UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

5G RADIO ACCESS NETWORK SLICING

A Degree Thesis

Submitted to the Faculty of the

Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona

Universitat Politècnica de Catalunya

by

Víctor Garcia Cruz

In partial fulfilment

of the requirements for the degree in

TELECOMMUNICATIONS TECHNOLOGIES AND SERVICES ENGINEERING

Advisor: Oriol Sallent Roig

Barcelona, July 2018

Abstract

With the introduction of 5G, it will be possible to simultaneously admit a wider range of applications and business models that was not possible with 4G. In order to carry this out it is necessary network slicing, because with a common network configuration the expected performance would not be obtained. In this way, 5G systems are allowed to manage logical networks with a specific functionality without losing the economy of scale of a common infrastructure.

The objective of this thesis is to study and analyse the operation of RAN slicing through a Matlabj simulator. First, different RAN slicing configurations are proposed controlling certain parameters of radio protocol Layers 2 and 3 that specify how are assigned the radio resources between slices. Next, these approaches are evaluated with some simulations of a multiservice scenario composed of two slices.

In addition, the appearance of a demonstration and a bus is incorporated to see how the network behaves when there is an increase in the demand of services in a specific moment and what effect can have the mobility of the sessions.

Resum

Amb la introducció del 5G es permetrà admetre simultàniament una gamma molt més àmplia d'aplicacions i models de negoci del que es podia fins al moment amb 4G. Però per poder dur això a terme és necessari el *network slicing*, ja que amb una configuració de xarxa comú no s'obtindria el rendiment esperat. D'aquesta manera, es permet als sistemes 5G administrar xarxes lògiques amb una funcionalitat específica sense perdre l'economia d'escala d'una infraestructura comuna.

L'objectiu d'aquesta tesi és estudiar i analitzar el funcionament del RAN *slicing* a través d'un simulador Matlab. Primer es proposen diferents configuracions de RAN *slicing* controlant certs paràmetres de les Capes 2 i 3 del protocol ràdio que s'especifiquen com estan assignats els recursos ràdio entre *slices.* A continuació, s'avaluen aquestes consideracions amb simulacions d'un escenari multiservei compost per dos *slices*.

A més, s'incorpora l'aparició d'una manifestació i d'un autobús per veure com es comporta la xarxa davant d'un increment de la demanda de serveis en un moment determinat i quin efecte pot arribar a tenir la mobilitat de les sessions.

Resumen

Con la introducción del 5G se permitirá admitir simultáneamente una gama mucho más amplia de aplicaciones y modelos de negocio de lo que se podía hasta el momento con 4G. Pero para poder llevar esto a cabo es necesario el *network slicing,* ya que con una configuración de red común no se obtendría el rendimiento esperado. De esta manera, se permite a los sistemas 5G administrar redes lógicas con una funcionalidad específica sin perder la economía de escala de una infraestructura común.

El objetivo de esta tesis es estudiar y analizar el funcionamiento del RAN *slicing* a través de un simulador Matlab. Primero se proponen diferentes configuraciones de RAN *slicing* controlando ciertos parámetros de las Capas 2 y 3 del protocolo radio que especifican como están asignados los recursos radio entre *slices.* A continuación, se evalúan estas consideraciones con la creación de un escenario multiservicio compuesto por dos *slices.*

A demás se incorpora la aparición de una manifestación y de un autobús para ver cómo se comporta la red ante un incremento de la demanda de servicios en un momento determinado y que efecto puede llegar a tener la movilidad de las sesiones.

Acknowledgements

First of all, I would like to express my gratitude to the project supervisor, Oriol Sallent, for his guidance and support during all the project helping me to achieve a correct development of the Bachelor's Thesis.

In addition, I would like to thank Sergio Garcia for giving me access to the UPC servers and thus to complete the simulations much faster.

Finally, I thank my family and friends for all the support they give me during all the degree.

Revision history and approval record

Revision	Date	Purpose
0	06/06/2018	Document creation
1	10/06/2018	Document revision
2	14/06/2018	Document revision
3	22/06/2018	Document revision
4	25/06/2018	Document revision
5	28/06/2018	Document revision
6	30/06/2018	Document revision

DOCUMENT DISTRIBUTION LIST

Name	e-mail
Garcia Cruz, Victor	victor.garcia.cruz@alu-etsetb-upc.edu
Sallent Roig, Oriol	sallent@tsc.upc.edu

Written by:		Reviewed and approved by:	
Date	02/07/2018	Date	02/07/2018
Name	Víctor Garcia Cruz	Name	Oriol Sallent Roig
Position	Project Author	Position	Project Supervisor

Table of contents

The table of contents must be detailed. Each chapter and main section in the thesis must be listed in the "Table of Contents" and each must be given a page number for the location of a particular text.

List of Figures

List of Tables:

1. Introduction

1.1. Statement of purpose

5G systems are intended to simultaneously support a wider range of applications scenarios and business models. This expected versatility comes with a high variety of requirements and expected performance that cannot always be met through a common network setting. In this respect, network slicing has become a fundamental capability for 5G networks to facilitate the cost-effective deployment and operation of multiple logical networks over a common physical network infrastructure in a way that each network slice can be customized and dimensioned to best serve the needs of specific applications and users.

This project is carried out at the Department of Signal Theory and Communication (TSC), which collaborates with 5G-PPP (Public Private Partnership). 5G PPP [1] is a joint initiative between the European Commission and European ICT industry that will deliver solutions, architectures, technologies and standards for the next generation communication infrastructures of the coming decade. 5G PPP divides the work into three different phases with some projects in each phase. The UPC in particular collaborates with the projects:

- SESAME in the Phase I.
- 5G ESSENCE in the Phase II.
- In the Phase III, some Trials will be done.

This project consists on analyzing and working with a Matlab code which simulates a multi-service scenario for RAN slice deployment in 5G context. The purpose of the project is to further understand the specification of RAN slice configuration parameters and see how the network responds to different scenarios.

The project main goals are:

1- To identify the fundamental design challenges for the realization of RAN slicing.

2- To formulate, develop and assess a selected case of RAN slice deployment.

1.2. Project requirements and specifications

Project requirements:

- To provide a particular system behavior.
- It has to provide a guaranteed level of network resources.
- Ensure isolation between different slices when it is required.

Project specifications:

- It has to satisfy the 3GPP normative specifications. [2]
- A set of configuration descriptors of the radio protocol layers L3, L2 and L1 is needed to specify the operation of each RAN slice.

1.3. Methods and procedures

The project has its initial point in two papers [3][4] published by the TSC Department and supported by the 5G-PPP project 5G ESSENCE. The first paper proposes a framework for the support and specification of RAN slices based on the definition of a set of configuration descriptors that characterize the features, policies and resources to be put in a place across the radio protocol layers of a next generation RAN node.

Based on this, the contribution of the second paper is to propose and analyze different options for configuring RAN slices by controlling certain parameters at radio protocol Layer 2 (L2) and Layer 3 (L3). Therefore, in this project we want to see what changes by modifying these parameters and we pretend to find out what happened when we add new sessions or we modify the scenario.

The basis of the Matlab code used in the project is the one associated to the second paper. In that Matlab code is where we will work and we will add some things and modify others to obtain the desired results.

1.4. Work Plan

In order to complete all the objectives of the project we divide it in 7 different parts, making the work easier and with a more organized structure. The 7 subprojects are:

1. Background
2. Simulator
3. Edited simulations
4. Quantitative evaluation
5. Protest appearance
6. Bus appearance
7. Conclusions and report

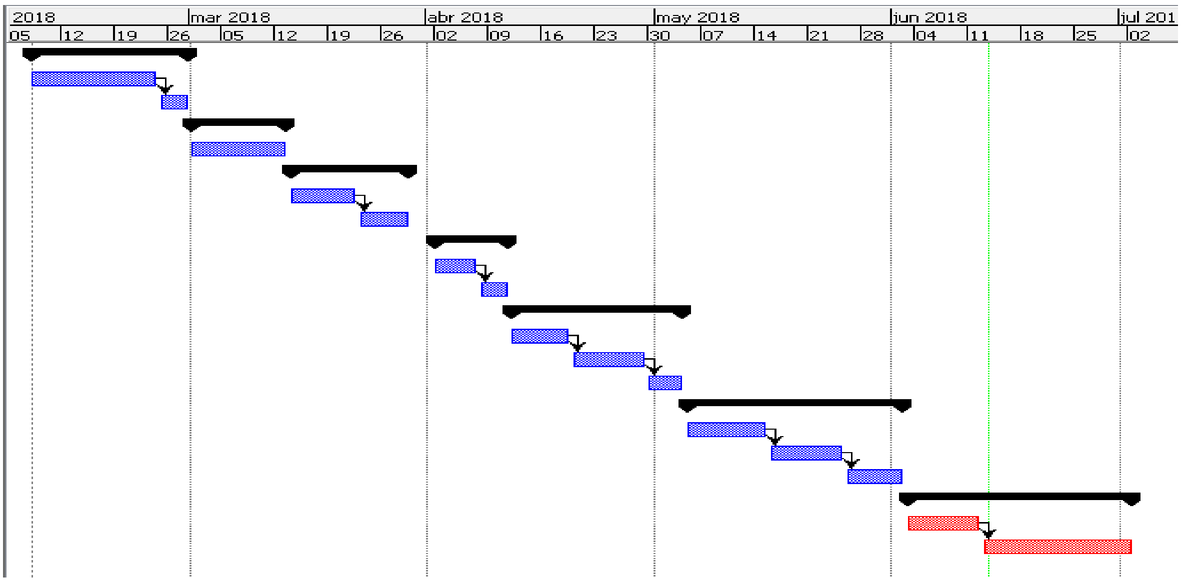
The Work Packages of all the project summarising each of these subprojects can be found in the Appendices (1).

1.4.1 Milestones

WP#	Task#	Short title	Milestone / deliverable	Date (week)
1	1.1	Read both papers	Some slices summarizing the most important things	24/02/2018
	1.2	Focus the project		28/02/2018
2	2.1	Get used to the simulator	**Project Proposal and Work Plan**	5/03/2018
			Be able to use the simulator	13/03/2018
3	3.1	Change the parameters and simulate	Create simulations with different configuration parameters	22/03/2018
	3.2	Study the results	Some slices showing the different results obtained with each simulation	29/03/2018
4	4.1	Compute some parameters	Do the necessary computation	7/04/2018
	4.2	Compare with the simulation results	Document with the calculations	11/04/2018
5	5.1	Add the protest appearance	Write the code necessary	19/04/2018
	5.2	Do simulations with different configurations		29/04/2018
	5.3	Study the results	Some slices with the conclusions I arrive	4/05/2018
			Critical Review	7/05/2018
6	6.1	Add the bus appearance (with Handover's rate)	Write the code necessary and compute the rate of Handovers	15/05/2018
	6.2	Do simulations with different configurations		25/05/2018
	6.3	Study the results	Performance evaluation	2/06/2018
7	7.1	Arrive to some conclusions		12/06/2018
	7.2	Write the final report	**Final Report**	2/07/2018

1.4.2 Gantt diagram

		Nombre	Duración	Inicio	Terminado	Predecesores
1		**Background**	21 days	**8/02/18 8:00**	**28/02/18 17:00**	
2		Read both projects	17 days	8/02/18 8:00	24/02/18 17:00	
3		Focus the project	4 days	25/02/18 8:00	28/02/18 17:00	2
4		**Simulator**	13 days	**1/03/18 8:00**	**13/03/18 17:00**	
5		Get used to the simulator	13 days	1/03/18 8:00	13/03/18 17:00	
6		**Edited simulations**	16 days	**14/03/18 8:00**	**29/03/18 17:00**	
7		Change some parameters and simulate	9 days	14/03/18 8:00	22/03/18 17:00	
8		Study the results	7 days	23/03/18 8:00	29/03/18 17:00	7
9		**Quantitative evaluation**	10 days	**2/04/18 8:00**	**11/04/18 17:00**	
10		Compute some parameters	6 days	2/04/18 8:00	7/04/18 17:00	
11		Compare with the simulation results	4 days	8/04/18 8:00	11/04/18 17:00	10
12		**Protest appearance**	23 days	**12/04/18 8:00**	**4/05/18 17:00**	
13		Add the protest apperance	8 days	12/04/18 8:00	19/04/18 17:00	
14		Do simulations with different configurations	10 days	20/04/18 8:00	29/04/18 17:00	13
15		Study the results	5 days	30/04/18 8:00	4/05/18 17:00	14
16		**Bus appearance**	29 days	**5/05/18 8:00**	**2/06/18 17:00**	
17		Add the bus appearance (with Handover's rate)	11 days	5/05/18 8:00	15/05/18 17:00	
18		Do simulations with different configurations	10 days	16/05/18 8:00	25/05/18 17:00	17
19		Study the results	8 days	26/05/18 8:00	2/06/18 17:00	18
20		**Conclusions and report**	29 days	**3/06/18 8:00**	**2/07/18 17:00**	
21		Arrive to some conclusions	10 days	3/06/18 8:00	12/06/18 17:00	
22		Write the final report	19 days	13/06/18 8:00	2/07/18 17:00	21



1.4.3 Deviations from the initial plan and incidences

During the project we have had some incidences that have made us modify some parts of the project. After getting used to the simulator, we decided to change some parameters related with how we defined the service mix of the services and the load of each slice to observe what happens and have a better knowledge of the system operation.

After that, the first idea was to create a new scenario, but when we saw that it lasted so much time to do a complete simulation (approximately 10 hours) we decided to use the scenario created adding some situations (bus appearance or protest appearance) to get different results.

In addition, we decided to add the quantitative evaluation of the results to be sure that what we obtained from the simulation make sense.

2. State of the art of the technology used or applied in this thesis:

2.1. Evolution until 5G [5]

The first commercial mobile communication network was launched in Japan in 1979 by NTT. Since then, the mobile communications have been in constant evolution, being able to do things unthinkable 30 years ago.

The first generation (1G) only incorporated the voice service and was an analogue technology based on FDMA. At that time, voice communications were of low quality and there was no security.

With the second generation (2G) was presented a digital technology, GSM (Global System for Mobile Communications), capable of adding data transmission, which gives rise to SMS and the concept of roaming. 2G used TDMA as an access technique and was based on circuit switching.

Before reaching 3G, two new standards, GPRS and EDGE, were developed, in which packet switching is introduced to provide high-speed data Internet. This improvement was given the name of 2.5G.

Figure 1.Mobile communication generations [6]

With 3G it was intended to offer a higher data rate and increase the number of applications capable of supporting. Voice calls continued to use circuit switching and data transmission packet switching. The access technique used is CDMA and is governed by the UMTS standard. A new standard appeared later, HSPA, which was the one that started offering high-speed Internet access or video calls.

The latest standardized generation is 4G with the LTE standard. It uses OFDMA and there is a considerable increase in speed. The services that this generation incorporates are IP telephony or video conference.

The future of mobile communications is the 5G that will mean a great revolution in terms of available applications. For this, it will be necessary to multiply the transmission speed and it is expected to start up in 2020.

