Design of an Algorithm for Slice Handover in 5G Networks

Intra and Inter Slice Handover Approach.

Prepared by:
Tawanda Muzanenhamo
MZNTAW004

Department of Electrical Engineering
University of Cape Town

Prepared for:
Associate Professor O.E. Falowo

Department of Electrical Engineering
University of Cape Town

June 2020

Submitted to the Department of Electrical Engineering at the University of Cape Town in partial fulfilment of the academic requirements for a Bachelor of Science degree in Electrical and Computer Engineering.

Key Words: 5G, Heterogenous Networks, Network Slices, Vertical Handover

Declaration

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used the IEEE convention for citation and referencing. Each contribution to, and quotation in, this final year project report from the work(s) of other people, has been attributed and has been cited and referenced.
3. This final year project report is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as their own work or part thereof

Name: **Tawanda Muzanenhamo**

Signature: ______________________ **Date:** 05 July 2020

Terms of Reference

ID:	OF2
SUPERVISOR:	O.E. Falowo
TITLE:	Design of an algorithm for slice handover in 5G networks
DESCRIPTION:	Handoff decision algorithms are designed for making handoff decisions when a call is to be handed over from one network to the another. Vertical handoff is necessary for joint management of radio resources among different radio access technologies in heterogenous wireless networks. The objective of this project is to design an algorithm for making handoff decisions in 5G networks and evaluate its performance.
DELIVERABLES:	Literature review, vertical handoff algorithm, simulation results, simulation code, results
SKILLS/REQUIREMENTS:	MATLAB, java, or any other programming language, EEE4121F/EEE4087F.
ELO1: Problem Solving: Identify, formulate, analyse and solve complex engineering problems creatively and innovatively	(Explain what needs to be done to meet ELO requirements). The student is expected to design and implement an algorithm for making handover decisions in 5G networks.
ELO4: Investigations, Experiments and Analysis: Demonstrate competence to design and conduct investigations and experiments	The student is expected to investigate the performance of the slice handover algorithm through numerical solutions.
EXTRA INFORMATION:	For students interested in pursuing a master's degree, the project can be expanded to an MSc dissertation.
AREA:	Wireless Networks.
Project suitable for ME/ECE/EE?	ECE/EE.

Acknowledgements

I would like to express my gratitude to the following people without whom this project and the last five years would not have been possible.

First and foremost, I would like to give my thanks to the Lord Almighty for his unlimited and unbounded grace that has enabled me to start and finish my degree. Lord thank you for keeping your promises, I will forever be grateful. [Jeremiah 29:11]

To my amazing parents Mr F. Muzanenhamo and Mrs S Muzanenhamo for the continuous support and sacrifices I wouldn't have made it this far without your support. Thank you for keeping me in your daily prayers, guiding me and being there for me all the time for the past 23 years.

To my best friend Shamiso Chikamhi for encouraging me to work hard and follow my dreams and for listening to me explain the complex engineering concepts for hours without understanding any word. Thank you for taking time to read through my paper and providing useful feedback.

To Takudzwa and Alban for teaching me how to take breaks and reminding me of what I am capable of.

To my supervisor Prof Olabisi Falowo for his tireless support, motivation, extensive knowledge and guidance throughout the project. Thank you Prof for your guidance and tireless support during the research and report writing, I will forever be grateful.

To ABSA Group for financing the last 3 years of undergraduate degree.

Abstract

5G wireless networks are envisioned to meet the rising demand for network services from users. User devices have evolved and demand different services from the network. The user demands can be categorised based on latency, reliability and bandwidth required. In order to meet the diverse requirements of users in a cost-effective manner whilst ensuring network resources are efficiently allocated to users, 5G networks are expected to utilise technologies like Software Defined Networks (SDN), Network Function Virtualisation (NFV) and Network Slicing.

With the introduction of diverse 5G application scenarios, new mobility management schemes must be implemented in Sliced 5G networks in order to guarantee seamless handover between network slices. Mobility management allows users to move from one coverage area to another without losing network connection. 5G networks follow the heterogenous networks architecture meaning different network slices can co-exist with each slice providing services tailored for specific Quality of Service (QoS) demands. Therefore, when users move from one coverage area to another, the call can be handed over to a slice catering for the same demands or a slice catering for different demands.

The aim of this project is to design an algorithm for making handover decisions in sliced 5G networks and evaluate the performance of the algorithm. The chosen network model for this project consist of three slices namely Enhanced Mobile Broadband, Massive Machine Type Communication (mMTC) and Ultra Reliable Low Latency Communication (uRLLC). An analytical model based on the Markov chain is used to model the call admission control algorithm in the network model.

This report details the design of the network model and the implementation of the vertical handover decision making algorithm . The algorithm performance is evaluated using connection level QoS metrics namely ne call blocking probability and handoff call dropping probability. The simulations are carried out to determine how the QoS metrics are affected by variations in different metrics like call arrival rate, capacity, new call threshold, required basic bandwidth unit and call departure rate. From the simulation results it is concluded that in overall, the implemented algorithms provide good QoS levels with the handoff call dropping probability being less than the new call blocking probability in all scenarios and the inter slice handover algorithm provides better QoS in the eMBB and mMTC slice whilst the intra slice handover algorithm provide better QoS in the uRLLC slice.

Table of Contents

List of Figures

List of Tables

List of Listings

List of Acronyms

1G	First Generation
2G	Second Generation
3G	Third Generation
3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
5GPPP	Fifth Generation Public Private Partnership
AHP	Analytic Hierarchical Process
AMC	Adaptive Modulation and Coding
AMPS	Advanced Mobile Phone System
BBU	Basic Bandwidth Unit
BS	Base Station
CAC	Call Admission Control
CAPEX	Capital Expenditure
CDMA	Code Division Multiple Access
CN	Core Network
D2D	Device to Device
DoS	Denial of Service
EBE	Engineering and Built Environment
EDGE	Enhanced Data Rates for GSM Evolution
EGPRS	Enhanced Global Packet Radio Service
ELECTRE	Elimination et Choix Traduisant La Realité
eMBB	Enhanced Mobile Broadband
ETSI	The European Telecommunication Standard Institutes
FDMA	Frequency Division Multiple Access
GPRS	Global Packet Radio Service
GPS	Global Position System
GRA	Grey Relational Analysis
GSM	Global System for Mobile Communications
HCDP	Handoff Call Dropping Probability
HSDPA	High Speed Downlink Packet Access
HSPA	High Speed Packet Access (HSPA)
iDEN	Integrated Digital Enhancement
IMT-2000	International Mobile Telephone 2000
InPs	Infrastructure Providers
ITU	International Telecommunications Unit
JCAC	Joint Call Admission Control
JRRM	Joint Radio Resource Management
LTE	Long Term Evolution
M2M	Machine to Machine
MADM	Multiple Attribute Decision Making
MC-CDMA	Multi Career Code Division Multiple Access
MCDM	Multiple Criteria Decision Making
MEC	Multiple Edge Computing

MEW	Multiplicative Exponent Weighing
MIMO	Massive Input Massive Output
MMS	Multimedia Services
mMT	Millimetre Wave Transmission
mMTC	Massive Machine Type Communication
MNOs	Mobile Network Operators
MVNO	Mobile Virtual Network Operators
NCBP	New Call Blocking Probability
NFV	Network Function Virtualisation
Nos	Network Operators
NTT	Nippon Telegraph and Technology
OFDMA	Orthogonal Frequency Division Multiple Access
ONF	Open Network Foundation
OS	Operating System
PDC	Personal Digital Cellular
PoA	Point of Attachment
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAN	Radio Access Networks
RATs	Radio Access Technologies
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RSS	Received Signal Strength
SAW	Simple Additive Weighting
SDN	Software Defined Network
SMS	Short Message Services
SPs	Service Providers
TDMA	Time Division Multiple Access
TV	Television
UE	User Equipment
UMTS	Universal Mobile Telecommunication Service
uRLLC	Ultra-Reliable Low Latency Communication
VHO	Vertical Handover
VIKOR	VIseKriterijumska Optimizacija I Kaompromisno Resenje
VMs	Virtual Machines
VNF	Virtual Network Functions
VPN	Virtual Private Networks
WCDMA	Wideband Code Division Multiple Access
WFV	Wireless Function Virtualisation

1. Introduction

1.1 Background to the study

The telecommunication industry has grown exponentially since the advent of the 1st Generation of wireless technologies. The growth of the industry can be attributed to the increase in the demand for network services from different users. Users requesting for network services have diverse requirements with some requiring services to stream videos and audios, services to send and receive multimedia files, and services to make online purchases. These services require network providers to provide users with good QoS levels whist efficiently utilising the network resources available. 5G wireless networks are expected to meet the rising and diverse requirements of users.

