and receiver, and finally the fundamental allocation of the radio resources for the created traffic performed in the scheduler. Furthermore, the execution time of the simulator have been measured.

In this thesis project, we designed a simplified traffic model in which users are randomly deployed in a single macro cell of a heterogeneous network and we have assumed that they do not cause any interference. In future, more advanced models can be considered, where the users populated either in one site or in multiple cell sites cause interference. Several methods can be helpful to reduce interference levels, for example beamforming techniques, utilizing the high frequency range (FR 2) available in 5G NR as it is less congested, etc.

In the scheduler part, we have considered two basic algorithms which either provide good fairness or maximum throughput. But implementing an optimal scheduling scheme that can balance fairness and throughput will be interesting. We have seen many studies that investigate all the possible combinations of scheduling schemes that can be optimal. For example, implementing proportional fair joint method algorithms. This method utilizes the strategy of a scheme that is channel unaware and also channel aware. Using these strategies together can lead to balance between throughput and fairness as required by many mobile network operators.

Overall, the target of a “final version” of the simulation tool is to reflect the dynamics of mobility situations with the time it takes to do handover, consider the dynamic interference situation due to the relative position between UE's and its respective traffic variations. The simulation tool shall also reflect the schedulers work to allocate the radio resources between the UE's and the simulation tool shall be able to reflect the latency in backhaul and fronthaul transport parts of the radio network. The aim is to create a simulator tool like this is to be capable to complement other already widely used simulators to reflect new situations that are expected to occur for 5G with e.g. mixes of new services with different priority and latency demands and/or to evaluate new HO methods.

References

- [1] “Cisco Visual Networking Index: *Global Mobile Data Traffic Forecast Update, 2016-2021*,” Tech. Rep., February 2017
- [2] “Cisco Visual Networking Index: *Forecast and Trends, 2017-2022*,” Tech. Rep., 2018.
- [3] S. Parkvall, E. Dahlman, A. Furuskär, and M. Frenne, 2017, *NR: The New 5G Radio Access Technology*, IEEE Communications Standards Magazine.
- [4] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia and P. Camarda, "Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey," in IEEE Communications Surveys & Tutorials, vol. 15, no. 2, pp. 678-700, Second Quarter 2013.
- [5] K. C. Beh, S. Armour and A. Doufexi, "Joint Time-Frequency Domain Proportional Fair Scheduler with HARQ for 3GPP LTE Systems," 2008 IEEE 68th Vehicular Technology Conference, Calgary, BC, 2008, pp. 1-5.
- [6] A. Pokhariyal et al., "HARQ Aware Frequency Domain Packet Scheduler with Different Degrees of Fairness for the UTRAN Long Term Evolution," 2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring, Dublin, 2007, pp. 2761-2765.
- [7] Techopedia.com. (n.d.). “What is a Core Network? - Definition from Techopedia”, [online] Available at: <https://www.techopedia.com/definition/6641/core-network> [Accessed 13 Mar. 2019].
- [8] Asante, J. and Olsson, J. (2018). “5G Simulation Framework”, Master of Science Thesis in Electrical Engineering Department of Electrical Engineering, Linköping University.
- [9] Moustafa, W., & Mugisha, R. (2017). “Scheduler Algorithms for MU-MIMO”, Department of Electrical and Information Technology Faculty of Engineering, LTH, Lund University SE-221 00 Lund, Sweden.
- [10] LTE eNodeB Scheduler and Different Scheduler Type - Techplayon. (2018). Retrieved from <http://www.techplayon.com/lte-enodeb-scheduler-and-different-scheduler-type/> [Accessed 6 Jun. 2019].

- [11] Habaebi, M., Dahawi, T., & Chebil, J. (2013). "Comparison between Scheduling Techniques in Long Term Evolution", Department of Electrical and Computer Engineering, Kuliyah of Engineering, International Islamic University Malaysia, Jalan Gombak 53100, Kuala Lumpur, Malaysia.
- [12] Wikipedia. Proportionally_fair. In https://en.wikipedia.org/wiki/Proportionally_fair. [Accessed 6 Jun. 2019].
- [13] The 3rd Generation Partnership Project (3GPP). "First 5G NR Specs Approved" Available at: http://www.3gpp.org/news-events/3gpp-news/1929-nsa_nr_5g [Accessed 16 Mar. 2019].
- [14] E. Dahlman, S. Parkvall, and J. Sköld (2016), *4G — LTE-Advanced Pro and the Road to 5G*, Academic Press, 2016.
- [15] E. Dahlman, (2018). *5G NR Webinar*. Available at: <https://www.youtube.com/watch?v=PDD7OGAO4tc> [Accessed 17 Mar. 2019].
- [16] Ericsson, (2019). 'Ericsson Mobility Report' Available at: <https://www.ericsson.com/assets/local/mobility-report/documents/2018/ericsson-mobility-report-june-2018.pdf> [Accessed 30 Mar. 2019].
- [17] GSA – Global mobile Suppliers Association (2016), 'Small Cell Network White Paper with input from Ericsson and Huawei' Available at: https://www.huawei.com/minisite/hwmbbf16/insights/small_cell_solution.pdf [Accessed 30 Mar. 2019].
- [18] Wikipedia, (2019). *QoS Class Identifier*. Available at: https://en.wikipedia.org/wiki/QoS_Class_Identifier [Accessed 31 Mar. 2019].
- [19] 3GPP Technical specification (2018), '5G; NR, Physical layer procedures for data- 3GPP TS 38.214 version 15.2.0 Release 15'. Available at: https://www.etsi.org/deliver/etsi_ts/138200_138299/138214/15.02.00_60/ts_138214v150200p.pdf [Accessed 7 Apr. 2019].
- [20] Wikipedia, (2019). *Link adaptation*. Available at: https://en.wikipedia.org/wiki/Link_adaptation [Accessed 7 Apr. 2019].