2.2. Introduction to 5G

It is expected that 5G technology would be one of the most important technological developments of our time. It will connect billions of "elements" that have not been connected before, adding artificial intelligence and data in cars, homes, buildings, factories or cities. Some capabilities expected for the 5G and that the current generation cannot cover are the following [7]:

- Increase by a factor 10 to 100 of the connected devices.
- Latency: It is intended to reduce the latency to 1ms, which is unnoticeable for the user.
- Data rates are expected to reach peaks of 10 GBit/s

The 4G systems were designed to provide a “one size fits all” mobile broadband solution but the 5G systems need support for network slicing. Network slicing [8] allows a network operator to provide dedicated virtual networks with specific functionality to the service or costumer over a common network infrastructure. The virtual networks are then customized to meet the specific needs of all the varied applications and services of 5G. This expected versatility comes with a high variety of requirements on network functionalities and expected performance.

Each of these network slices can be tailored to fulfil a couple of proposes [3]:

- To provide a particular system behaviour through the use of specific control plane (CP) [9] or user plane (UP) [10] functions to best support specific service/application domains. For instance, a User Equipment (UE) for smart metering applications can be served through a network slice with radio access.
- To provide a particular tenant with a given level of guaranteed network resources and isolation with regard to the operation of other concurrent slices.

The realization of network slices considers support for specific features and resources both in the 5G Core Network (5GC) part, referred to as Core Network (CN) slice, and in the New Generation Radio Access Network (NG-RAN) part, referred to as RAN slice. The realization of RAN slices is particularly challenging because it requires addressing how the pool of radio resources available to one NG-RAN node can be configured and operated to simultaneously deliver multiple and diverse RAN behaviours, and this is what we focus on.

2.3 RAN Slicing [3][4]

A first step to understand the impact of network slicing to the 5G RAN design has been given by identifying RAN-specific requirements [11] needed to fulfil the network slicing vision. Some requirements are the following:

- Utilization of RAN resources should be maximized.
- RAN should be slice-aware, making it possible to prioritize different service and signalling in a different way.
- RAN should support mechanisms for traffic differentiation in order to be able to treat different slices differently or different services within the multi service slices.
- RAN should support protection mechanisms for slice isolation so that events within one slice do not have a negative impact on another slice.

In order to fulfil with these requirements a set of new blocks of information, configuration descriptors and protocol features has to be introduced across the protocol layers of a NG-RAN node. The proposed overall framework for RAN slicing support within a NG-RAN node is illustrated in Fig. 2 and explained in the following.

<u>UE Context</u>**:** is instantiated within the RAN at the time the UE becomes active. It is a block of information that contains all the necessary data required to maintain the RAN services towards the UE. It is added a RAN slice identifier to the use it as a pointer to a new block of information denoted as RAN Slice Context.

<u>RAN Slice Context</u>**:** it contains all the data necessary to support the operation of a particular RAN slice along with the *Slice_ID(s)* that are served though the RAN slice. A *RAN Slice Descriptor* is introduced as the baseline descriptor and it includes at least the *RAN_Slice_ID*, the list of associated *Slice_ID(s)* and *PLMN_ID(s)*, and a set of pointers to the configuration descriptors of the underlying radio protocol layers 3, 2 and 1 (L3, L2, L1) for the realization of the RAN slice.

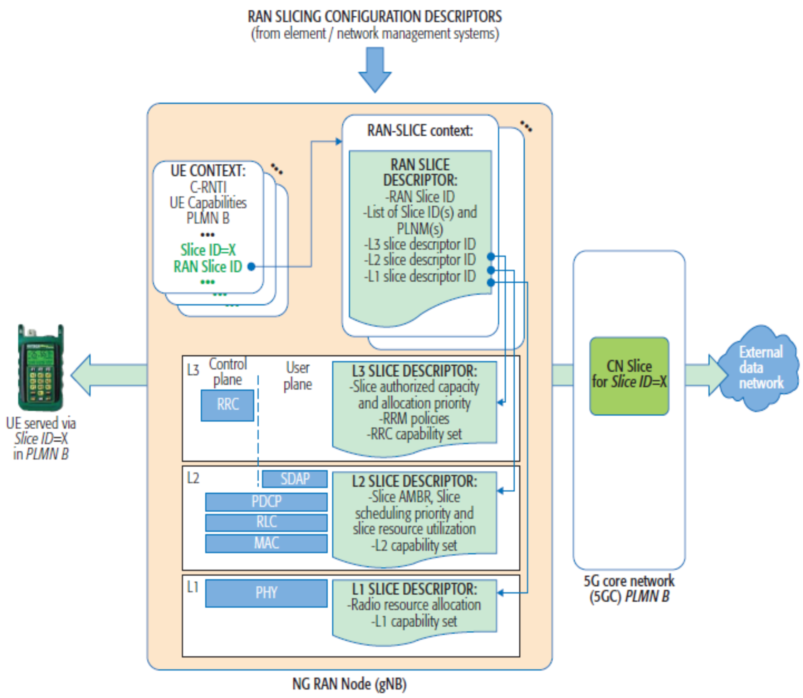


Figure 2. Framework for the realization of RAN slices in a NG-RAN node

A more detailed definition of the layer descriptors is given below.

2.3.1 L3 Configuration

A *L3 Slice Descriptor* is necessary to specify the capacity allocation for the RAN slice, the RRM policies that govern the operation of the slice and the capability set of the RRC protocol in use. When multiple RAN slices are realized over shared radio resources, the RRM functions for RBC, RAC and CMC have to assure that each RAN slice gets the expected amount of resources and handles any conflicts that might appear across slices. To dictate the operation of these functions the following parameters are proposed:

- **Slice Authorized Capacity:** this can be a combination of resource-oriented (absolute or relative occupation levels of the consumed radio resources) and rate-oriented (rate limits on the aggregate bit rate of the entire set of admitted guaranteed bit rate (GBR) RBs within the slice). It is used by the RAC for the admission or rejection of RBs. It is the main mean control of L3.

- **Slice Allocation Priority:** this parameter allows for conflict resolution among UE/RB resource requirements across slices that cannot be solved based only on the Slice Authorized Capacity parameters. The priority and pre-emption policies at UE/RB level are solved through the allocation and retention priority (ARP) parameter included in the QoS profile.

If multiple RAN slices are configured to share the same set of common logical channels, the following extended features have to be incorporated within the RRC protocol:

- Protocol fields to allow UEs to discriminate among signalling from different slices.
- System information block (SIB) messages to advertise the *Slice_ID(s)* that can be reached from the cell.
- Paging configuration features allowing paging cycles to be organized considering the specific needs to each *Slice-ID.*

2.3.2 L2 Configuration

L2 comprises a Medium Access Control (MAC) sub-layer for multiplexing and scheduling the packet transmission of the DRBs over a set of transport channels exposed by L1. Considering that the current MAC operation is based on individual UE and DRB specific QoS profiles, it is necessary:

- Define the Packet Scheduling (PS) behaviours to be enforced on the traffic aggregate of DRBs of the same slice.
- Specify the capability set of the applicable L2 sub-layers processing functions.

Therefore, the proposed *L2 Slice Descriptor* includes the following parameters to dictate the operation of the MAC scheduler and yield isolation:

- **Slice-AMBR:** to limit the aggregate bit rate of all non-GBR RBs associated with the slice.
- **Slice Scheduling Priority:** to handle short-term traffic congestion between RBs with the same QoS profile.
- **Slice Resource Utilization:** used to establish constraints on the amount of physical-layer resources scheduled by the MAC that are consumed by the slice (minimum % of PRBs that the PS guarantees to the slice for allocating transmissions of both GBR and non-GBR bearers). It is the main mean control of L2.

2.3.3 L1 Configuration

The new physical layer for 5G NR is being defined with the goal to provide high flexibility for the use of different waveforms and adaptable time-frequency frame structures.

L1 provides L2 with transfer services in the form of transport channels, which define how the data is transferred. *L1 Slice Descriptor* is necessary to specify both L1 transfer service capabilities and specific radio resource allocation.

Considering that the L1 optimal settings can differ per slice type, the L1 descriptor intends to establish a partitioning of the L1 radio resource structure so that different L1 optimization settings can be simultaneously applied. The mixing of L1 slices could be achieved through the use of the different OFDM numerologies.

Some important parameters that L1 defines are:

- **Radio Resource Allocation**: number of PRBs
- **Numerology :** subcarrier separation (Δf)

2.4 Simulation tool

I was provided with a Matlab code that simulates a multi RAN slices deployment scenario and it computes some parameters that helps us to understand better the operation of the network.

In this section I explain the code of the main file of the simulation step by step.

Code summary:

- FOR (iteration through all the average session generation rates)
  - Declaration of constants:
    - Slicing algorithm
    - AC algorithms
    - General scenario parameters (cell radio, number of cells, number of RBs, number of slices, …)
    - Propagation model parameters (frequency)
    - Spectral Efficiency computations parameters (SINRmin, Smax, …)
    - Service profiles
    - Parameter for the scheduling of Non-GBR UEs
    - Traffic parameters
    - Admission parameters
    - Traffic spatial parameters (Gaussian distribution)
    - Capacity share parameters per slice
    - SAGBR for the total scenario
    - SAGBR per cell
  - Setting of some actions:
    - Distribute and initialize the cells (*init_BS* [1])
    - Apply the spatial distribution to determine the traffic at each cell. In this case, we disable the traffic generation based on hotspots and we specify the traffic per slice as:
      - Slice 1→ It takes the value of the first FOR (from 0.5 to 3)
      - Slice 2→ 2 sessions/s
    - Specify the service mix for each cell and slice
    - Define the parameters for the Slicing at PS and Slicing at Spectrum Planning.
    - Define the parameters for the No Slicing Case
    - Initialize the radio parameters of each cell and schedule the next session arrival rates (*init_radio_and_next_arrivals*)

[1] It is in cursive because is the function that it is called.

- FOR (all the simulation time): in each of the following steps the first two lines of code are two FOR loops. The first one iterate through all the cells and the second one through all the slices.
  - Check session finalisations
    - If the process is ended
      - Remove UE from the list.
  - Check session starts
    - If the session can be admitted
      - Generate the new UE (*init_UE*)
    - If the session cannot be admitted
      - Count a blocking
  - Check activity of the UEs
  - If some change has occurred
    - Run *compute_occupation*
  - Update the averages of: - Average number of RBs used by each slice
    - Average RB utilisation per slice
    - Estimate bit rate per RB achieved in the cell
    - Average bit rate assigned per tenant
    - Compute data volume
    - Check congestion status per slice
    - Estimate of bit rate per RB
    - Check congestion status
  - Compute: - Average RB utilisation of each slice at multi-cell level
    - The aggregate bit rate of each tenant at multi-cell level
  - Compute and update the deltaC parameters.
  - Reduce the size of the variable list_UEs to include only the actual entries
  - Measure final statistics and plots
    - Average along the whole simulations is saved in the vaiable stats.
    - We generate an output file (.mat) for each value of the vector variation.

As it can be seen, the main file call other functions to compute some parameters or do some actions. To have a better framework of the simulator, I do a block diagram showing the connection between the main file and the other functions (Figure 3).

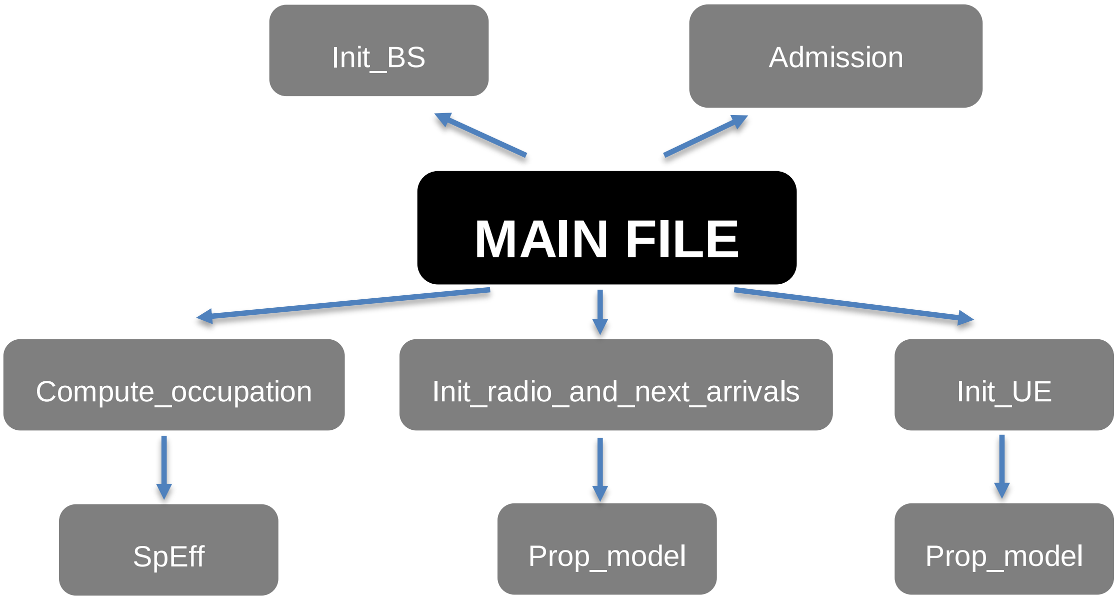


Figure 3. Simulator structure

A short description of each function:

- **Init_BS:** It is used in the main files to initialize a new cell. It has as an input parameter the variable *config.*
- **Init_radio_and_next_arrivals:** It is used in the main files to initialize the radio parameters of each cell and schedule the next session arrival rates. The input parameter is *config.*
- **Admission:** Check if a new session can be admitted or not. The input parameters are: *slice, Rbreq, config* and *service.* It returns the parameter *admission_result,* which can have the values 0 or 1.
- **Init_UE:** It is used to initialize a new UE. The input parameters are: *config, BS, cell, slice.*
- **Compute_occupation:** This function is called in the main file if some change is produced. First, it computes the SINR and the number of required RBs by each UE to determine the number of assigned RBs. Finally, it computes the final bit rate assigned to each slice.
- **SpEff:** It computes the spectral efficiency as a function of the SINR in linear. This function is called in *compute_occupation.*
- **Prop_model:** It computes the total loss in dB.

3. Project development and results:

In this section I am going to explain the different experiments we have done in this project. As we have realised more than one experiment and each experiment has its own results, we have included these results here to make a clearer understanding.

3.1 Simulation scenario

The scenario that it is proposed consists on a multiservice scenario that has deployed a NG-RAN which is configured with two RAN slices. The RAN Slice 1 is for the eMBB services and RAN Slice 2 for the Mission Critical (MC) [12] services. The considered services of each RAN Slice are summarized in Table 1.

RAN Slice ID	Service	Type	5QI (Priority)	ARP	GFBR	Service mix	Average session generation rate
1	Premium – Video HD	GBR	2 (40)	2	10 Mb/s	10%	Varied from 0.5 to 3 sessions/s
	Premium – Data	Non-GBR	6 (60)	2	N/A	20%	
	Basic – Video	GBR	2 (40)	3	1 Mb/s	30%	
	Basic – Data	Non-GBR	8 (80)	3	N/A	40%	
2	MC Video	GBR	2 (40)	2	2 Mb/s	10%	2 sessions/s
	MC PTT	GBR	65 (7)	1	10 kb/s	50%	
	MC Data	Non-GBR	70 (55)	3	N/A	40%	

Table 1. Services of each RAN Slice

The table gives us some information about the considered services:

- **5G QoS Identifier (5QI):** it is a scalar that is mapped to specific QoS characteristics in terms of a priority in the scheduling process. The lower the value (indicated in parenthesis) the higher the priority.
- **Allocation Retention and Priority (ARP):** it defines the importance of a resource request and allows deciding if a new session may be rejected in case of resource limitations. Lower values represent higher priorities.
- **Guaranteed Flow Bit Rate (GFBR):** it specifies the bit rate to provide to a GBR service session.
- **Service mix:** how much rate of sessions of the RAN that are from that service.
- **Average session generation rate:** Traffic generation assumes that services generate sessions following a Poisson process with the average rate indicated for each slice.