In order to meet the diverse requirements of users, 5G wireless networks employ different technologies like Network Function Virtualisation (NFV), Software Defined Networking (SDN and Network Slicing. Network slicing is the logical separation of the physical network infrastructure into different independent virtual networks which are configured to fulfil the diverse requirements requested by specific users. Network slicing improves the utilisation of radio resources by ensuring that only the resources necessary for the call are made available. Network slicing succeeds in ensuring that 5G networks efficiently utilise network resources. However, in order to provide good QoS services, 5G is expected to provide seamless and continuous connectivity to users even at cell edges and thus mobility management is considered as a key challenge in 5G wireless networks.

Mobility management is the ability to maintain connection whilst the call is being transferred from one cell to another. An example of scenario when mobility management is key is when a user move from one location to another whilst on a call, it is the networks duty to maintain the call whist transferring it from one cell to another. When a call is transferred to a cell supporting the same network technology, the process is referred to as horizontal handoff and when a call is transferred to a cell supporting a different network technology, the process is referred to as vertical handover. In the context of 5G networks calls can be transferred to slices supporting the same use cases in the process referred to as intra slice handover or to slices supporting different use cases in the process referred to as inter slice handover.

In order to provide seamless connectivity during handover, 5G networks must implement efficient algorithms in order to ensure that less calls are blocked or dropped during handoffs and the number of handoffs performed is minimised. Vertical handover algorithms in 5G networks incorporates different bandwidth reservation policies in order to give priority to handoff calls and maintain good levels of QoS.

1.2 Objectives of this study

1.2.1 Problems to be investigated

The number of users requesting for different network services is expected to increase exponentially with the deployment of 5G wireless networks. This increase in the number of users means more calls are likely to be blocked or dropped as users move from one coverage area to another thus there is need to implement efficient call admission control algorithms in order to minimise the number calls are blocked or dropped and resources are efficiently utilised in the network. Furthermore, network users are known to be more intolerant to dropped handoff calls thus there is need to integrate handoff call prioritisation schemes in the call admission control algorithm hence there is need to investigate the effects of user mobility on the networks with handoff call prioritisation.

1.2.2 Purpose of the study

The purpose of this study is to design and evaluate the performance of an algorithm that can be used to make handover decisions in 5G wireless networks. An investigation is made on how the performance of the proposed handover algorithm is affected by parameters such as call arrival rate, call departure rate, capacity , threshold and required basic bandwidth unit. The results obtained from the simulations will be analysed in order to determine if the proposed algorithm is feasible. In order to be feasible, the proposed handover algorithm must provide good QoS by ensuring that the call blocking and dropping probabilities are minimised.

1.3 Scope and Limitations

This project will focus on the implementation of an algorithm to make handover decisions in sliced 5G networks. The algorithm is evaluated based on a network model consisting of three network slices. The performance of the algorithm is evaluated using the call blocking and dropping probabilities. Sliced 5G networks are complex thus this report does not delve into the physical infrastructure on the networks and how resources are allocated to the slices. In terms of time, there are 16 weeks to complete this project thus some call admission control algorithms cannot be implemented due to their complexity.

1.4 Plan of development

Figure 1.4.1 shows the plan for development for the entire project.

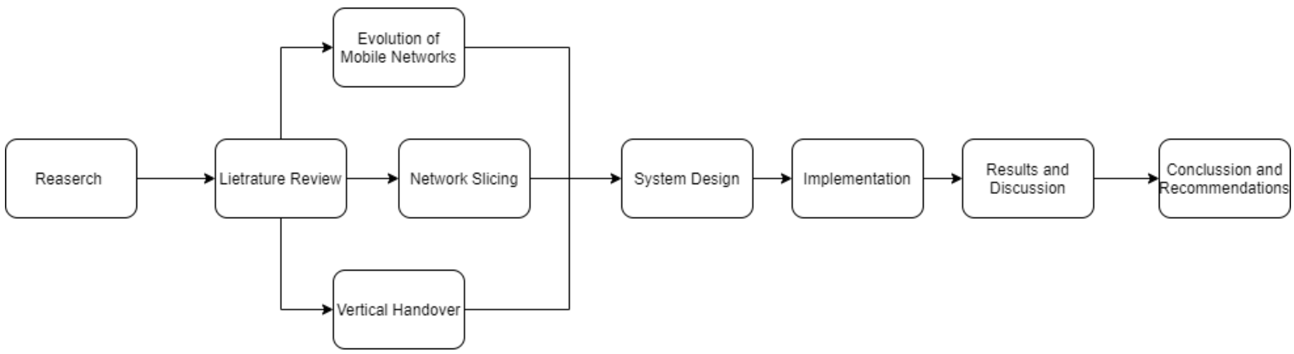

Figure 1.4.1: Plan of Development

1. **Research:** This involves conducting intensive research on the thesis topic in order to develop a clear understanding of the topic and derive the objectives of the project are outlined in Chapter 1

2. **Literature Review:** In Chapter 2, a detailed review on the evolution of mobile networks focusing on the technologies implemented and services provided by the network generations is provided. The review further expands on the concept of Network Slicing focusing on the challenges, requirements and enabling technologies. Furthermore, the

review also delves into the concept of Vertical Handover focusing on the existing algorithms in literature.

3. **System Design:** Chapter 3 describes the network model used in this project. The call admission control algorithm implemented in this project is detailed using flow chart diagrams. Furthermore, the markov chain model used for performance evaluation is described. Lastly, the metrics used to evaluate the performance of the algorithm are discussed.

4. **Implementation:** In Chapter 4, the implementation of the call admission control algorithm described in the System Design section is described. The tools and approach used to implement the algorithm are detailed and the building blocks in program used to simulate the algorithm are explained.

5. **Results and Discussion:** Chapter 5 presence the results obtained from simulating the algorithm implemented in the Simulation Design section. A comparative analysis of the algorithm is used to evaluate the performance of the algorithm under different conditions.

6. **Conclusions:** In Chapter 6, the conclusions drawn from the results obtained from the simulations in Chapter 5 are presented.

7. **Recommendations:** In Chapter 7, various areas of future work are recommended.

8. **References:** Chapter 8 lists all the papers referenced in the report.

9. **Appendices:** Chapter 9 presence the information that supports the analysis and validates the conclusions made in the report.

10. **Ethics in Research Projects:** Chapter 10 presence the Engineering and Built Environment (EBE) Faculty ethics assessment form.

2. Literature Review

Mobile communication can be traced back to the time when Guglielmo Marconi first pioneered long distant radio transmission. The technology behind mobile communication has since evolved remarkably since the advent of Guglielmo. The most notable advancements in wireless technologies is in relation to the number of subscribers and mobile technologies [1]. The advancements in wireless technologies comprises of several generations and is still advancing. The generation refers to the change in the system in relation to speed, technologies implemented and frequency. Each generation is characterised by capacity, techniques and features which differentiate from the preceding generation. 1G is the first generation of wireless technology. The key technique behind 1G was frequency reuse and it was used for voice calls only [2]. 2G is the second generation of wireless technologies which was based on digital technologies like Time Division Multiple Access Technologies (TDMA). 3G is the third generation of wireless technology which provided multimedia services and high data rates. 4G is the fourth generation of wireless technology which integrated 3G with fixed internet in order to support wireless mobile internet [1]. 5G is the fifth generation of wireless technology which is envisaged to provide high bandwidth, low latency and always on connection. This section briefly introduces the evolution of wireless networks, outlining the key characteristics of each generation. It then expands the theory on 5G network slicing and vertical handover.