- [21] Afroz, F., Barua, SH. and Sandrasegaran, K. (2014), “*Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system,*” International Journal of Wireless & Mobile Networks (IJWMN) Vol. 6, No. 5.
- [22] 3GPP. *5G; Study on channel model for frequencies from 0.5 to 100 GHz*. Technical Report (TR) 38.901, 3rd Generation Partnership Project (3GPP), January 2018. Version 14.3.0.
- [23] Techplayon (2019) ‘*LTE TDD Special Subframe and Its significance for Cell Size*’ Available at: <http://www.techplayon.com/lte-tdd-special-subframe-and-its-significance-for-cell-size/> [Accessed 14 Apr. 2019].
- [24] Preet, Kanwar & Rekhi, Mohit & Luthra, Sukhvinder & Malik, Rahul & , Atri. (2012). *Throughput Calculation for LTE TDD and FDD Systems*.
- [25] Haziza, F. (2007) *Scheduling Algorithms* [Powerpoint slides]. Uppsala University delivered Spring 2007. Available at: https://www.it.uu.se/edu/course/homepage/oskomp/vt07/lectures/scheduling_algorithms/handout.pdf [Accessed 6 Jun. 2019].
- [26] Máthé, L. (2016). *Traffic Models for Voice over LTE*. Budapest: Budapest University of Technology and Economics.
- [27] Laner, M., Svoboda, P., Nikaein, N., & Rupp, M. *Traffic Models for Machine Type Communications*. Austria: Vienna University of Technology.
- [28] Campos, J. (2017). *Understanding the 5G NR Physical Layer*. Retrieved from https://www.keysight.com/upload/cmc_upload/All/Understanding_the_5G_NR_Physical_Layer.pdf
- [29] W. Koehrsen, "The Poisson Distribution and Poisson Process Explained", Towards Data Science, 2019. [Online]. Available: <https://towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459>. [Accessed: 21- May- 2019].
- [30] Electronics Notes, "Erlang". [Online]. Available: <https://www.electronics-notes.com/articles/connectivity/erlang/what-is-an-erlang-formula.php>. [Accessed: 21- May- 2019].
- [31] "Erlang (unit)", En.wikipedia.org. [Online]. Available: [https://en.wikipedia.org/wiki/Erlang_\(unit\)](https://en.wikipedia.org/wiki/Erlang_(unit)). [Accessed: 21- May- 2019].
- [32] Cisco, "Traffic Analysis", 2011. [Online]. Available: https://www.cisco.com/c/en/us/td/docs/ios/solutions_docs/voip_solutions/TA_ISD.html. [Accessed: 21- May- 2019].
- [33] Confidence Intervals - using t distribution [Video file]. (2015, March 30). Retrieved from <https://www.youtube.com/watch?v=UmAJtEo6cQ> [Accessed: 23- May- 2019].

- [34] T. R. Robbins, D. J. Medeiros and T. P. Harrison, "Does the Erlang C model fit in real call centers?" Proceedings of the 2010 Winter Simulation Conference, Baltimore, MD, 2010, pp. 2853-2864.doi: 10.1109/WSC.2010.5678980
- [35] T. R. Robbins, "Evaluating the fit of the Erlang A model in high traffic call centers," 2016 Winter Simulation Conference (WSC), Washington, DC, 2016, pp. 1790-1801.doi: 10.1109/WSC.2016.7822226
- [36] "Poisson distribution", En.wikipedia.org. [Online]. Available: https://en.wikipedia.org/wiki/Poisson_distribution [Accessed: 10 - Jun- 2019].
- [37] A. S. Pagès, "Link Level Performance Evaluation and Link Abstraction for LTE/LTE-Advanced Downlink. " Ph.D. Dissertation, Polytechnic University of Catalonia, 2015.
- [38] Massive MIMO highlights - Ericsson. (2018). Retrieved from <https://www.ericsson.com/en/news/2018/1/massive-mimo-highlights> [Accessed: 18 - Jun- 2019].



LUND
UNIVERSITY

Series of Master's theses
Department of Electrical and Information Technology
LU/LTH-EIT 2019-706
<http://www.eit.lth.se>