3.1.1 Simulation parameters

The deployment assumes a gNB with a single cell configured with a channel of 100 MHz organized in 275 Physical Resources Blocks (PRBs) composed by 12 subcarriers with a separation Δf=30 kHz. In Table 2 are presented the considered simulation parameters.

Parameter	Value
Cell radius	115 m
Path loss and shadowing model	Urban micro-cell model with hexagonal layout (details in [15])
Shadowing standard deviation	3 dB in Line of Sight (LOS) 4 dB in Non Line of Sight (NLOS)
Base station antenna gain	5 dB
Frequency	3.6 GHz
Transmitted power per PRB	16.6 dBm
Number of PRBs	275
UE noise figure	9 dB
Link-level model to map SINR and bit rate	Maximum spectral efficiency of 8.8 b/s/Hz
Average session duration	120 s
Activity factor of Non-GBR services	0.2
Averaging period for measuring PRB occupation	30 s
Simulation duration	20000 s

Table 2. Simulation Parameters

3.1.2 RAN Slice configurations

To have a more completed analysis we considered three different configurations of the slices by controlling certain parameters at radio protocol Layers 2 and 3. These parameters allows configuring the RAC and PS functionalities that specify how the different radio resources are allocated to the slices. The configurations are shown in Table 3.

	Configuration #0 (Slice agnostic)		Configuration #1 (Slice-aware L3)		Configuration #2 (Slice-aware L2 & L3)	
Control parameters	**RAN Slice ID=1**	**RAN Slice ID=2**	**RAN Slice ID=1**	**RAN Slice ID=2**	**RAN Slice ID=1**	**RAN Slice ID=2**
L3- Maximum % of PRBs for the admission of GBR DRBs	70%		50%	20%	50%	20%
L2- Minimum % of PRBs that the PS guaranteed to the slice	N/A	N/A	N/A	N/A	70%	30%
L1- Number of PRBs	275 PRBs		275 PRBs		275 PRBs	
L1- Numerology: Subcarrier separation (Δf)	30 kHz		30 kHz		30 kHz	

Table 3. RAN Slice Configurations

- **Configuration #0 (Slice-agnostic):** The L3 RAC function does not make distinctions among slices establishing a limit of 70% of PRBs for all the GBR DRBs of the two slices. At L2 the PS operates on the basis of the 5QI parameter.
- **Configuration #1 (Slice-aware L3):** Now the RAC function distincts among slices, giving a 50% of capacity for GBR bearers to RAN Slice 1 and a 20% to Slice 2. The PS continue not making differentiations among slices.
- **Configuration #2 (Slice-aware L2 & L3):** L3 considers the same as Configuration #1 and the PS will ensure at least 70% of PRBs for Slice 1 and 30% for Slice 2.

The RAC takes into consideration the different priorities associated with the ARP of each DRB. More specifically, to admit a GBR DRB of RAN slice *s* requesting a guaranteed bit rate $GFBR_i$ with an ARP value ARP_i, the following condition must be fulfilled:

$$\rho_{occ}(ARP_i, s) + \Delta\rho(GFBR_i) \leq \rho_{max}(s)$$

where: - $\rho_{occ}(ARP_i, s)$ measures the average % of PRBs occupied by the GBR bearers of slice *s* that have an ARP lower or equal than ARP_i.

- $\Delta\rho(GFBR_i)$ is the estimated % of PRBs needed to provide a bit rate equal to $GFBR_i$.

- $\rho_{max}(s)$ is the admission control limit.

The resource allocation process at the PS first distributes the PRBs among the admitted GBR bearers and then the remaining PRBs are allocated among the active non-GBR DRBs. In case that a slice has no active non-GBR DRBs, the remaining PRBs are distributed among the non-GBR bearers of the other slice according to their priority level.

3.1.3 Simulation results

With this scenario a simulation for each of the configurations is done to see the difference between them. The Key Performance Indicators (KPI) of the GBR services is the blocking rate, which measures the percentage of GBR DRBs that are rejected by the admission control. In the case of non-GBR services, the main KPI considered is the average throughput obtained by each DRB.

<u>GBR services:</u>

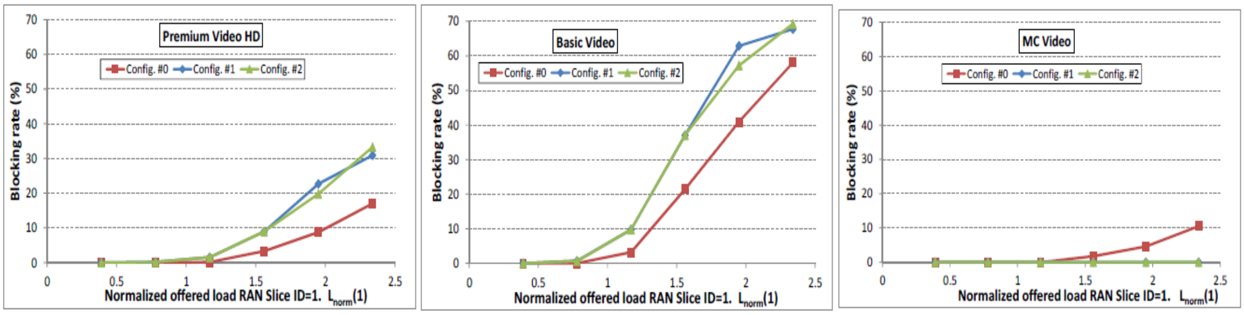


Figure 4. Blocking rate of Premium Video, Basic Video and MC Video respectively

In Figure 4 it can be seen that the blocking rate of these services increases with the total offered load of their slice, being significant when the load is higher than 1. Also, we can see that the blocking rate is higher for Basic Video than for the Premium because the latter has a lower ARP value so it has more priority.

The blocking rate is primarily affected by the admission control of the L3 parameter, and this is why we do not see almost any difference between Configurations #1 and #2, because they have the same configuration of L3. Regarding MC Video, with both Configurations #1 and #2, the blocking rate is 0% because the 20% that it is configured for the slice 2 is enough to serve both the MC Video and MC PTT services.

With the Configuration #0, the blocking rate of Premium and Basic Video decreases in relation to the other configurations but this is at the expense of a degradation in the blocking rate of MC Video. This is explained by the fact that this configuration does not distinct between RAN slices and just configures a total admission control limit of 70%, allowing to Premium and Basic Videos consuming more than the limit of 50% imposed by Configurations #1 and #2 and remaining less than the 20% of PRBs. This leads to some blockings of MC Video sessions.

The MC PTT service is not shown because it is the service with the highest priority so the blocking rate is always 0%.

Non-GBR services:

Non-GBR services do not pass any admission control check and they make use of the PRBs that are available after having performed the PRB allocation to the GBR services.

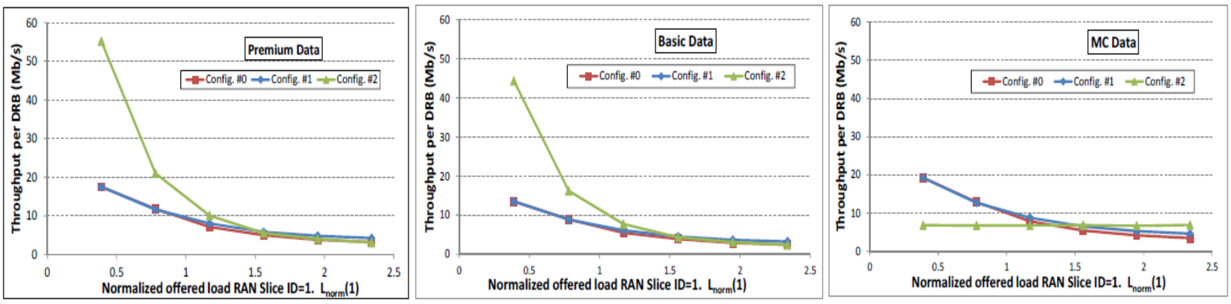


Figure 5. Average throughput per DRB of Premium Data, Basic Data and MC Data respectively

In Figure 5 it is observed that the throughput per DRB of the Premium and Basic Data services decreases as the load increases with the three configurations. There are two reasons for that. First, if the load increases we have more GBR sessions so there are less available PRBs for the non-GBR services. In addition, if the load increases we also have more non-GBR sessions, so the available PRBs have to be distributed among more non-GBR DRBs.

The higher priority associated to the 5QI value of Premium Data achieves that its throughput is higher than Basic Data gets. There are almost no differences in the throughput of Configurations #0 and #1 because none makes use of the L2 control to ensure a minimum amount of PRBs per slice at the PS. Therefore, the remaining PRBs in the cell after having allocated the GBR DRBs are shared between both slices.

However, with Configuration #2 the throughput of the two services of RAN Slice 1 is considerably increased because it ensures at least the 70% of the PRBs in the cell for serving all its DRBs (GBR and non-GBR).

Regarding MC Data, Figure 4 reveals that with Configuration #2 the throughput is constant but with the other two configurations the throughput decreases as the load increases because the available PRBs are shared among the non-GBR services of the two slices.

3.2 Simulations with different configurations

In this section we have changed some parameters related to the service mix of the services and the load of each slice to see how the network reacts. The KPI computed are the same as before so I can compare the results obtained. Here we explain two different edited simulations but in the Appendices (2) there is some more that has helped us to reach the conclusions.

3.2.1 Edited simulation 1

In the first edition we give more priority to MC Video and I have changed the service mix of both slices. In RAN Slice 1 all the services have the same percentage and in RAN Slice 2 the probability of session of MC Video has been increased. In Table 4 are shown the parameters that have been changed with the new values.

RAN Slice ID	Service	ARP	Service mix
1	Premium – Video HD	2	25%
	Premium – Data	2	25%
	Basic – Video	3	25%
	Basic – Data	3	25%
2	MC Video	1	20%
	MC PTT	1	40%
	MC Data	3	40%

Table 4. Service configuration of edited simulation 1.

The results obtained are the following:

GBR services:

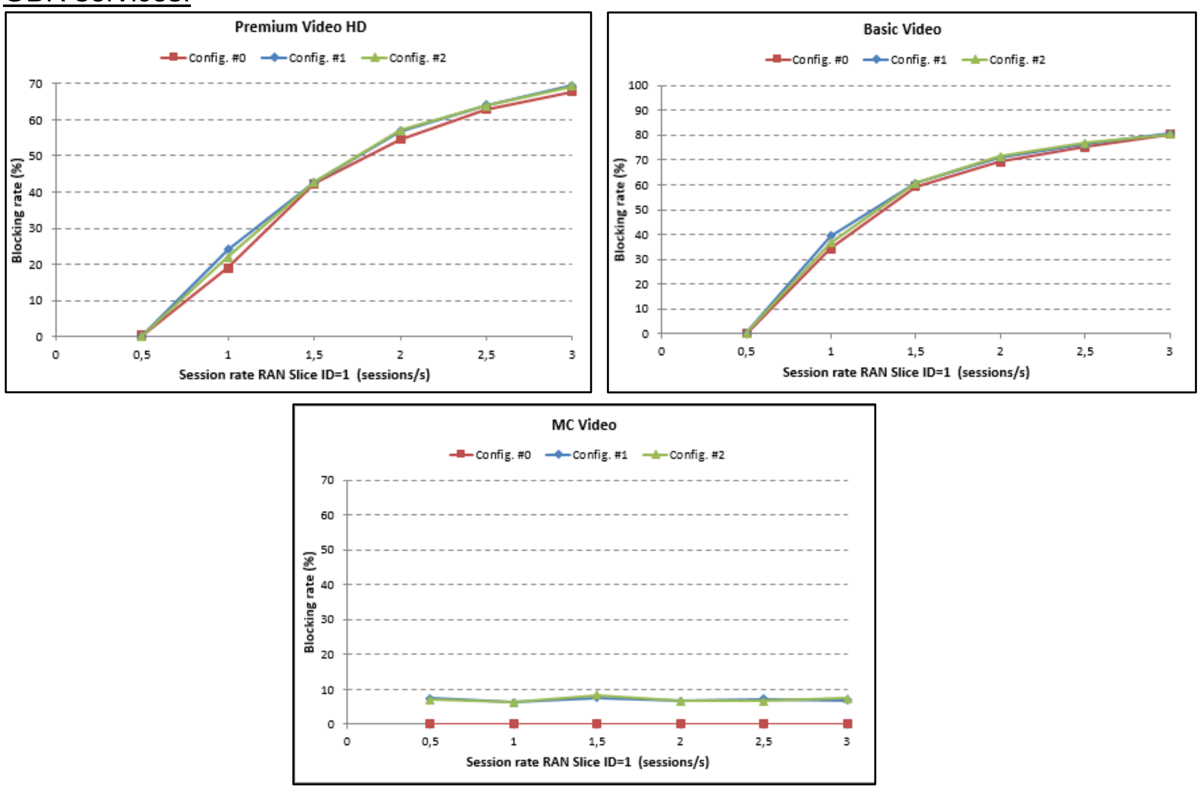


Figure 6. Blocking rate with edited simulation 1

Setting the highest priority to MC Video implies that with Configuration #0 the blocking rate of this service is 0% because L3 parameter does not distinct among slices so the PRBs are first assigned to this service. However, with the configurations #1 and #2, the 20% defined by the RAC parameter is not enough to support the sessions generated from MC Video and MC PTT (the two services with ARP=1) so a small blocking rate appears.

The fact of changing the service mix has its consequences too. We can see in Figure 6 that with a low session load it already appears a small blocking rate, reaching a maximum of 70% in Premium Video and 80% in Basic Video when the load is 3 sessions/s. The blocking rate has increased a lot with respect the original simulation. It can be appreciated that the three configurations follow more or less the same evolution of the blocking rate for the RAN Slice 1 services but there is something of coincidence in this fact. There are mainly two reasons of the increase of the blocking rate of these services:

1. The fact of increasing the priority and the probability of MC Video makes that in Configuration #0 more PRBs go to RAN Slice 2 and there are less to assign to the RAN Slice 1 services, increasing the blocking rate.
2. More sessions of Premium Video than in the original configuration are generated by how we have defined the new service mix (25%), so that with the 50% of PRBs for Slice 1 defined by the RAC control parameter of Configurations #1 and #2, is not enough and we have a greater blocking rate.