2.1 Evolution of Mobile Networks.

2.1.1 First Generation (1G)

The first generation (1G) of wireless networks was launched by Nippon Telegraph and Technology (NTT) in the metropolitan area of Tokyo in Japan in 1979 and was made available in Europe and North America in 1982. The key technique behind 1G was frequency reuse and it was based on a technology known as Advanced Mobile Phone System (AMPS) and was mainly used for voice services [1]. AMPS made use of frequency modulation to transmit radio signals using Frequency Division Multiple Access (FDMA) with a channel capacity of 30KHz and a frequency band ranging from 824MHz to 894MHz and supported a speed up of up to 2.4kbps [1]. The introduction 1G brought about an improvement in voice communication between users however, the voice communications in 1G were susceptible to eavesdropping because any receiver with an all radio could listen into the conversation [3], furthermore 1G networks had low capacity and low data which impacted performance leading to poor Quality of Service (QoS).

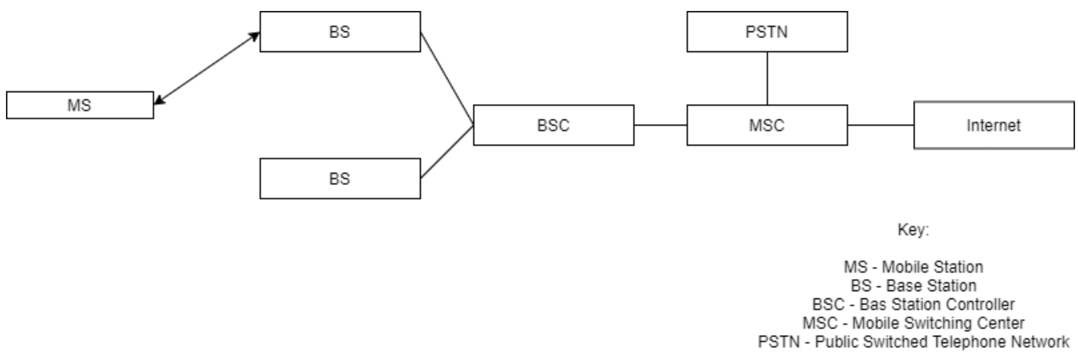

Figure 2.1.1 AMPS Architecture [1]

2.1.2 Second Generation (2G)

2G is the second generation of wireless networks which is based on the digital technology and it first introduced late 1980s. 2G quickly gained popularity because it addressed the security issues associated with 1G and in addition to providing voice services, 2G also provided services such as short message services (SMS), Multimedia Services (MMS). Compared to 1G, 2G implemented digital modulation techniques such as Time Division Multiple Access Techniques (TDMA) and Code Division Multiple Access (CDMA) [1]. TDMA facilitates many users to share the same frequency channel by dividing the signal into different timeslots thereby allowing users to transmit in rapid succession. CDMA implements a channel splitting technique like the one implemented in TDMA and assigns a code to every call thereby making it possible to distinguish calls [4]. Different TDMA technologies like Global System for Mobile Communications (GSM), Personal Digital Cellular (PDC), Integrated Digital Enhancement (iDEN), IS-136 and CDMA technologies like IS-95 were used in 2G wireless networks.

The European Telecommunication Standard Institutes (ETSI) developed GSM. GSM implements a combination of TDMA and FDMA as its multiple access techniques and it was the most used 2G standard. GSM was the first 2G technology to support international roaming and this enabled mobile users to use the same mobile device around the world.

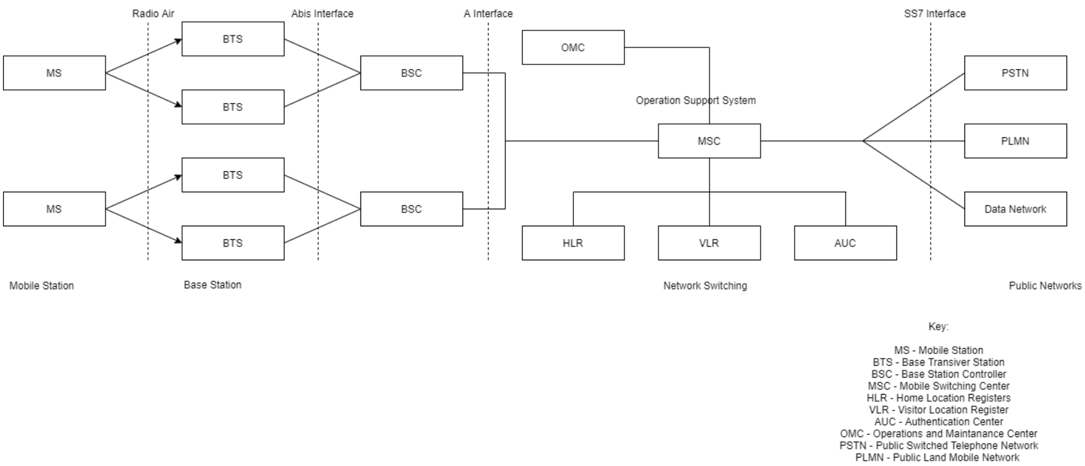

Figure 2.1.2: GSM Architecture [1].

However, GSM could not provide high data rates as a result it was further improved to Global Packet Radio Service (GPRS) and Enhanced GPRS (EGPRS) commonly known as 2.5G and 2.75G respectively. 2.5G and 2.75G implemented a packet switched domain on top of the circuit switched domain implemented in 2G, this led to new services like web browsing and enabled Mobile Network Operators (MNOs) to bill users based on data sent rather than connection time [5]. The need to further increase data rates and improve user QoS led to the introduction of Enhanced Data Rate for GSM Evolution (EDGE). However, the introduction of EDGE was not enough to handle the growing demand in the number of internet user as well as the complex data.

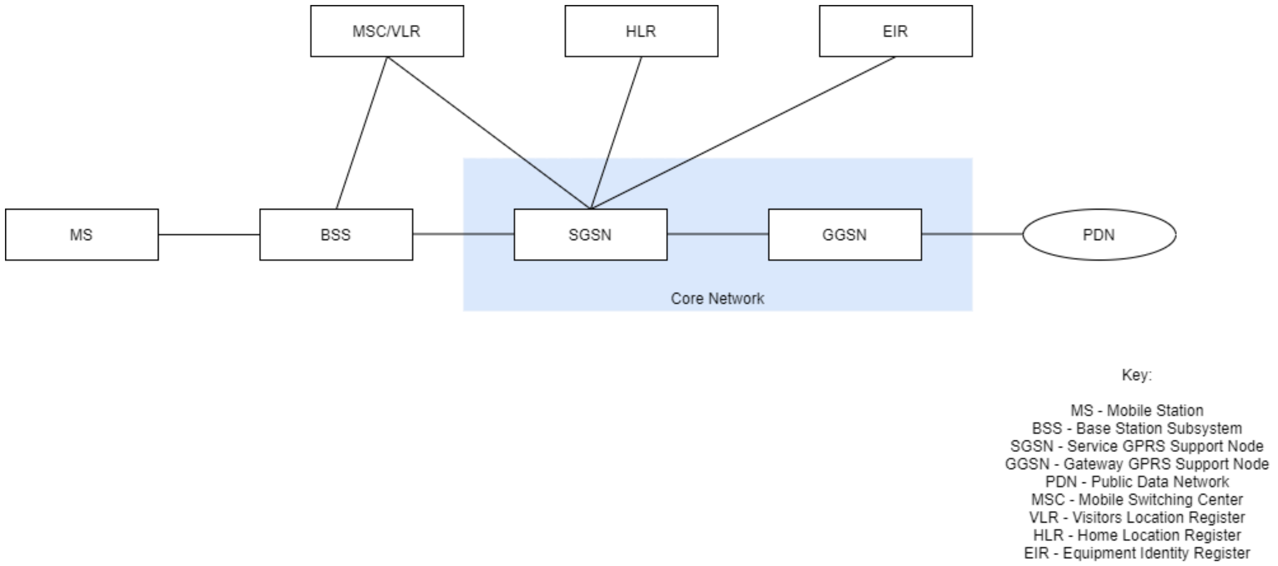

Figure 2.1.3: GPRS Architecture [6]

2.1.3 Third Generation (3G)

The third generation of wireless mobile communication systems was launched in 2000. The goal of 3G was to address the limitations of 2G by providing high data rates and increasing network capacity. Along with voice communication, 3G provided users with new services like mobile TV, video conferencing, Global Position System (GPS), video calls, web browsing and email in all mobile environments [2].