<u>Non-GBR services:</u>

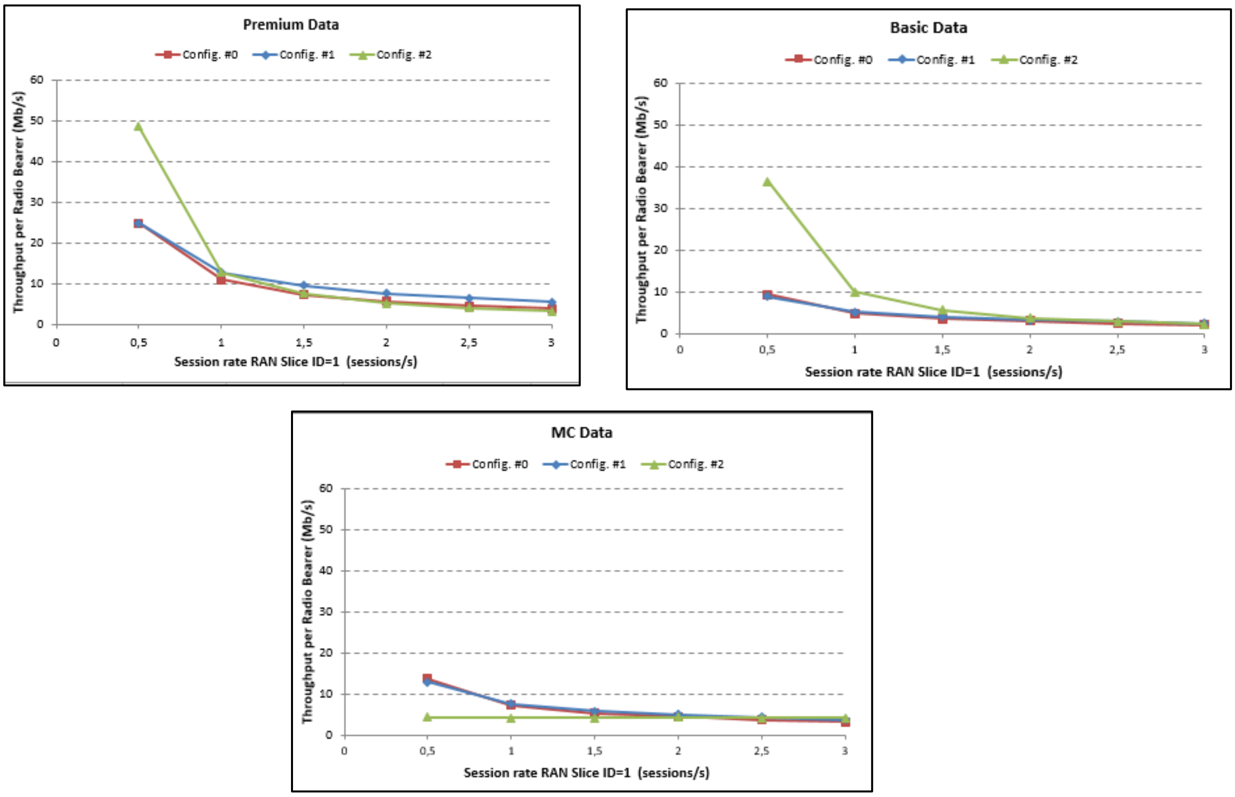


Figure 7. Average throughput with edited simulation 1

The results of the non-GBR services are practically the same than in the original simulation just changing a little the average throughput.

With the Configuration #2, which makes distinction among slices, the three services has a very little decrease in the throughput. Both slices are influenced by the same reason, an increment of the sessions of a GBR service that makes there are less PRBs for the non-GBR services. In RAN Slice 1 is Premium Video who causes this situation whereas for RAN Slice 2 is MC Video.

With the other two configurations we can observe that the throughput for Premium Data has been increased at the expense of decreasing the Basic Data and MC Data throughput23The reasons for this are:

- There are more sessions originated from Premium Data than from MC Data so when the PRBs are going to be distributed there are more Premium Data sessions to be assigned.
- Premium Data has a lower 5QI value than Basic Data, so the ratio between these two services has to match the ratio between priorities of Table 1 (80/60).

Assessment:

To sum up, changing to this service configuration is not profitable for almost any service because we increase a lot the sessions of Premium Video, the service with the highest value of GFBR, so all the PRBs required to this service makes that all the others that have less priority will be harmed. In addition, the increase of sessions and priority of MC Video guarantees better performance for this service but the others are also harmed.

3.2.2 Edited simulation 2

In the second edited simulation we focus on the changes that happen when the average session generation rate changes. In this case, we change the configuration of each slice by the opposite, that is to say, RAN Slice 1 has 2 sessions/s and RAN Slice 2 varies from 0.5 to 3 sessions/s. The service mix configuration is the same as the original simulation and the ARP of MC Video changes to 1. The Table 5 shows the parameter configuration:

RAN Slice ID	Service	ARP	Service mix	Average session generation rate
1	Premium – Video HD	2	10%	2 sessions/s
	Premium – Data	2	20%	
	Basic – Video	3	30%	
	Basic – Data	3	40%	
2	MC Video	1	10%	Varied from 0.5 to 3 sessions/s
	MC PTT	1	50%	
	MC Data	3	40%	

Table 5. Service configuration of edited simulation 2

The results obtained are the following:

<u>GBR services:</u>

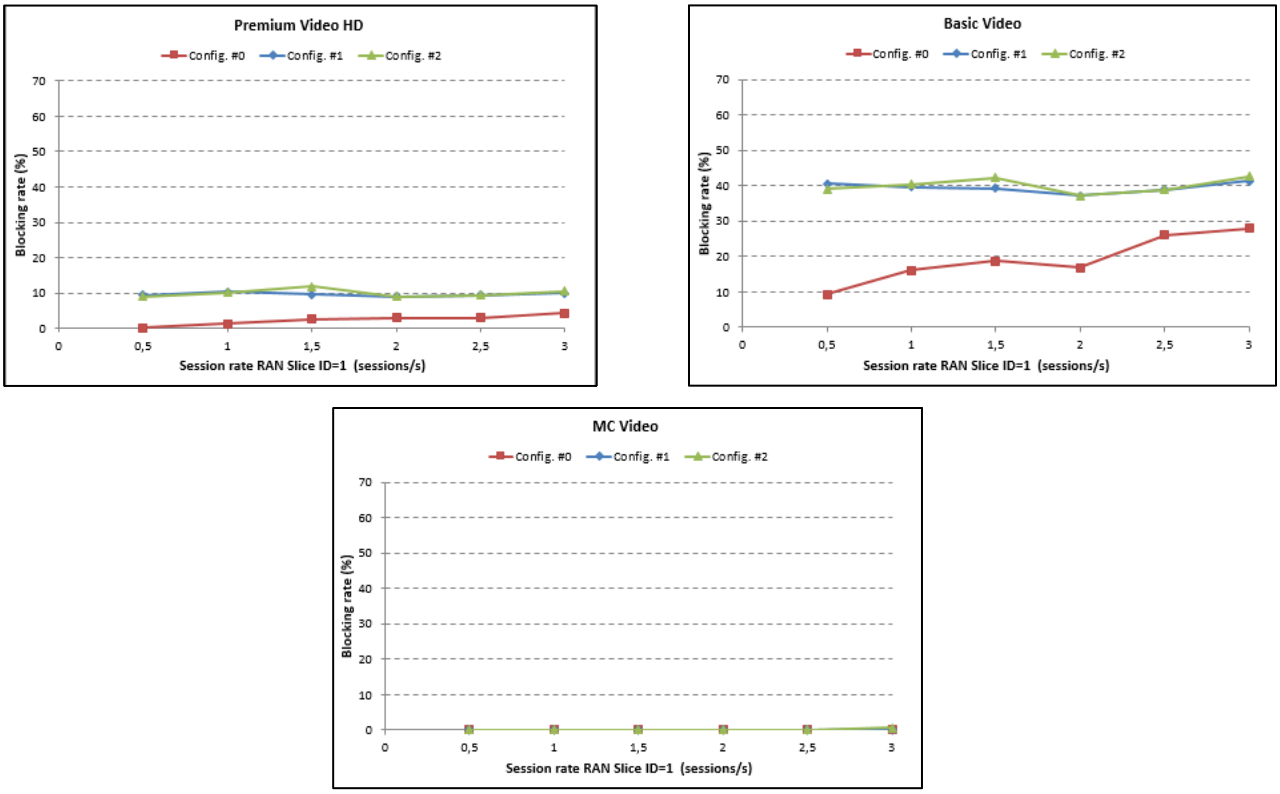


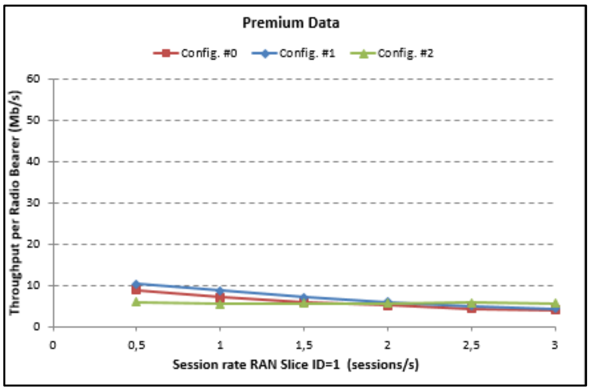
Figure 8. Blocking rate with edited simulation 2

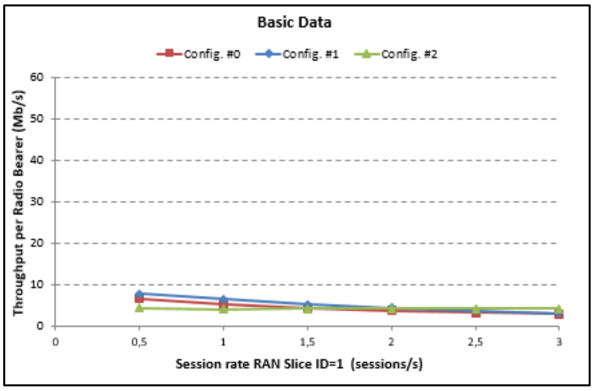
In Figure 8 it can be seen how MC Video does not suffer any blocking rate with any configuration because is the service together with MC PTT with the highest priority and there are few sessions (10%) of it, making that the number of available PRBs are enough for all the sessions.

For Premium and Basic Video, with the Configurations #1 and #2, the value of the blocking rate is the same as in the original simulation when the generation rate of the Slice 1 was 2 sessions/s. It is almost constant with these two configurations because they do distinction among slices and the Slice 1 has always the same load.

With the Configuration #0, the increase of the blocking rate as the load increases is almost insignificant in the case of Premium Video service. This is because the amount of PRBs needed for the two services with more priority than it (MC Video and MC PTT) is small so there are so available PRBs to get almost a blocking rate of 0%. In the case of Basic Video the increment is a bit higher because it has a higher value of ARP, less priority, so there are less available PRBs for it.

Non-GBR services:





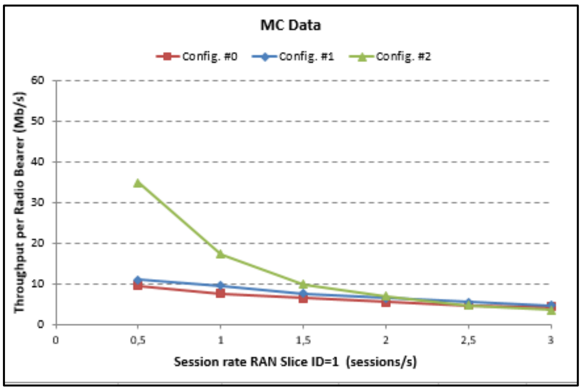


Figure 9. Average throughput with edited simulation 2

The throughput of Premium and Basic Video is practically the same with the three configurations. Configuration #2 is constant because ensures a minimum number of PRBs for each slice but the other two configurations suffer a little decrease. The reason of having a throughput so low is because we need a lot of PRBs for the GBR services of Slice 1 and there are less available for these non-GBR services. As with the GBR services, the value of the throughput is the same we had in the original simulation when the load of Slice 1 was 2 sessions/s.

However, MC Data service has a high throughput with Configuration #2 when the load is small because the amount of PRBs need it for MC Video and MC PTT is small so the number of available PRBs is high. With the other two configurations, which not make distinction among slices, the evolution is very similar to the services of Slice 1 because all have to share the same amount of PRBs following the ratio of the 5QI value.

Assessment:

In conclusion, we can say that this change on the service configuration is profitable for the GBR services because the blocking rate decreases since the PRBs needed for the RAN Slice 2 services is small, including for high session load values. However, for the non-GBR services the performance is worse, especially for the RAN Slice 1 services. The reason is that Slice 1 has always the same session load (2 sessions/s) and with this load the amount of PRBs that are used by GBR services is high. MC Data, but, has more throughput with Configuration #2 when the load is low because are more free PRBs to assign.

3.3 Quantitative evaluation

In this section the objective is proved the results obtained with the simulations doing a quantitative evaluation of them, computing the same parameters. The results are not going to be exact because in the simulation there is an important part of randomness (e.g. the services generate sessions following a Poisson process) but it has to have some similarity.

The first step is compute the general parameters of the simulation, which are shown in Table 5:

General Parameters	**Calculation**	**Value**
Bit rate, B	PRBs x nº subcarriers x subcarrier separation = 275 x 12 x 30 kHz	99 MHz
Maximum Spectral efficiency, S	Obtained from the paper [3]	8.8 b/s/Hz
Average Spectral efficiency, $\overline{S}$	Obtained from the simulations	5.7 b/s/Hz
Average bit rate, $\overline{Rb}$	B x $\overline{S}$	564.3 Mb/s
Session duration	Obtained from the paper	120 s

Table 6. Computation of general parameters

Specifically, we will try to prove the results obtained with the edited simulation 1 (3.2.1).

3.3.1 Blocking rate

I start computing the blocking rate of the Premium Video service. I study the cases that the load of RAN Slice 1 goes from 0.5 to 3 session/s with steps of 0.5. The first thing is compute what is the bit rate needed for each service depending on the load value. Also we compute the needed for the two services that have more priority, MC Video and MC PTT, because depending the configuration we will need it.

RAN Slice	**Services**	**Service mix**	**Load**	**λ**	**Θ=λ·T**	**$\Delta\rho$ = Θ·GFBR**
1	Premium Video	25%	0.5 session/s	0.125	15	150 Mb/s
			1 session/s	0.25	30	300 Mb/s
			1.5 sessions/s	0.375	45	450 Mb/s
			2 sessions/s	0.5	60	600 Mb/s
			2.5 sessions/s	0.625	75	750 Mb/s
			3 sessions/s	0.75	90	900 Mb/s
2	MC Video	20%	2 sessions/s	0.4	48	96 Mb/s
	MC PTT	40%		0.8	96	960 Kb/s

Table 7. Computation of the bit rate needed for Premium Video, MC Video and MC PTT

Θ is the traffic offered to the cell, λ is the total call rate of the service per second (number of UEs).

Depending on the configuration, the maximum bit rate that can be used for the GBR DRBs is different. Configuration #0 shares the same limitation of 70% for the two slices while Configurations #1 and #2 restrict a 50% for RAN Slice 1 and 20% for RAN Slice 2. In the next table this is summarized with the values.