3G is based on the International Mobile Telephone 2000 (IMT-2000) standard which was formulated by the International Telecommunication Unit (ITU) with the goal to provide wireless communication across the globe. The IMT-2000 standard led to the evolution of standards used in preceding wireless generations. The 3G evolution for GSM led to the Wideband CDMA (WCDMA) which is also known as the Universal Mobile Telecommunication Service (UMTS) [7]. UMTS emerged from Europe and was developed by an organisation called 3rd Generation Partnership Project (3GPP). The enabling technologies in UMTS are TDMA and WCDMA. In WCDMA, users are allocated frames with dynamic data rates and each frame is partitioned into time slots with each time slot being either uplink or down link. WCDMA architecture is made up of the Radio Access Network (RAN) and the Core Network (CN). The CN is responsible for switching, routing calls and all the data connections whereas the RAN is responsible for all radio functions.

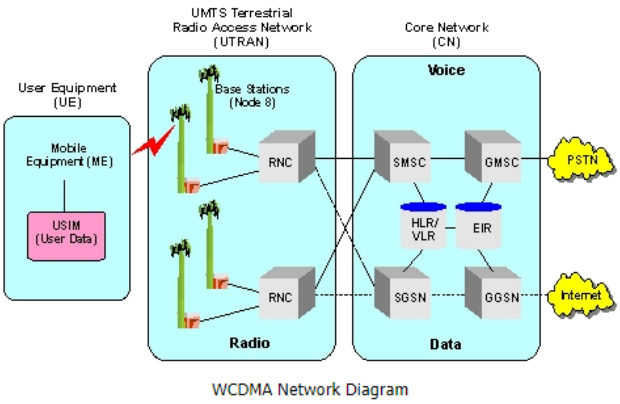

Figure 2.1.4: WCDMA Network Diagram [1]

3.5G High Speed Downlink Packet Access (HSDPA).

3.5G is a mobile telephony protocol which enable high data transfer rates. It is a packet-based data service implemented in the WCDMA and it provided downlink data rates of up to 10Mbit/s [7]. 3.5G most popular implementations includes Massive Input Massive Output (MIMO), Adaptive Modulation and Coding (AMC) and fast cell search. With AMC being used to increase the downlink data rates.

3.75G High Speed Uplink Packet Access (HSPA).

3.75G is another technology introduced under UMTS. It is related to the 3.5G and the two technologies complement each other. Services like email and real time gaming improved massively in HSPA due to the fast data rates in the uplink enabling the services to respond in real time.

2.1.4 Fourth Generation (4G)

The fourth generation (4G) of wireless networks was introduced in the late 2000s with the aim of provisioning an all IP based network system. Being an all IP system means that the system is a packet-based network hence all the data is transferred using the same method which is independent of the access or transport technology. The reason behind the transition to an All-IP is to establish a common platform for all existing technologies [8].4G provides high data rates, higher channel capacity, higher efficiency, low cost for voice services and data services over IP [1].

4G wireless networks assimilated existing wireless technologies like Orthogonal Frequency Division Multiple Access (OFDMA), Multi-Carrier Code Division Multiple Access (MC-CDMA) to provide seamless and quicker handover between technologies.

An implementation of 4G known as Long Term Evolution (LTE) was introduced by the 3GPP. LTE provided high data rates to technologies like MIMO by using the frequency spectrum ranging from 1.25MHz to 20MHz [9]. LTE was further upgraded to LTE-Advanced (LTE) and LTE-pro.

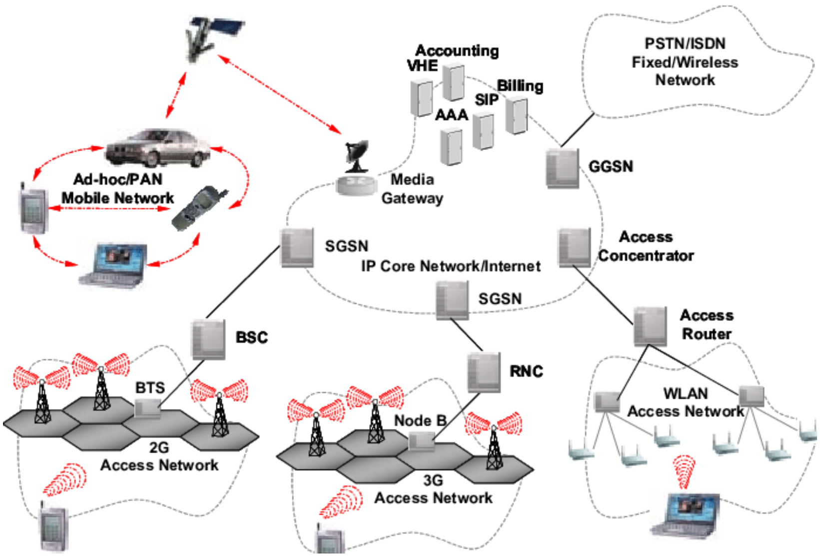

Figure 2.1.5: The generic 4G mobile network architecture [9].

2.1.5 Fifth Generation (5G)

Fifth generation (5G) is the latest iteration of wireless networks, it is expected to be deployed by the end of 2020. 5G is an All-IP network system and it is envisaged to provide services to a wide range of sectors including but not limited to defence agencies, health care institutions, energy utilities and manufacturing systems [10]. Being an All-IP system, 5G will satisfy user QoS requirements in different applications, furthermore 5G systems should provide wide area coverage such that users will always have seamless connectivity anywhere even in locations at cell edges [11]. 5G wireless networks will integrate technologies like Network Function Virtualisation (NFV), Software Defined Networks (SDN), Multiple-Access Edge Computing (MEC), Millimetre Wave Transmission (mMT), Next Generation Protocols (NGP). In order to provide services tailored for specific user's QoS demands, 5G will integrate the network slicing technology which allows Network Operators (NOs) to virtualise networks on shared infrastructure thereby allowing service providers to flexibly deploy their services.

The goal of 5G wireless technologies is to provide seamless connectivity across multiple devices and applications. 5G network requirements include.

- Maintain very high speeds for upload and download
- Provide Bi-Directional bandwidth
- Provide remote diagnostics
- Support different Modulation techniques.

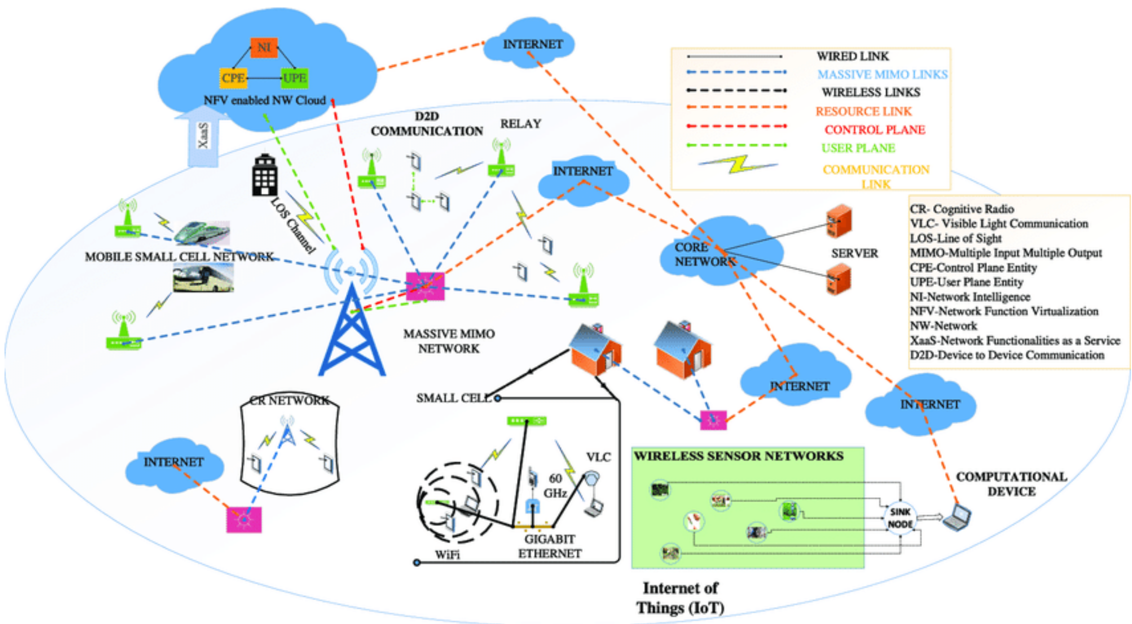

Figure 2.1.6: 5G Network Architecture [12].