	Configuration #0	Configuration #1	Configuration #2
Maximum for Slice 1	395.01 Mb/s	282.15 Mb/s	
Maximum for Slice 2	395.01 Mb/s	112.8 Mb/s	

Table 8. Limitation of bit rate for GBR services of each configuration

With that values we can compute the value of the blocking rate for Premium Video at every moment. It is computed using the following equation:

$$blocking\ rate = \frac{\rho_{occ}(ARPi,s) + \Delta\rho(GFBRi) - \rho_{max}(s)}{\Delta\rho(GFBRi)} x100$$

The parameters of the equation are measured in Mb/s. This equation does not differentiate between slices with Configuration #0 but it does with the other two. How Configurations #1 and #2 have the same restriction for a GBR service, the value of the blocking rate will be the same.

Service	Load	$\Delta\rho$ = Θ·GFBR	Blocking rate with conf. #0	Real value	Blocking rate with conf. #1 and #2	Real value
Premium Video	0.5 session/s	150 Mb/s	0%	0%	0%	0%
	1 session/s	300 Mb/s	1%	19%	6%	21%
	1.5 sessions/s	450 Mb/s	33.8%	41%	37.3%	42%
	2 sessions/s	600 Mb/s	50.3%	54%	53%	57%
	2.5 sessions/s	750 Mb/s	60.2%	62%	62.4%	63%
	3 sessions/s	900 Mb/s	66.9%	68%	68.7%	69%

Table 9. Blocking rate computation of Premium Video

As we expected, although there are a bit differences between the values obtained with the simulator and the computed we can take them for valid. As the load increases the values obtained are more similar. The only value that is very difference is with a load of 1 session/s because in the simulation is possible that more sessions were generated than the ones we contemplate on the calculations.

With all the parameters we have, we can also compute the blocking rate of MC Video. If we look what this service needs and what it can be used, the needed is always smaller than the other. Therefore, with the three configurations we have a blocking rate of 0%. However, if we look the simulator results (Figure 5) we see how appears a small blocking rate for Configurations #1 and #2 because in the simulation this 20% is not enough to initiate all the sessions.

The next step would be compute the same for Basic Video. We do the same steps and use the same equation to compute the blocking rate. The difference now if that we take into account all the PRBs that are used by Premium Video, assuming that is part of the parameter $\Delta\rho(GFBRi)$ because if we put it as $\rho_{occ}(ARPi,s)$ the blocking rate of Basic Video would be 100% for almost all the loads. The tables summarizing the calculations are below:

RAN Slice	Services	Service mix	Load	λ	Θ=λ·T	$\Delta\rho$ = Θ·GFBR
1	Basic Video	25%	0.5 session/s	0.125	15	15 Mb/s
			1 session/s	0.25	30	30 Mb/s
			1.5 sessions/s	0.375	45	45 Mb/s
			2 sessions/s	0.5	60	60 Mb/s
			2.5 sessions/s	0.625	75	75 Mb/s
			3 sessions/s	0.75	90	90 Mb/s

Table 10. Computation of the bit rate needed for Basic Video

The restrictions of each configuration continue being the same as the shown in Table 7.

Service	Load	$\Delta\rho$ = Θ·GFBR	Blocking rate with conf. #0	Real value	Blocking rate with conf. #1 and #2	Real value
Basic Video	0.5 session/s	15 Mb/s	0%	0%	0%	0%
	1 session/s	30 Mb/s	9.7%	34%	14.5%	37%
	1.5 sessions/s	45 Mb/s	39.8%	59%	43%	60%
	2 sessions/s	60 Mb/s	54.8%	69%	57.3%	70%
	2.5 sessions/s	75 Mb/s	63.9%	75%	65.8%	76%
	3 sessions/s	90 Mb/s	69.8%	80%	71.5%	80%

Table 11. Blocking rate computation of Basic Video

The results now are somewhat more distant than with Premium Video. The reason is that we are assuming something that is not true at all, because is like both Premium and Basic Video services pretend to initiate their sessions at the same time but Premium Video has more priority so their sessions are already initiated when the sessions of Basic Video are going to start. Despite this, the calculated values tend to approach the real ones as the load is higher and between the three configurations the difference is small since as we see in Figure 5 the three lines are practically equal.

3.3.2 Average throughput

When the PRBs are already assigned to the GBR services is when the remaining PRBs are distributed among the non-GBR services. It is a bit more difficult to prove theoretically the throughput per DRB that is using each service because is the last thing to do and the randomness makes that the session generation of the simulation could be very different from the theoretical. For that reason, we will compute the average throughput at some easy point.

Premium Data and Basic Data has the same configuration, the only difference is the priority that Premium has more, but both will have the same offered load. The number of UEs that each service initiate depending on the session load are:

RAN Slice	Services	Service mix	Load	λ	Activity factor	Θ=λ·T·activity factor
1	Premium Data and Basic Data	25%	0.5 session/s	0.125	0.2	3
			1 session/s	0.25		6
			1.5 sessions/s	0.375		9
			2 sessions/s	0.5		12
			2.5 sessions/s	0.625		15
			3 sessions/s	0.75		18
2	MC Data	40%	2 sessions/s	0.8	0.2	19.2

Table 12. Number of UEs initiated by the non-GBR services

Now we will compute some values of some moments of the simulations. The first hypothesis that we do is that GBR services uses the maximum PRBs they can use. This means that in each configuration remains the total bit rate minus the limitations of the Table 7, but configuration #2 guarantees a number of PRBs per slice, so the remaining for each configuration is the following:

	Configuration #0	Configuration #1	Configuration #2
RAN Slice 1	169.29 Mb/s	169.29 Mb/s	112.86 Mb/s
RAN Slice 2			56.43 Mb/s

Table 13. Remaining bit rate of each configuration

We assume that this situation happens with a load of 3 sessions/s. In that moment there are 18 UEs of both Premium and Basic Data and 19.2 of MC Data. To obtain the throughput per PRB we divided the available bit rate by the number of UEs. To know how are distributed among the three services we have to look the value of the parameter 5QI of Table 1, which match the ratio between these services (i.e. between Basic and Premium Data the relation is 60/80). Therefore, the results are shown in Table 14:

	Conf. #0	Real value	Conf. #1	Real value	Conf. #2	Real value
Premium Data	3.1 Mb/s	3.03 Mb/s	3.1 Mb/s	3.4 Mb/s	3.13 Mb/s	3.18 Mb/s
Basic Data	2.3 Mb/s	2.27 Mb/s	2.3 Mb/s	2.56 Mb/s	2.35 Mb/s	2.4 Mb/s
MC Data	3.35 Mb/s	3.3 Mb/s	3.35 Mb/s	3.6 Mb/s	2.94 Mb/s	4.17 Mb/s

Table 14. Average throughput per UE of Premium Data, Basic Data and MC Data

The results obtained are almost the same as the simulation results so we can validate the results. The only value that differs more from the real is the one of MC Data using the configuration #2. This happens because in the simulation the GBR services of slice 2 do not arrive to use all the PRBs available for them so there is more remaining bit rate than what we have contemplated theoretically.

We have achieved the marked objectives and we can validate the results obtained in the simulation.

3.4 Simulation adding a protest (no mobility)

In this section we want to study the impact of adding, in a period of time, an increment of the generation of sessions. To do it, we suppose that there is a protest of almost one hour in the middle of the simulation without taking into account the mobility of the sessions. Specifically, the protest is added between the time 9000 and 12000s of the simulation. We assume that the people in the protest only use services of the Slice 1 and the session generation rate of the protest is 4 sessions/s. The service configuration used it is the same as the Edited simulation 1 (Table 4).

In order to see and understand clearly the impact of the protest, the results are presented in another way. Every 500 seconds it is computed the average of PRBs assigned to each service and then it is plotted in a graph, taking the values every 1000 seconds to not have too many bars. When we completed the simulation, we obtained a result for each configuration and each different session rate but we have chosen the most significant results.

The results below (Figure 9) belong to the simulation using the configuration #2 (PS) and a load of 1 session/s. Between No Slicing and AC Configurations there are almost no difference but with Configuration #2 (PS) there is some change. There are more results with other configurations and other session loads in the Appendices (2). Although a service mix it is defined we must be taken into account that the generation of sessions follow a Poisson process, with what there is something of randomness.

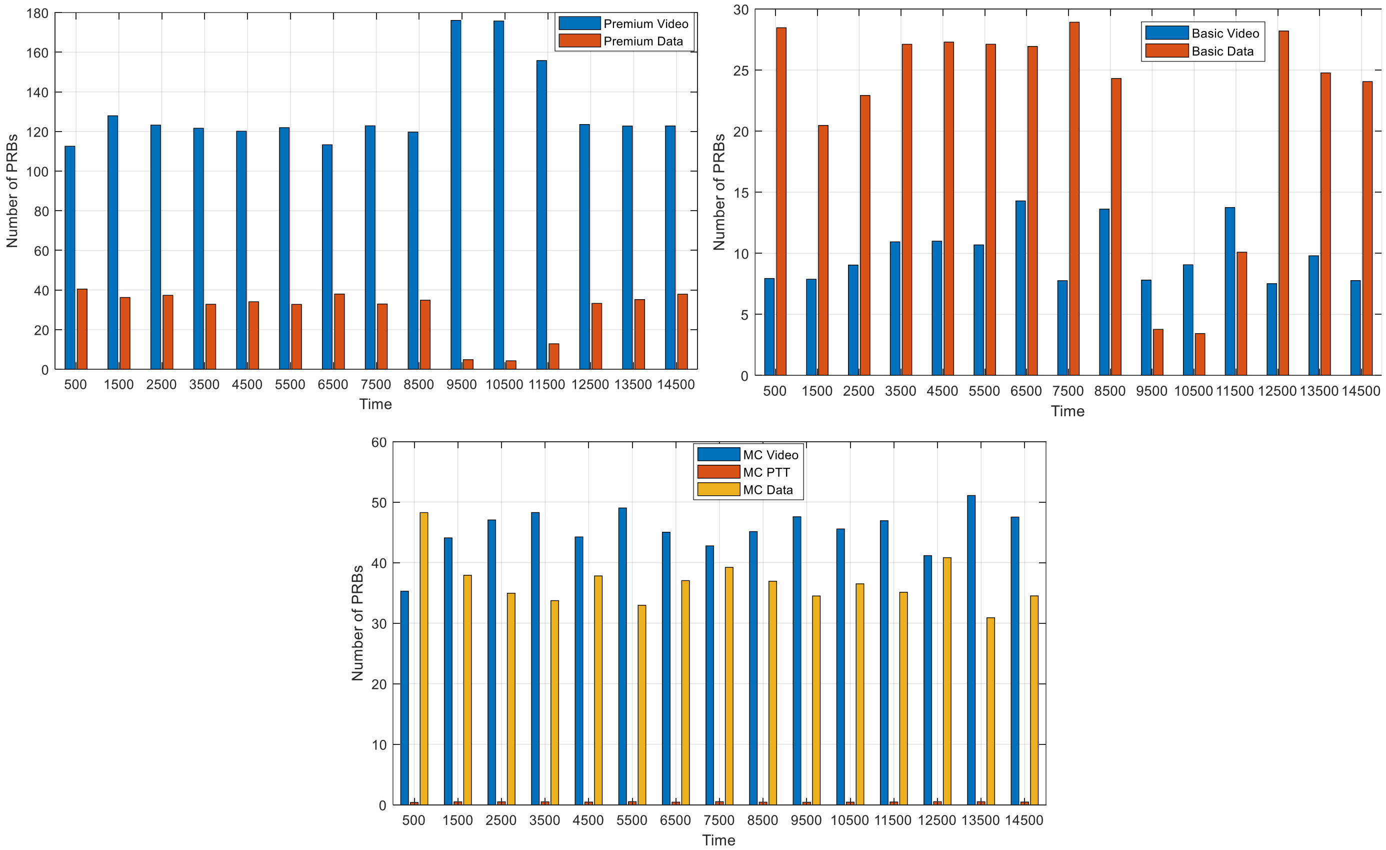


Figure 10. Assignment of PRBs with protest appearance (PS configuration)

In Figure 10 we can see clearly that when the protest appears there is a change in how the PRBs are distributed among the services, mainly of the Slice 1. The amount of PRBs assigned to the Premium Video increases considerably causing a decrease in another services for many reasons:

- When the protest starts, the generation rate of the services of the Slice 1 pass from 1 to 5 sessions/s, making that the four services of this slice pretend to initiate more sessions.
- As the service mix is distributes equally, we can assume that there are more or less the same attempts to initiate sessions from the four services. However, the fact that Premium Video is the service with the highest priority, makes that the PRBs are first assigned to this service leaving a small amount available for the others. This is why Premium Video is the only service that has more PRBs assigned when the protest arrives causing a high decrease in the other services.
- The fact that the load of the normal simulation is 1 session/s helps that at the moment the protest starts, we could see the change better because there are more free PRBs to assign.

Also, we can see how the relation between the GBR and the non-GBR services is inversely proportional. When the GBR services need more PRBs it remains less for the non-GBR, but the sum of all the PRBs of the four services is more or less the same. This happens because the PRBs are first assigned to the GBR services.

If we look at the performance of the services of Slice 2, it can be seen how they do not suffer any change. The reason of this is that Configuration #2 guarantees a minimum of PRBs for each slice so they always have a minimum of PRBs. On the other hand, with the other two configurations, the arrival of the protest does provoke an impact on the Slice 2.

This impact can be seen in Figure 11, which shows the results with the Configuration #0 (No Slicing). The only service injured is MC Data because MC Video and MC PTT are defined with ARP equals to one, highest priority, so after assigning the PRBs to these two services is when the services of the protest are taken into account. The PRBs are assigned to MC Data at the same moment of Premium and Basic Data. Although MC Data is who has lowest 5QI value of the non-GBR services, there are more sessions of the Slice 1 services and few available PRBs to distribute, so MC Data is strongly affected.

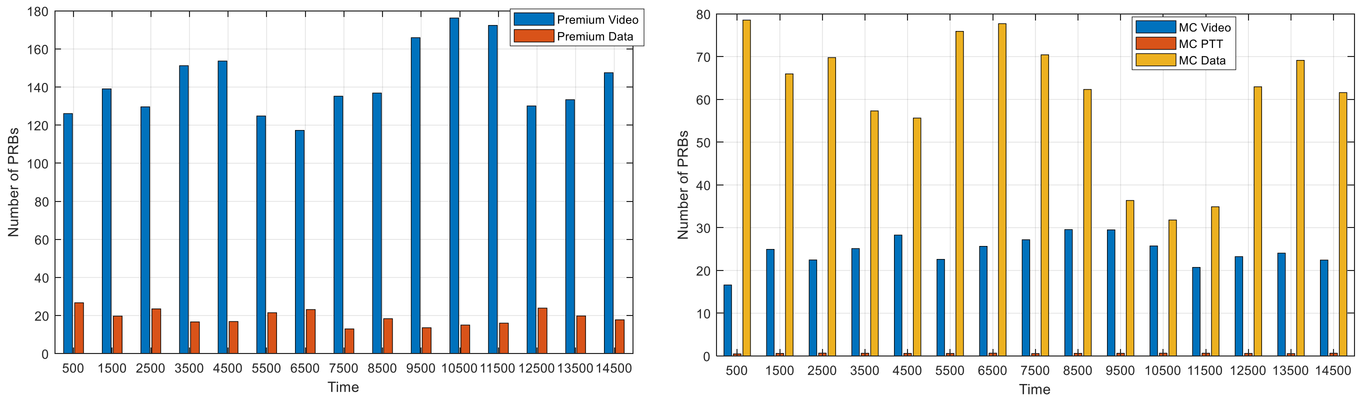


Figure 11. Assignment of PRBs with protest appearance (No Slicing configuration)

3.5 Simulation adding a bus (mobility)

The next step is study the impact of an increment of the sessions having into consideration the mobility of the sessions, the Handover. This means that a session that it is initiated in another cell and arrives to our cell, the first thing to do is assigned the necessary to not interrupt the call.

In this case, we assume that in an instant of the simulation, between 5000 and 5600s (10 minutes), arrives a bus plenty of calls. As before, all the services in the bus are from Slice 1 and the session generation rate is 4 sessions/s.

I assume a hard Handover, which means that at every moment the mobile is only connected to one base. To compute the rate of Handovers I suppose the following considerations:

- The bus has a constant velocity of v = 60 km/h = 16.67 m/s
- Cell radio: R = 115 m
- Duration of the sessions: T_m=120 s
- Remaining session duration of the sessions initiated in another cell: 50 s

The steps to compute the rate are [13]:

$$\alpha_m = \frac{2R}{vT_m} = 0{,}115 \rightarrow P_h = \frac{1 - e^{-\alpha_m}}{\alpha_m} = 0{,}944 \rightarrow \lambda_h \approx \frac{P_h}{1 - P_h}\lambda \approx \mathbf{68\ sessions/s}$$

Therefore, when the bus arrives first are initiated these 68 sessions of 50 seconds of duration and then all the sessions that are starting in the bus (4 sessions/s) and outside (it is variable depending on the simulation) with a duration of 120s.

In order to see the results clearer I have changed the service mix with respect to the previous simulations. I give more percentage to the Premium Video service, which is the service of Slice 1 with the highest priority, to see how the network can arrive to get congested. The configuration service is shown in Table 5:

RAN Slice ID	Service	ARP	Service mix
1	Premium – Video HD	2	40%
	Premium – Data	2	20%
	Basic – Video	3	20%
	Basic – Data	3	20%
2	MC Video	1	20%
	MC PTT	1	40%
	MC Data	3	40%

Table 15. Service configuration with bus appearance

The results are shown in the same way as the last section (3.4), computing the amount of PRBs that are assigned to each service. The difference now is the time step that we use to compute the average, 100 seconds, since the bus stays only 600 seconds and it is preferable to have a smaller time step to know more information while the bus is in the cell. The part of the simulation that can be seen in the results is from 4400 to 6400 seconds.

The next figure, Figure 12, corresponds to the simulation do it with Configuration #1 (AC) and the case that the normal simulation has a load of 1 session/s.

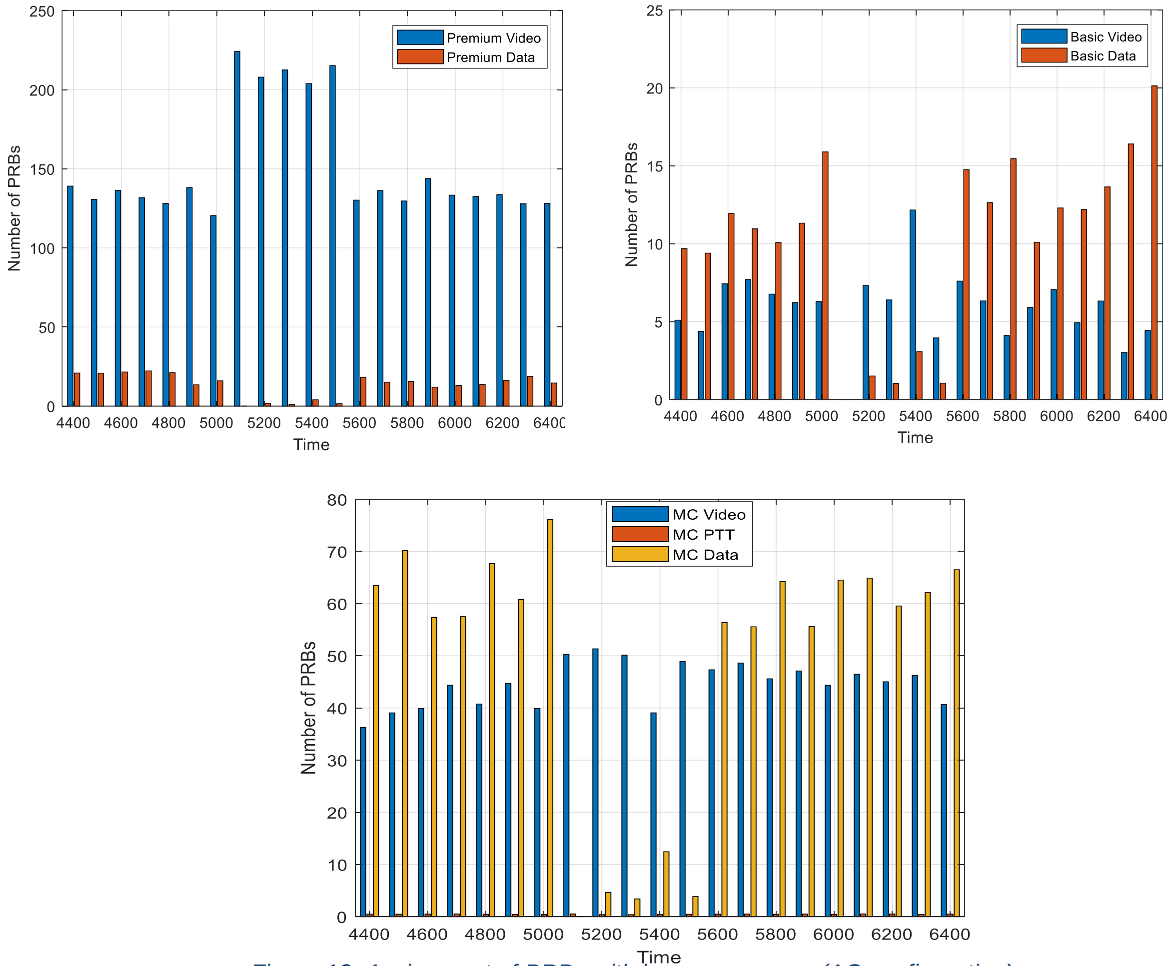


Figure 12. Assignment of PRBs with bus appearance (AC configuration)

The appearance of the bus has a big impact on the performance of each service. The amount of resources that it is needed for the Handover sessions of the Premium Video causes that the other services do not receive any PRB, only MC Video and MC PTT because have more priority. This happens because these Handover sessions skip the admission control of the RAC parameter restricting in a 50% the maximum % of PRBs for GBR bearers of Slice 1, and so they can use a high number of PRBs. Then, when the Handover sessions end but the bus still continue in the cell, it can be seen how some PRBs are assigned to all the services but in less quantity because most of the PRBs are needed for Premium Video service.

In the instant 5400s, it is appreciated how Basic Video has more PRBs assigned than in the rest of the simulation. The reasons are that in this moment we can see how MC Video uses less PRBs, remaining more available PRBs to assign to services with a higher ARP. In addition, Basic Video has suffered an increase of its sessions with the bus appearance so if there are more available PRBs they will be used by Basic Video, overcoming the PRBs use it in the rest of the simulation.

If we study the results obtained with Configuration #0 (No Slicing), Figure 13, we see that they are a bit different. Now, the moment that Premium Video needs more PRBs is not initializing the Handover's sessions but when the bus is about to leave. The reason is because in this simulation the randomness of generating sessions create more Premium Video sessions in that moment than with AC configuration, reaching a point of congestion that all the PRBs are used between MC Video, MC Data and Premium Data.

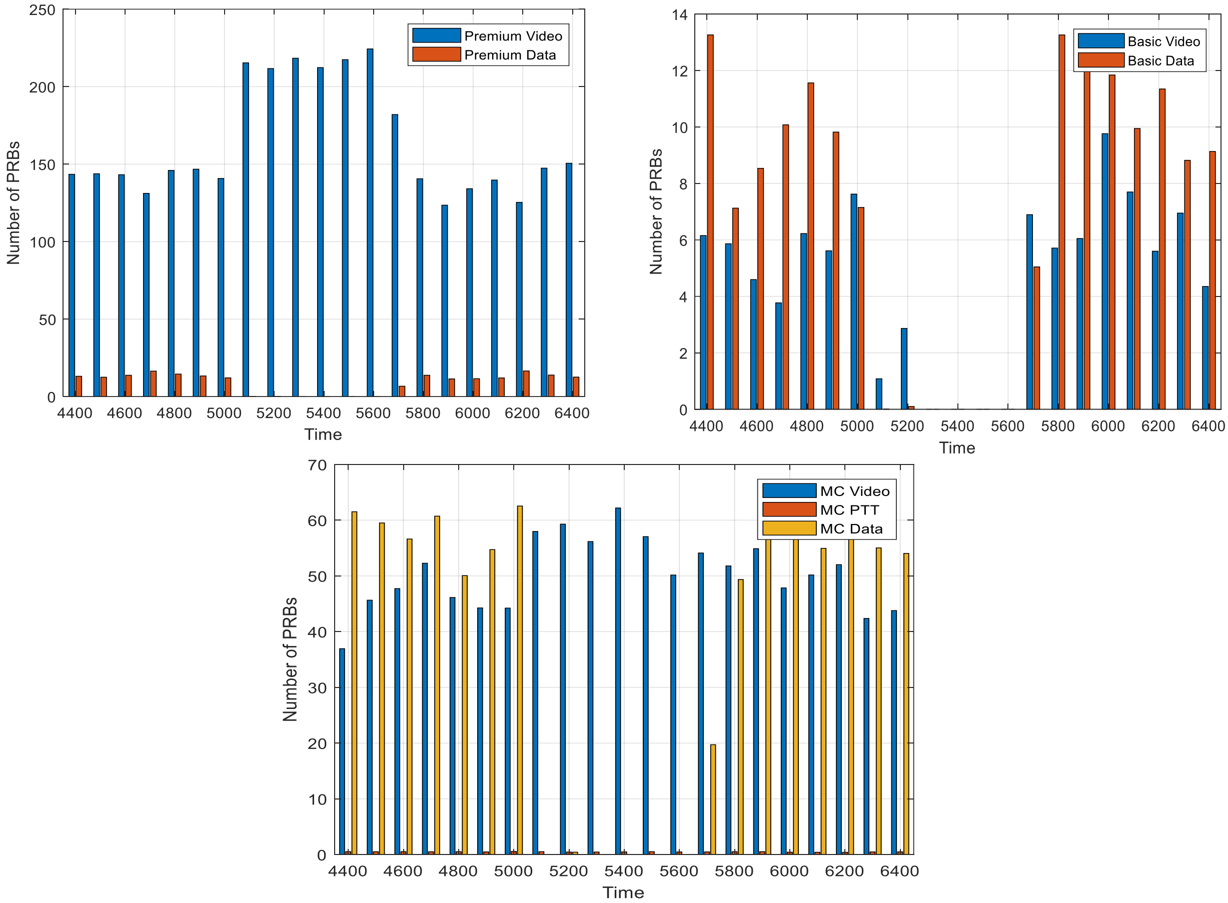


Figure 13. Assignment of PRBs with bus appearance (No Slicing configuration)

With these two configurations we see how MC Data is harmed a lot when the bus is in the cell. However, with the PS configuration, it is clearly seen how the fact of guaranteeing a minimum % of PRBs for each slice achieves that MC Data continue having the same number of PRBs when the bus arrives. It is shown in Figure 13.

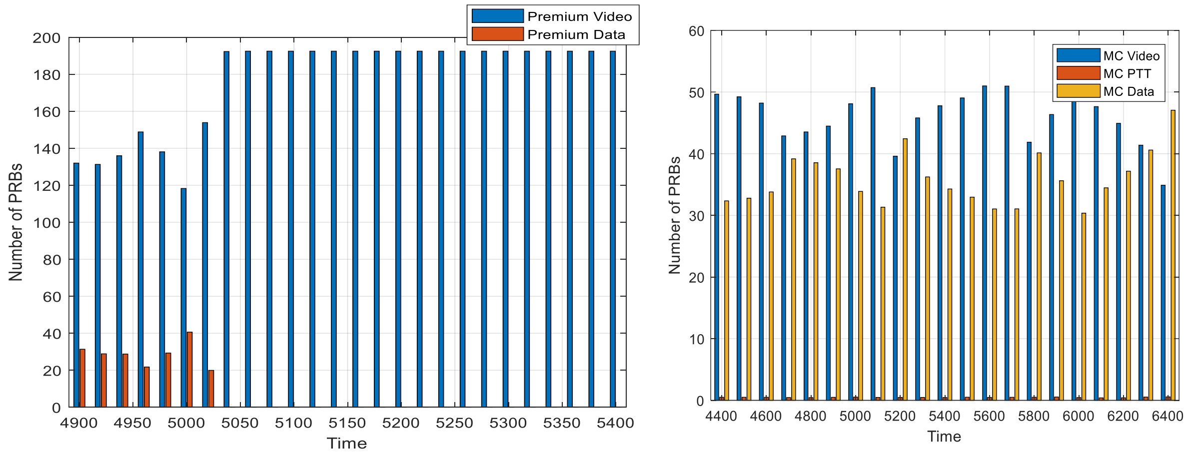


Figure 14. Assignment of PRBs with bus appearance (PS configuration)

With the 30% of PRBs guaranteed for Slice 2, RAN Slice 1 only can use the 70% remaining, that is 192.5 PRBs. In Figure 14.a it can be seen clearly how Premium Video uses all these 192.5 PRBs so the rest of services of this slice has a blocking rate of 100%. It can be appreciated how this graph has a smaller time step than the others because we want to emphasize how when the bus arrives Premium Video is using all the PRBs possible at every moment.

<u>Assessment:</u>

Adding the bus we see how the network can get to collapse, arriving to no assign any PRB to some services. To avoid that the increase of services of one slice does not affect the other slice, it is needed the PS parameter (Configuration #2). However, with this configuration the services with less priority of the same slice would be so disadvantaged because it will be very difficult than in a moment with the bus, Premium Video needs minus than the maximum available PRBs (192.5). With AC Configuration, if in a moment the process initiates less Premium Video sessions, is possible that remain some PRBs for the other services of the same slice. As the Handover sessions skip the RAC limitation, the % of the RAC parameter is not important, only if the configuration distinct among slices or not. The No Slicing configuration is with whom more PRBs are assigned to Premium Video service but at the expense of degrading the performance of the other slice. Therefore, depending on what interests us the most it will be used one configuration or another.

4. Budget

As this thesis was not about building any prototype, we only need to take into account the software cost and the amount of hours dedicated.

We used different software, but Matlab [14] is the only that has a cost, being the following:

Software license	Cost
Matlab for academic use	500€

Table 16. Total software cost [14]

I have been working on the thesis a total of 20 weeks working approximately 20 hours per week, so it took me a total of 400 hours. The cost, evaluated as a junior engineer, is the following:

Number of total hours	Cost/hour	Total cost
400 h	10 €/h	4000 €

Table 17. Total hour cost

The total cost of the project is 4500€.

5. Conclusions and future development:

5G technology would be one of the most important developments of our time changing the idea we have of mobile communications, increasing the amount of applications available to use. To carry out the 5G systems it is needed support for network slicing, specifically the realization of RAN Slicing.