Generation	1G	2G	3G	4G	5G
Year Deployed	1979	1991	2001	2009	2020
Speed	2.4Kbps	64Kpbs	2Mbps	200Mbps-1Gbps	10 – 50Gbps
Access Technology	AMPS, FDMA	GSMA, TDMA	CDMA-200 (UMTS, EDGE)	Wi-Max, LTE, Wi-Fi, OFDMA	BDMA, FBMC
Frequency Band	800MHz	200KHz	5MHz	1.4MHz to 20MHz	1.8-2.6GHz
Services	Voice only	Voice, Data, SMS	Voice, Data, SMS, MMS, Video Calling	Smart wearable devices, Online Gaming, HD Tv	Smart wearable devices with AI capabilities, ultra HD video
Handoff	Horizontal	Horizontal	Horizontal and Vertical	Horizontal and Vertical	Horizontal and Vertical
Multiplexing	FDMA	TDMA and CDMA	CDMA	OFDMA	OFDMA
Switching	Circuit Switching	Circuit switching and packet switching	Packet Switching	Packet Switching	Packet Switching
Forward Error correction	N/A	N/A	Turbo Codes	Turbo Codes	Low Density Parity Check Codes (LDPC

Table 2.1.1: Summary of Mobile wireless networks evolution

2.2 Fifth Generation (5G) Wireless Technology

This section expands 2.1.5 by further explaining the architecture of 5G wireless networks, requirements, and the challenges of 5G networks.

2.2.1 Fifth Generation (5G) Requirements

For 5G to meet the diverse user and application requirements, there are several requirements that the network must meet. The requirements are derived from the applications that will require service from the network.

To ensure that 5G provides the best services and meet the demands of businesses and users, the requirements outlined must be met.

- **Latency:** Latency refers to the time for data to from source to destination. It is very important in applications like self-driving cars where response time can have impact on the outcome [13].

5G is expected to provide a latency of 10ms for general purpose applications and a latency of 1ms for applications requiring low latency. Low latency enhances user QoS and maintains high connectivity.

- **Spectrum:** 5G networks are envisioned to operate at higher frequencies than its preceding networks. The spectrum of 5G networks is expected to be in the range of 1GHz to 100GHz. The distance of propagation for higher frequencies is shorter than that of lower frequencies hence 5G networks will implement higher networks in ultra-dense areas.

- **Traffic Handling:** The age of 5G networks coincides with the rapid increase in the number of devices requiring network connection, it has been reported that by 2021 the number of smart phones and machine to machine (M2M) devices will reach 8 billion and 13 billion respectively [10]. The increase in devices leads to an exponential increase in data requirements hence 5G networks are expected to improve all the technologies involved such that the network can provide seamless connectivity.

- **Throughput:** Throughput refers to the amount of data that was transferred successfully from one place to another in each period. 5G networks are expected to provide peak data rates ranging from 10Gbps to 50Gbps for low mobility users whilst also improving the cell edge data rate by ensuring that users always get close to peak data rates despite of their locations.

- **Energy Efficiency:** 5G networks are expected to be backward compatible with its preceding networks e.g. 4G. One of the drawbacks of 4G networks is energy consumption, hence 5G is expected to address the energy consumption problem without sacrificing user QoS. Efficient energy utilisation in 5G networks is achieved by harnessing power from sources like wind. The energy will be used in small cell communication whilst a combination of energy sources can be used in macro cells [14].

- **Interoperability:** 5G is expected to use multiple Radio Access technologies (RATs). Like GSM, RATs are expected to continue operating whilst 5G networks provide new features to enhance the services they can provide to users.

- **Security:** 5G networks are expected to provide diverse services to users hence there should be a dedicated mechanism that will help ensure that user security is not compromised.

- **Cost:** 5G networks are expected to provide higher capacity, faster connection and low cost.

- **Flexibility:** 5G wireless networks are expected to provide different network slices which offer different services through the use of network virtualisation. This requires a high degree of flexibility in the architectural design because there is need for network resources to be allocated dynamically [15].

- **Mobility:** Mobility is the ability of users to maintain connection whilst changing their geographical position. 5G is expected to provide connection for all users moving at different speeds e.g. pedestrians (2km/hr), cyclists (10km/hr), motorists (80km/hr), high speed trains (500km/hr).

2.2.2 5G Network Architecture

A one size fits all architecture is not viable for 5G networks because the network is expected to offer different slices with each slice having different requirements. 5G network engineers are faced with the need to build a system that ensures that the QoS requirements of each user are met whilst making sure that the resources available are efficiently utilised.

A series of recent studies has indicated that the research on 5G networks is influenced by two notions: *the evolutionary and service-oriented view* [10]. The evolutionary view addresses issues with latency, throughput and bandwidth. On the radio access side, the evolutionary view implements technologies like MIMO, Beamforming and Millimetre wave. The service-oriented view addresses both the radio access and core network side of 5G networks. In order to address a wide range of network requirements, the service-oriented view includes technologies line SDN, NFV and Network Slicing.

- **Millimetre Wave:** The rate of data consumption is in wireless networks is increasing exponentially whilst the resources available remain constant. This presents a problem in mobile networks because at the current rate of increase, services provided to users decrease in quality and other users might not get access to resources resulting in a sharp decrease in the QoS and user Quality of Experience (QoE). 5G networks addresses the problem by increasing frequency and the most viable way of obtaining more frequency is moving the frequency towards the millimetre wave spectrum [16]. Millimetre waves broadcast waves in the range of 30-300GHz and wavelength ranging from 1 to 10 millimetres. Faster data rates are guaranteed by allocating larger bandwidths, furthermore the technology provides a flexible spectrum which enables the dynamic allocation of bandwidth [17]. However, millimetre waves suffer from poor propagation across building hence the need for small cells.

- **Small Cells:** The small cell technology is more about getting around an obstacle instead of facing it [18]. Small cells (femto and Pico) are small base stations with low power consumption during operation. In order to maintain an always on connectivity, thousands of small cells can be deployed across a city to provide a denser network. Small cells are made up of smaller antennas hence they can be deployed on top of buildings or poles thereby providing land utilisation. Installing small cells close to each other enables frequency reuse for different services. Despite the advantages associated with deploying small cells, there is a drawback in their infrastructure complexity which can affect their deployment in remote areas.

- **Massive MIMO:** The concept of massive MIMO involves having more antennas fitted on a single array thereby increasing the capacity of the mobile terminal without acquiring another spectrum because base stations will be able to send and receive signals from multiple subscribers simultaneously. Using MIMO results in an increase in the spectrum efficiency, however the presence of more antennas in a single array introduces interference hence there is need for new technologies that can fully harness the spectrum efficiency whilst minimising interference.

- **Beamforming:** Beamforming addresses the interference issue introduced in Massive MIMO by identifying the most efficient data delivery route to a subscriber [19]. The technology of beamforming makes use of the millimetre and small cell technologies. Beamforming reduces interference by concentrating beams in order to focus signals in the desired direction instead of scattering them in all directions. Beamforming introduces coordination in the way in which packets are exchanged and selects the path with the least interference.

- **Device to Device (D2D) Communication:** D2D communication refers to the direct communication between devices in proximity without any data paths or routing. The main advantage of D2D communication is ultra-low latency due to shorter travel path.

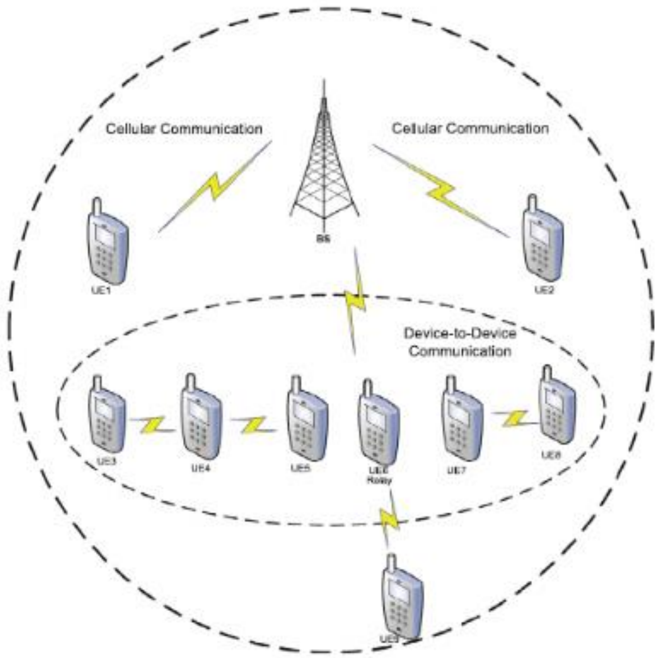

Figure 2.2.1: An Overview of Device to Device Communication [20]

- **Full Duplex:** Full duplex makes use of the same frequency channel to simultaneously transmit and receive data thereby reducing latency and increasing throughput. The amount of time needed by full duplex to transmit a signal is half of the time needed by the current age of base stations because they a separate receiver and transmitter. The main drawbacks to implementing full duplex is interference caused by the echo and the collisions caused by the two-way communications. In order to mitigate the collisions, techniques like digital cancellation can be implemented in full duplex systems.

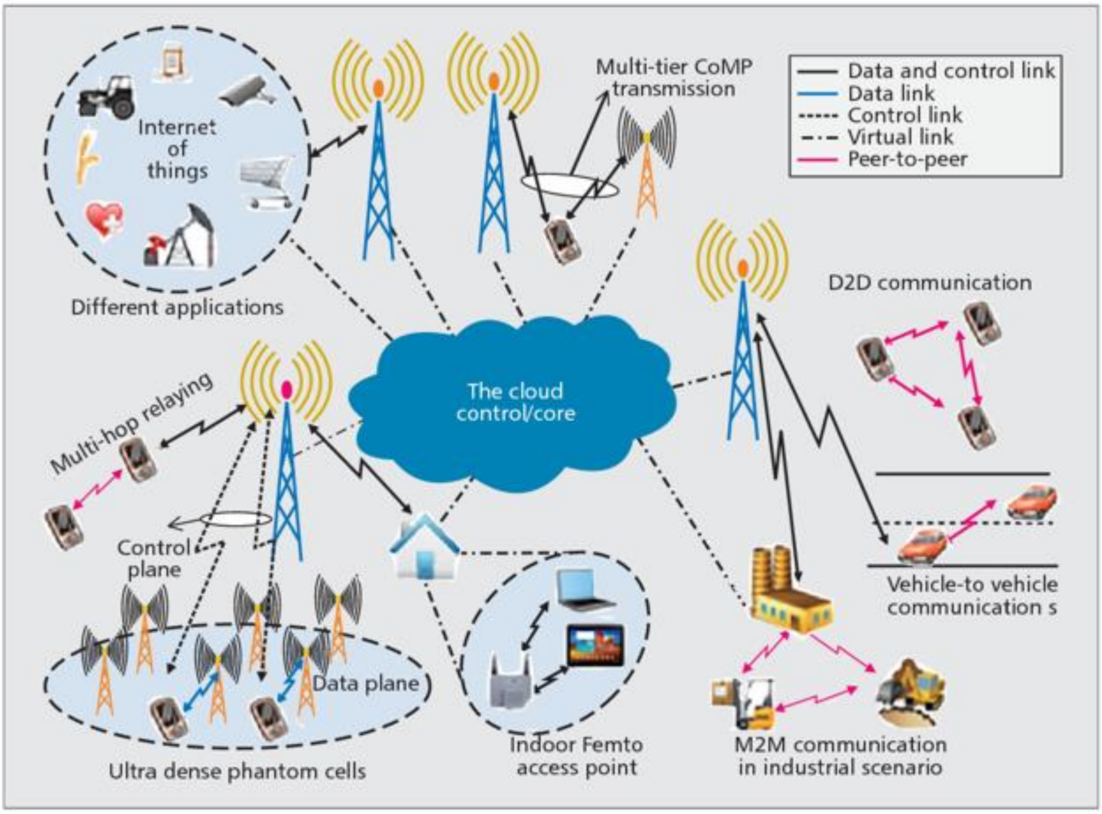

Figure 2.2.2: A general 5G Network Architecture [12]

2.2.3 5G Challenges.

- **Machine to Machine (M2M) communication:** 5G networks are envisaged to serve multiple devices ranging from small devices e.g. mobile phones to large devices e.g. home appliances, smart grid systems and health monitoring systems. The introduction of large devices will result in an increase in the amount of traffic connected to the network as well as an increase in the data rates and bandwidths required, furthermore, large devices consume more power because they are always on. M2M communication give rise to the need for new and efficient ways to manage bandwidth in order to accommodate all devices [21].

- **Capacity:** 5G wireless networks are expected to increase the network capacity whilst limiting the increase in the cost of operation. Technologies that can be integrated into 5G networks to help increase network capacity include the use of small and macro cells to separate the control and data plane.

- **Small Cell Deployment:** The small cell deployment in other countries have been slowed down by the excessive obligations placed on the mobile operators by the local authorities. Some of the constrains placed on the deployment of small cells includes long permitting processes, outdated regulations and excessive fees [22].

- **Privacy and Security:** One of the key requirements of 5G is fast data rates, this means that the network can provide access to multiple devices at the same time therefore there is need to ensure that user data is secured and protected from cyber-attacks.

➢ **Availability of Compatible Devices:** The deployment of 5G networks will coincide with the need for new devices that can fully utilise the network whist also being backward compatible with previous network generations.

➢ **Spectrum Challenges:** Telecommunications companies are faced with a challenge to efficiently allocate the global spectrum. Efficient allocation of the spectrum results in higher bandwidth utilisation.

➢ **Deploying Cost:** Deploying 5G networks involves building new infrastructure, and this can prove to be costly for telecommunication companies.

2.3 Network Slicing.

The fifth-generation era of wireless technologies is touted as the generation of mobile networks that will support different use cases and provide specific services to different users in order to meet different QoS requirements. The concept of network slicing in the context of 5G networks is defined as a process whereby network resources are packaged and assigned to different users with different requirements in an isolated manner [23]. A slice is defined as a collection of resources packaged in order to meet specific user requirements in terms of QoS offered by the slice. The goal of network slicing is to improve resource utilisation by providing only the resources required by the user.

The enabling aspect of network slicing is virtualisation of network resources which enables the sharing of physical infrastructure amongst different service providers in a flexible and dynamic manner thereby allowing them to efficiently utilise the network resources. Network slicing is envisaged to be more efficient when it comes to the allocation of resources because the idea of virtual networking belongs Software Defined Networks (SDN) architecture which aims to migrate modern networks towards a software-based paradigm.

5G networks can have dedicated slices configured for the major use cases in the network e.g. Enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communication (uRLLC), Massive Machine Type Communication (mMTC) as depicted in Figure 2.3.1. The slices are isolated from each other in order to provide users with the best QoS and security.