We define some configuration parameters that dictate the operation of PS and RAC function at Layer 2 and Layer 3 respectively and a multi-scenario with two Slices and different services each slice. After doing some simulations with different configurations and different cases, some are in the Appendices (2), we conclude that:

- L2 control is needed to isolate the non-GBR DRBs of different slices, avoiding that an increase of services of one slice impacts negatively on the throughput of the other slice. The impact that has L2 parameter in the GBR services is insignificant.
- L3 control is important for the GBR services. Making distinction among slices (Configurations #1 and #2) ensure the isolation of the different slices, and in case of overload situation in one slice the GBR services of the other would not be affected.
- Configuration #0 (No Slicing) allows that if a service increases its services considerably, this service would be benefited having more PRBs assigned at the expense of degrading the throughput of the other slice. With the other configurations the service is more restricted in the amount of PRBs it can use.
- If we add an increase of the services of one slice in a period of time, depending how big the increase is, all the PRBs can be used by some services leaving nothing for the others. This happens because some sessions skip the admission control.
- Adding the mobility of the sessions makes that when the phenomenon starts, the amount of PRBs needed in that moment for the most priority service of the slice is so high that the others are collapsed without assigning any PRB to them. Depending the configuration there will be more services affected or less.
- Because of some aleatory processes, the simulation is needed and we cannot compute the exact results theoretically because we consider all in an ideal case. However, we can achieve an approximation of them.

Finally, even though we have done some advances in the study of the RAN Slice realization in 5G, there is still a lot of work to do in this area. Basically, the work would be focused on continue doing more simulations to understand it better, some ideas of other simulations could be:

- Change the radius of the cell to a larger one and thus have more traffic load offered and see how it affects to the results.
- Change the % of the configurations trying to get with which the increase of load of a service in a moment of time affects the minimum possible to the other services.
- Add another cell and thus take into account the average PRB utilisation of each slice and the aggregate bit rate of each tenant at multi cell level.
- Create a new scenario with another slices and different services.

Bibliography:

[1] 5G-PPP [Online] Available: https://5g-ppp.eu/ [Accessed: 14-Jun-2018]

[2] 3GPP TS 22.261 v15.0.0, "Service requirements for the 5G system; Stage 1 (Release 15)", March, 2017

[3] R. Ferrús, O. Sallent, J. Pérez-Romero, R. Agustí, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration", IEEE Communications Magazine, September, 2017.

[4] J. Pérez-Romero, O. Sallent, R. Ferrús, R. Agustí, "On the Configuration of Radio Resource Management in a Sliced RAN", 2018.

[5] Universidad Internacional de Valencia, "Evolución de la red de comunicación móvil, del 1G al 5G", [Online] Available: https://www.universidadviu.es/evolucion-la-red-comunicacion-movil-del-1g-al-5g/ [Accessed: 18-Jun-2018]

[6] J. Pérez-Romero, O. Sallent, F.Casadevall "RCOM Tema 1 Introducción a las comunicaciones móviles", ETSETB, UPC, 2018.

[7] Nokia, "5G uses Cases" [Online] Available: https://www.ramonmillan.com/documentos/bibliografia/5GUseCases_Nokia.pdf

[8] 5G.co.uk "What is Network Slicing?" [Online] Available: https://5g.co.uk/guides/what-is-network-slicing/ [Accessed: 18-Jun-2018]

[9] TechTarget, "Control plane (CP)" Online, Available:
https://searchsdn.techtarget.com/definition/control-plane-CP [Accessed: 22-Jun-2018]

[10] Dialogic, "User Plane" [Online] Available: https://www.dialogic.com/glossary/user-plane [Accessed: 22-Jun-2018]

[11] I. Leonardo da Silva, G. Mildh, A. Trogolo, E. Buracchini, P. Spapis, A. Kaloxylos, G. Zimmemann, N. Bayer, "On the impact of network slicing on 5G radio Access networks", IEEE Communications Magazines, September, 2016

[12] 3GPP, "Mission Critical Services in 3GPP", [Online] Available: http://www.3gpp.org/NEWS-EVENTS/3GPP-NEWS/1875-MC_SERVICES [Accessed: 24-Jun-2018]

[13] J. Pérez-Romero, O. Sallent, F.Casadevall "RCOM Tema 5: Sistemas celulares", ETSETB, UPC, 2018.

[14] MathWorks, “Pricing and Licensing”, [Online], Available: https://es.mathworks.com/ pricing-licensing.html?intendeduse=edu&s_tid=htb_learn_ gtwy_ct a2 [Accessed: 30-Jun-2018]

[15] 3GPP TR 36.814 v9.0.0, “E-UTRA; Further advancements for E-UTRA physical layer aspects (Release 9)”, March 2010

Appendices:

1. Work Packages

Project: 5G Radio Access Network Slicing	WP ref: 1	
Major constituent: Read documentation	Sheet 1 of 7	
Short description: Read the necessary documentation to understand the context in which we are going to work.	Planned start date: 8/02/18 Planned end date: 28/02/18 Start event: 8/02/18 End event: 28/02/18	
Internal task T1: Read both papers Internal task T2: Define how we are going to focus the project	Deliverables:	Dates:

Project: 5G Radio Access Network Slicing	WP ref: 2	
Major constituent: Simulation	Sheet 2 of 7	
Short description: Work with the simulator and get some practice in order to get used to it.	Planned start date: 1/03/18 Planned end date: 11/03/18 Start event: 1/03/18 End event: 13/03/18	
Internal task T1: Get used to the simulator	Deliverables:	Dates:

Project: 5G Radio Access Network Slicing	WP ref: 3	
Major constituent: Simulation	Sheet 3 of 7	
Short description: Do some changes in the parameters related to how we define the service mix of the services and the priority of each service.	Planned start date: 14/03/18 Planned end date: 29/03/18 Start event: End event:	
Internal task T1: Change the parameters and do the simulation Internal task T2: Study the results	Deliverables:	Dates:

Project: 5G Radio Access Network Slicing	WP ref: 4	
Major constituent: Quantitative evaluation	Sheet 4 of 7	
Short description: Compute some parameters and statistics to validate the results obtained with the simulations.	Planned start date: 2/04/18 Planned end date: 11/04/18 Start event: End event:	
Internal task T1: Compute some parameters Internal task T2: Compare with the simulation results	Deliverables:	Dates:

Project: 5G Radio Access Network Slicing	WP ref: 5	
Major constituent: Simulation (add protest appearance)	Sheet 5 of 7	
Short description: Add the appearance of a protest in a long moment of the simulation. The protest increases the demand of services so I want to observe what happen in this moment of the simulation.	Planned start date: 12/04/18 Planned end date: 4/05/18 Start event: End event:	
Internal task T1: Add the protest appearance Internal task T2: Do simulations with different configurations Internal task T3: Study the results	Deliverables:	Dates:

Project: 5G Radio Access Network Slicing	WP ref: 6	
Major constituent: Simulation (add bus appearance)	Sheet 6 of 7	
Short description: Add the appearance of a bus in some moments of the simulation. Adding the bus we have to take into account the mobility of the bus, adding the sessions that are initiated in another cell and are not finished (Handovers).	Planned start date: 5/05/18 Planned end date: 2/06/18 Start event: End event:	
Internal task T1: Add the bus appearance with the rate of handovers Internal task T2: Do simulations with different configurations Internal task T3: Study the results	Deliverables:	Dates:

Project: 5G Radio Access Network Slicing	WP ref: 7	
Major constituent: Conclusions and report	Sheet 7 of 7	
Short description: Collect all the results we obtained and write the final report of the project.	Planned start date: 3/06/18 Planned end date: 2/07/18 Start event: End event:	
Internal task T1: Arrive to some conclusions Internal task T2: Write the final report	Deliverables:	Dates:

2. Simulations

Apart from the simulations we explained in the principal part of the thesis we have done more simulations that helped us to arrive to the conclusions. Here we will explain briefly some of the other simulations.

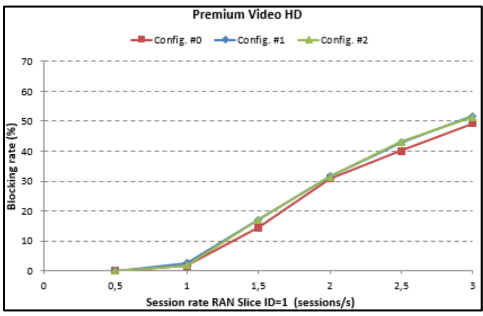
2.1 Simulation 1

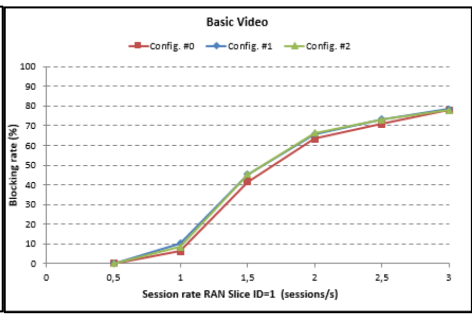
After doing the edited simulation 1 (3.2.1), we changed the service mix of RAN Slice 1 giving less rate to both Premium Services. In the next table it is summarized:

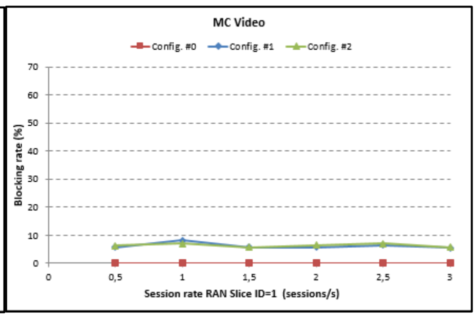
RAN Slice ID	Service	ARP	Service mix
1	Premium – Video HD	2	15%
	Premium – Data	2	15%
	Basic – Video	3	35%
	Basic – Data	3	35%
2	MC Video	1	20%
	MC PTT	1	40%
	MC Data	3	40%

The results are below:

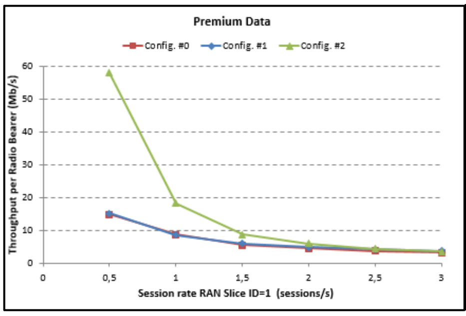
<u>GBR Services:</u>

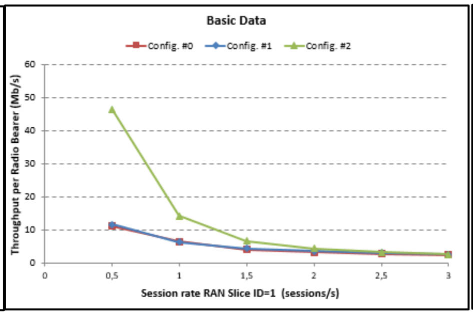


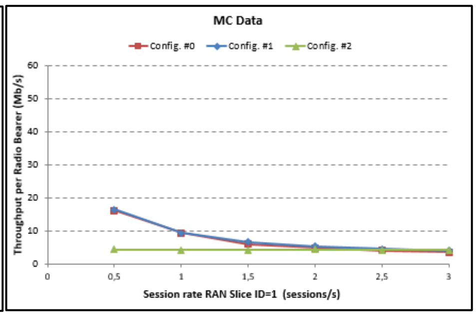




Non-GBR Services:







Comments:

The blocking rate of Premium Video decreases compared with the results obtained in 3.2.1 because now there are less sessions of this service by how we have defined the service mix. We do not see any difference between the three configurations because the amount of PRBs need it for MC Video causes that with Configuration #0 the blocking rate of Video services of Slice 1 increases.

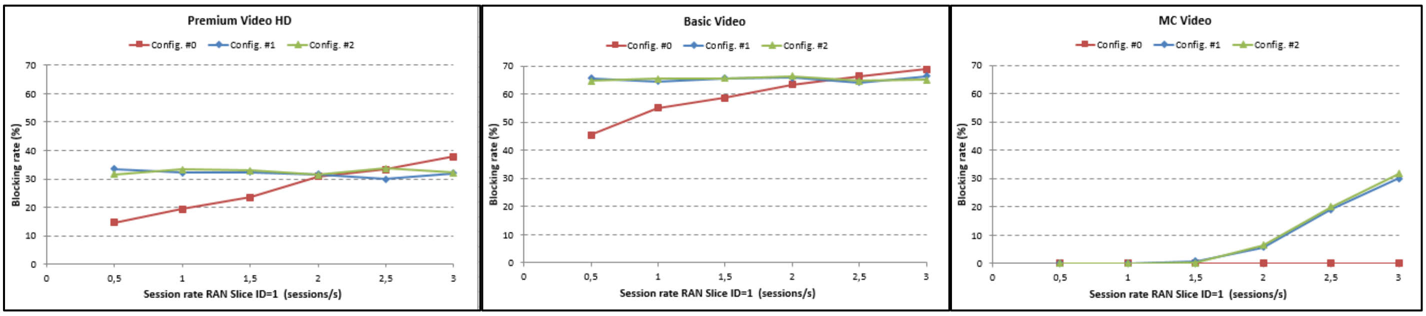
For the non-GBR services happen the same as the GBR, because we have less sessions of Premium Data the throughput is a bit higher for this service.

2.2 Simulation 2

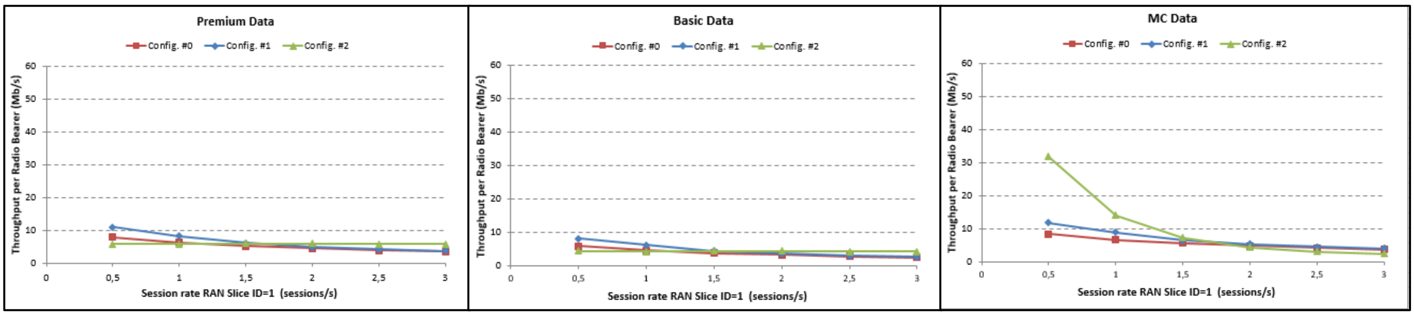
With the same configuration of service mix as before, we changed the average session generation rate setting 2 sessions/s for RAN Slice 1 and RAN Slice varies from 0.5 to 3 sessions/s.

RAN Slice	Service	ARP	Service mix	Average session generation rate
1	Premium – Video HD	2	15%	2 sessions/s
	Premium – Data	2	15%	
	Basic – Video	3	35%	
	Basic – Data	3	35%	
2	MC Video	1	20%	Varied from 0.5 to 3 sessions/s
	MC PTT	1	40%	
	MC Data	3	40%	

GBR Services:



Non-GBR Services:



Comments:

Compared with the other simulations, the novelty of this simulation is the evolution of the blocking rate of Premium Video and Basic Video with Configuration #0. As the session load increases the blocking rate increases too, going from having a lower value than Configurations #1 and #2 to a higher one. The explanation is the increase of PRBs needed for MC Video as the load increases, remaining less for the other services and consequently increasing the blocking rate.

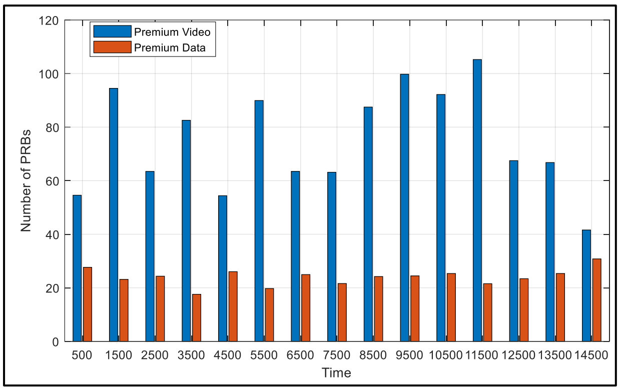
2.3 Simulation 3

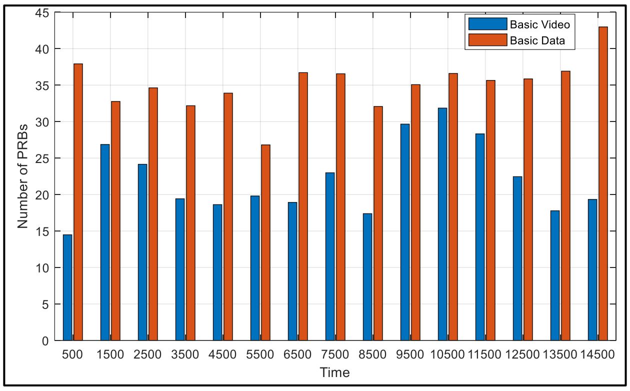
From the section of the protest appearance we also did a simulation with the original configuration and adding the protest in the same interval of time. The configuration is shown in the next table:

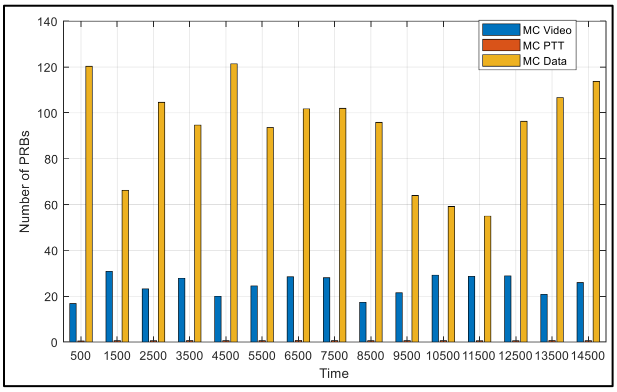
RAN Slice ID	Service	ARP	Service mix	Average session generation rate
1	Premium – Video HD	2	10%	Varied from 0.5 to 3 sessions/s
	Premium – Data	2	20%	
	Basic – Video	3	30%	
	Basic – Data	3	40%	
2	MC Video	2	10%	2 sessions/s
	MC PTT	1	50%	
	MC Data	3	40%	

The results are the following:

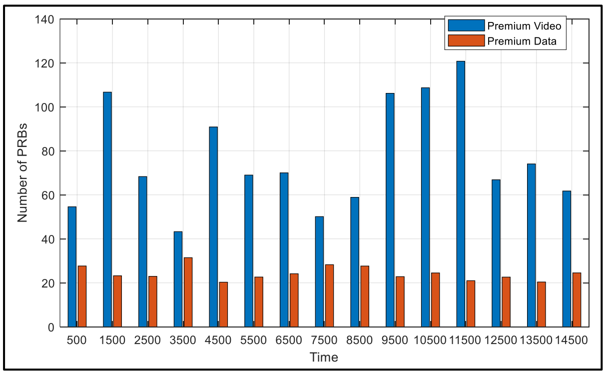
Load 1 session/s and Configuration #0:

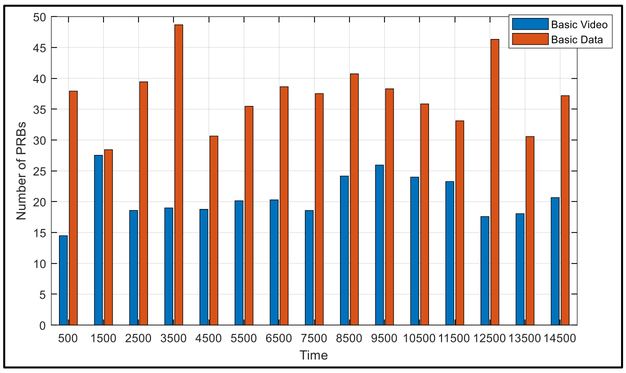


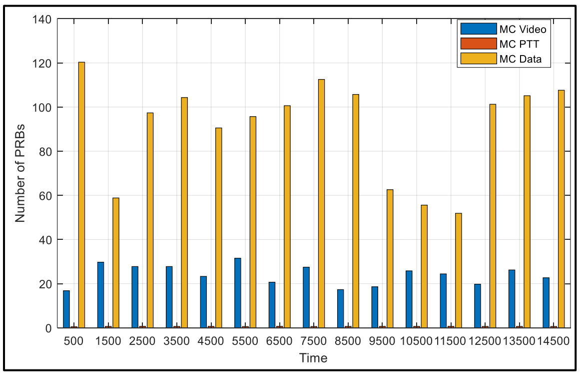




Load 1 session/s and Configuration #1:





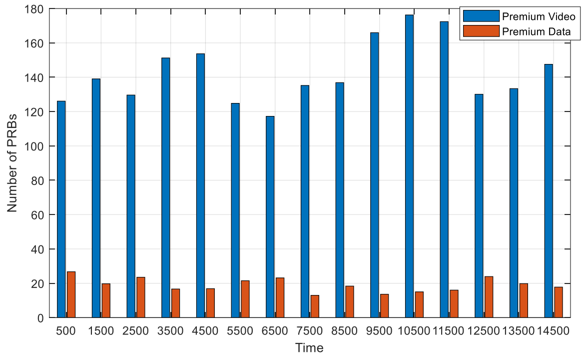


With both configurations we can see something similar. When the protest arrives there is an increase of the amount of PRBs assigned to Premium Video. As now Premium Video is only a 10% of all the sessions of Slice 1, it does not use all the available PRBs for GBR services so there is also an increase of Basic Video PRBs. The only service harmed is MC Data. The reason is because after distributing all the PRBs among the GBR services the available PRBs have to be distributed among the non-GBR services following the 5QI value. However, as the number of sessions of Premium Data and Basic Data are much more than MC Data because they are generated in the protest, so finally the two services of Slice 1 received the same throughput more or less than before the protest but MC Data suffers a decrease in its throughput.

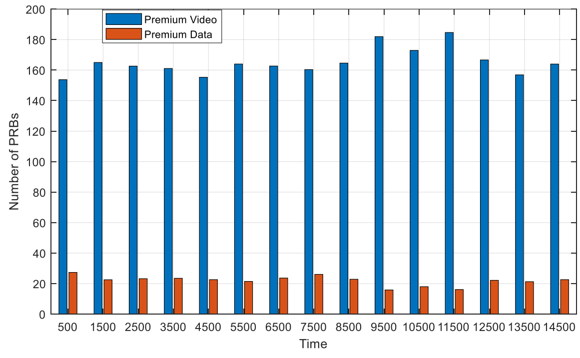
2.4 Simulation 4

It is also interesting to see how as the session load is higher the effect of the protest is smaller. We will focus on the increment of Premium Video sessions with Configuration #0.

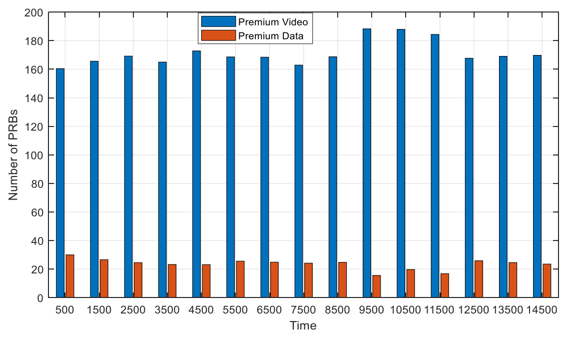
Load 1 session/s



Load 2 sessions/s



Load 3 sessions/s



3. Another way of computing Handover rate

There is another method to compute the Handover rate from the duration of the call that it has been taken in the cell (t_H).

The steps are the following:

$$E[t_H] = T_m(1 - P_h) \rightarrow \mathrm{Ph} = 1 - \frac{E[th]}{Tm} \rightarrow \lambda_h = \frac{P_h}{1-P_h}\lambda$$

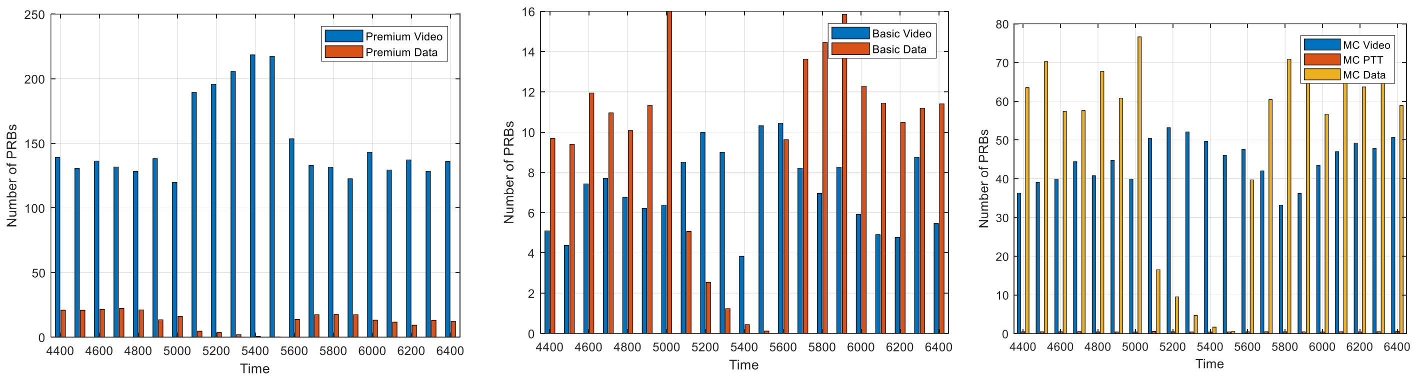
We assume some values of t_H and we compute the rate of Handover.

t_H=30s → λ_h=12 sessions/s

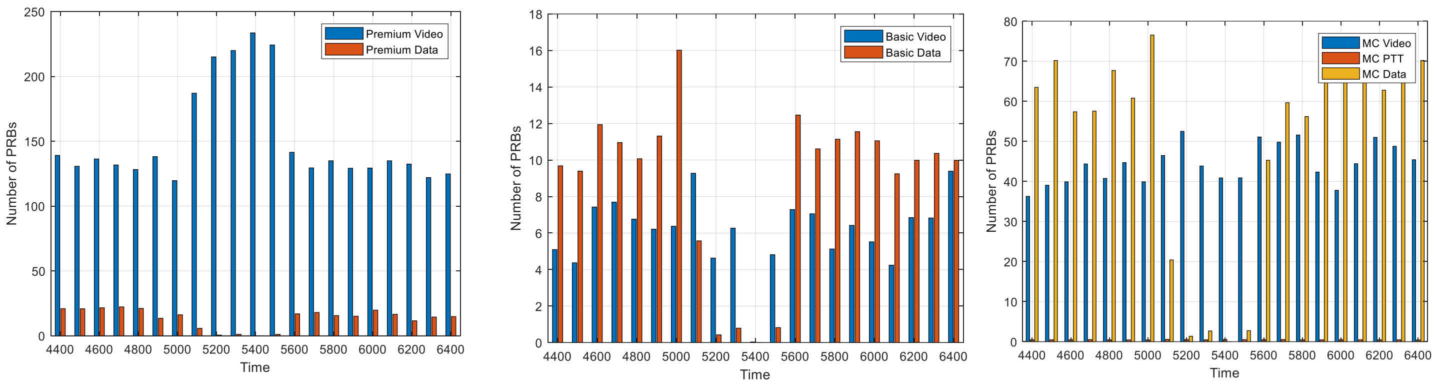
t_H=60s → λ_h=4 sessions/s

We did the same simulation as in 3.5 only changing this Handover rate and the results with Configuration #1 are the following:

<u>Remaining session duration= 30s (λ_h=12 sessions/s)</u>



<u>Remaining session duration= 60s (λ_h=4 sessions/s)</u>



It can be seen how now when the bus arrives the Handovers do not cause the same as in 3.5 because the rate is smaller, so the amount of sessions we have to initiate at the beginning do not use all the available PRBs. Nevertheless, as the bus stays a bit in the cell all the sessions generated from Premium Video also congested the network. Between these two cases the difference is practically insignificant in the moment of initializing the Handovers.

Glossary

3GPP: 3rd Generation Partnership Project

5G: 5th Generation of mobile communications

5G- PPP: 5G Public Private Partnership

5GC: 5G Core Network

5QI: 5G QoS Identifier

ARP: Allocation Retention and Priority

CDMA: Code Division Multiple Access

CMC: Connection Mobility Control

CN: Core Network

CP: Control Plane

DRB: Data Radio Bearers

EDGE: Enhanced Data Rates for GSM Evolution

eMBB: enhanced Mobile Broadband

FDMA: Frequency Division Multiple Access

GBR: Guaranteed Bit Rate

GFBR: Guaranteed Flow Bit Rate

GPRS: General Packet Radio Service

GSM: Global System for Mobile communications

HSPA: High-Speed Packet Access

ICT: Information and communication technology

IP: Internet Protocol

KPI: Key Performance Indicator

LTE: Long Term Evolution

L1: Layer 1

L2: Layer 2

L3: Layer 3

MAC: Medium Access Control

MC: Mission Critical

NG-RAN: New Generation RAN

NTT: Nippon Telegraph and Telephone

OFDMA: Orthogonal Frequency-Division Multiple Access

PLMN: Public Land Mobile Network

PRB: Physical Resource Block

PS: Public Safety

PS: Packet Scheduling

PTT: Push To Talk

QoS: Quality of Service

RAC: Radio Admission Control

RAN: Radio Access Network

RAT: Radio Access Technology

RB: Radio Bearer

RBC: Radio Bearer Control

RRC: Radio Resource Control

RRM: Radio Resource Management

SIB: System Information Block

SNR: Signal to Noise Ratio

SINR: Signal to Interference-plus-Noise Ratio

TDMA: Time Division Multiple Access

UE: User Equipment

UMTS: Universal Mobile Telecommunications System

UP: User Plane