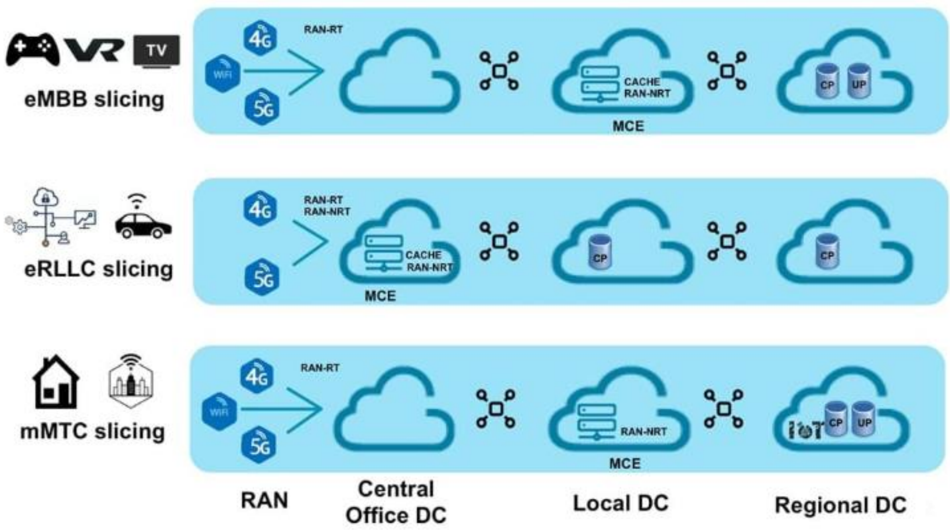

Figure 2.3.1: 5G Network Slices [24]

2.3.1 5G Network Use Cases (Slices)

The ITU and the Fifth Generation Public Private Partnership (5GPPP) have divided the 5G service use cases into three main categories based on the requirements of each use case. The use cases are categorised as eMBB and mMTC. Each use case has unique requirements which are specific to that use case.

- **Ultra-Reliable Low Latency Communication (uRLLC)**

uRLLC slice is designed to cater for applications that can be categorised as mission critical i.e. applications that require low latency, high reliability and security. Mission critical applications require a continuous stream of data and are less tolerant to delays hence the ideal latency for applications is less than 1 millisecond and the data transmission rate of up to 10Gbps. Applications that can utilise the uRLLC slice include telemedicine applications and self-driving cars.

- **Enhanced Mobile Broadband (eMBB)**

eMBB slice is designed to improve the limitations of the previous generation of wireless networks (4G). The main goal of eMBB is to enhance connectivity at cell edges, provide faster and always available networks with an uplink and downlink of 1Gbps and latency of less than 1 millisecond. Applications that can utilise eMBB include Augmented Reality (AR), Virtual Reality (VR) and high-quality video streaming. The main requirements of the applications are high data rates and low latency.

- **Massive Machine Type Communication (mMTC)**

mMTC slice is designed to cater for numerous Internet of Things (IoT) devices. IoT devices are always connected and they are characterised by low throughput and low data transmission rate in the uplink. The end to end delay requirements of IoT devices is low hence mMTC slice offers latency of less than 10 milliseconds with a minimum bit rate of 0.5Mbps. Applications that can utilise the mMTC slice include smart home and smart city devices.

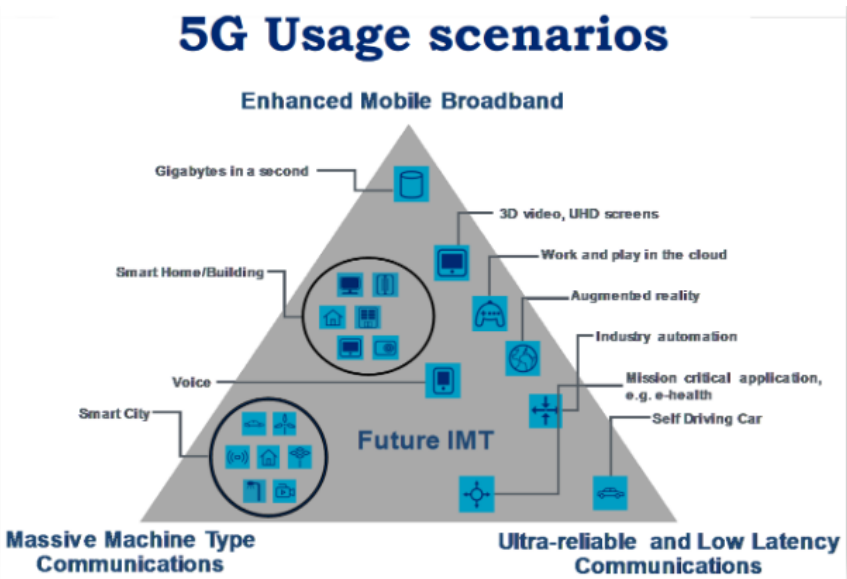

Figure 2.3.2: 5G Usage Scenarios [25]

2.3.2 5G Network Slicing Enabling Technologies.

Network slicing leverages Wireless Function Virtualisation (WFV), Multiple Edge Computing (MEC), Software Defined Network (SDN), Network Function Virtualisation (NFV) in order to provide a network service where slices can be created dynamically.

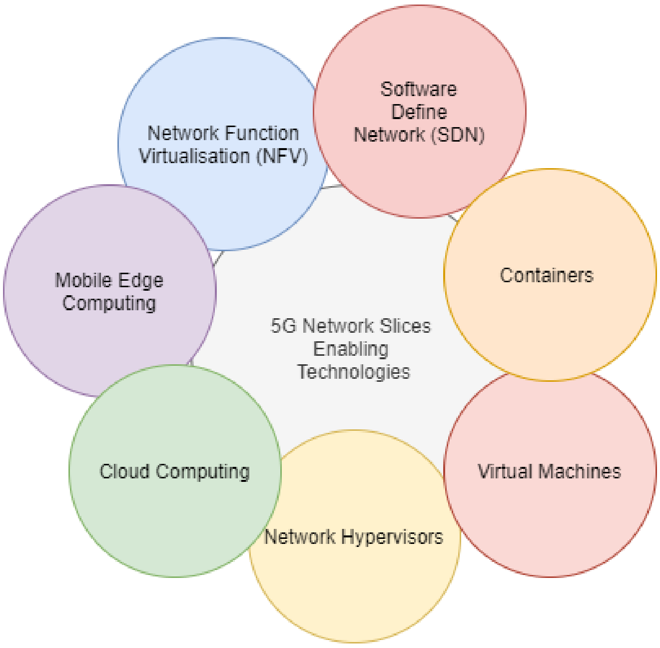

Figure 2.3.3: 5G Network Slicing enabling Technologies [26].

- **Software Defined Network (SDN):** The Open Network Foundation (ONF) defined SDN as the physical separation of network control plane from the forwarding plane. The separation provides flexibility and affords the control with a global view of the entire network whilst also providing ways of responding to network changes [26]. SDN is used to construct a virtualised control plane which leverages the gap between network management and service allocation using intelligent network management techniques. Standard Southbound Interfaces like OpFlex [27], OpenFlow [28] allows direct programming to the network control. The SDN controller uses a set of predefined rules to dynamically manage the slices by grouping slices that belong to the same context [26].

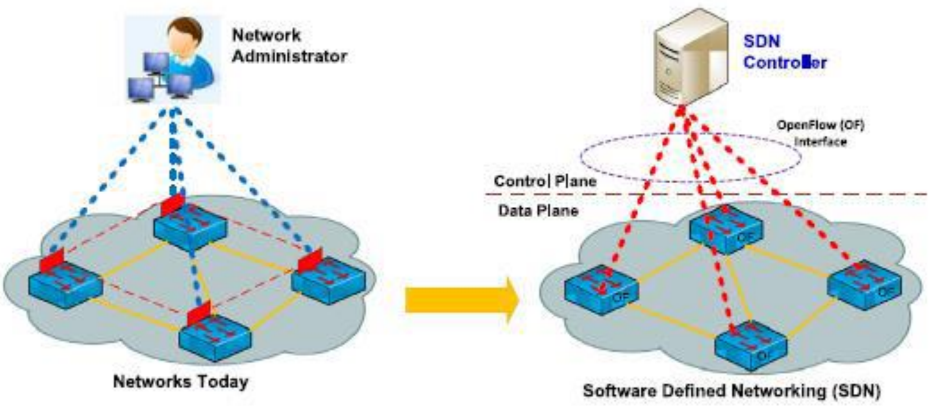

Figure 2.3.4: A comparison of SDN and today's network operation [26]

- **Network Function Virtualisation (NFV):** NFV is a technology that can be used to virtualise network functions e.g. Virtual Private Networks (VPN) and Firewalls on top of commodity hardware thereby breaking the traditional way of using hardware and software [26]. NFV moves the network functions to the services providers and they are executed on shared infrastructure thereby making it easy to manage the adding, removing or updating of a function for subscribers without physically visiting the subscriber's location. In the context of 5G, NFV guarantees performance of the Virtual Network Functions (VNFs) operations including low latency and low failure rate [26].

- **Multiple Edge Computing (MEC):** MEC is a technology that is used to provide application and content providers with the ability to use cloud computing services at the edge of the mobile network [29]. MEC facilitates the delivery of ultra-low latency to mission critical applications by processing the data close to where it is being consumed. Network resources utilisation is improved by the implementation of MEC through hosting applications at the network edge.

- **Cloud Computing:** Cloud computing refers to the technology which is used to deliver computing services including but not limited to servers, storage databases and software over the internet ('The Cloud'). According to Mujumbi *et al* [30], the service provider can be categorised into either Infrastructure Providers (InPs) and Service Providers (SPs). The primary role of InPs is to manage the cloud platforms and lease the services based on usage. SPs are tasked with renting resources from InPs and distributing them to different users.

- **Network Hypervisors:** Network hypervisors are used to abstract the physical infrastructure into virtual network slices.

- **Virtual Machines (VMs):** Virtual Machines allows for virtualisation of physical resources thereby allowing the user to execute an independent Operating System (OS) with its functions isolated from the host despite sharing critical resources like storage, network and memory.

- **Containers:** Containers are based on the idea of OS-level virtualisation and can be used as alternatives to hypervisor based VMs [26]. Containers are run on the OS and provide higher application density. Docker, Linux-Vserver, OpenVZ are examples of containers. In the context of 5G networks, VMs and Containers can be used to run VNFs chained together, thereby creating a 5G network slice.

2.3.3 Network Slicing Requirements

5G wireless networks are envisaged to be flexible, demand oriented and scalable such that they can meet the stringent requirements of different users. The network slices are expected to be designed beforehand with specific users in mind. 5G networks are expected to cater for multiple Radio Access Technologies (RATs), this leads to the need to redesign the Radio Access Network (RAN) and Core Network (CN) such that they can support and fully harness the benefits of network slicing.

Due to the programmability of the CN, network operators are able to address the needs of the users whilst also reducing their Capital Expenditure (CAPEX), Operational Expenditure [26] and network revenue. Network operators have more control over how resources are allocated to users and they can implement methods like machine learning algorithms to allocate resources to users based on user history.

2.3.4 Challenges of Network Slicing

The concept of network slicing is not without any challenges. Many of the challenges faced are due to the concept being a new topic in academia and industry hence there are conflicts when it comes to the best way to implement the technology.

- **Security:** 5G wireless networks provide services to intelligent devices hence there is need for new intelligent security protocols that can shield the network from cyber-attacks like Denial of Service (DoS). The security in 5G networks can be either security due to network virtualisation or security in the management of slices.

- **Mobility Management:** Mobility management is used to ensure to ensure continuity of ongoing calls whilst users move at different velocities. Seamless and fast handover is a requirement for applications requiring low latency. In order to implement seamless handover, there is need to keep track of the user's location and predict the direction they will most likely take as well as the time at a certain RAT. The implementation is a challenge because it requires synchronization amongst different Mobile Virtual Network Operators (MVNOs).

- **Resource Allocation:** 5G networks are dynamic in nature thus it is difficult to allocate resources in the network. The resource allocation techniques implemented in 5G networks are responsible for making decisions on the resources that will be virtualised, prioritising slices and allowing coordination amongst MVNOs.

- **Isolation:** Isolation in network slices is a concept of ensuring that changes associated with one slice cannot affect another slice. Implementing isolation in network slices is difficult because the channel quality changes in an unpredictable manner.

2.4 Heterogenous Networks

The emergence of heterogenous networks can be attributed to the need to solve the node deployment density problem faced by carriers. Heterogenous networks integrates low power nodes on the back of macro cells in order to reduce macrolevel density and improve QoS at cell edges [31]. Pico and Femto cells are examples of low power nodes used in heterogenous networks. The cells are placed in areas based on the traffic density knowledge of that area and they can be deployed anywhere due to their small physical size.

Homogenous networks allow mobile terminals to access one RAT at a time. This leads to poor radio resource utilisation. Heterogenous networks mitigates the poor radio resource utilisation problem introduced in homogenous networks by allowing a mobile terminal to access multiple RATs thereby improving resource utilisation and QoS.

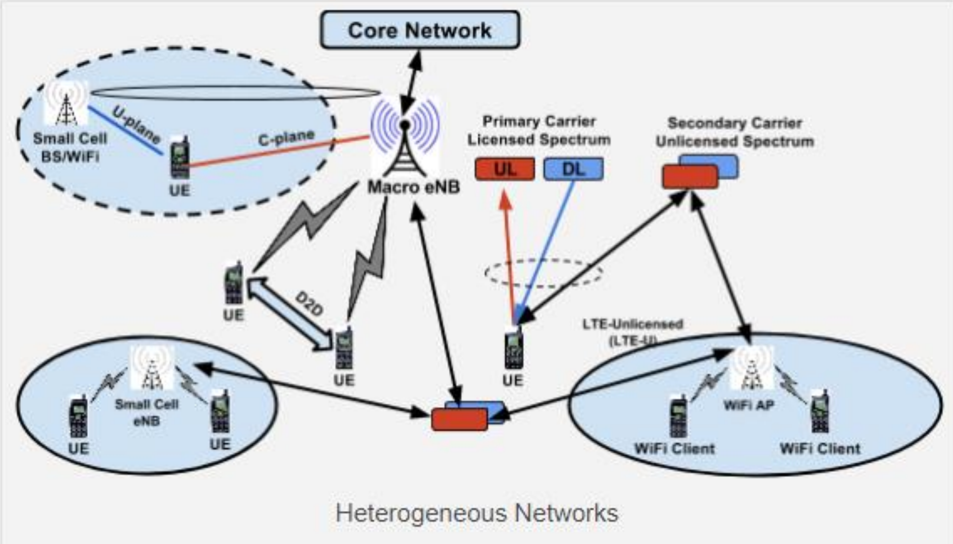

Figure 2.4.1: 5G Network Illustration [32]

Figure 2.4.1 illustrates a heterogenous network in 5G networks which has data plane separated from the control plane. The control plane controls the data by providing connectivity to the User Equipment (UE) and the data plane provides data transport. The architecture illustrated provides high spectral and energy efficiency.

2.4.1 Radio Resource Management (RRM)

The concept of RRM helps with the efficient managing of the scarce radio resources in wireless networks whilst also ensuring that the QoS requirements are met. Different techniques of RRM are implemented in order to manage how radio resources are allocated to different RATs. The RRM techniques can be classified into, handover control, call admission control, scheduling and power control.

- **Handover Control:** When users move from one coverage area to another, handover control guarantees that the user remains connected when the UE changes it's point of attachment.

- **Call Admission Control (CAC):** CAC is used to decide whether a call is admitted or declined in a resource constrained network without compromising the QoS of calls already admitted in the network.

- **Scheduling:** Scheduling is used to give priority to users so that the network can meet different QoS requirements whilst fully utilising network resources.

- **Power Control:** The increase in the demand for network resources can be solved by increasing the network capacity. However, an increase in network capacity results in an increase in the power consumed which leads to the degradation of QoS due to the interference hence there is a trade-off between network capacity and interference. Efficient power control algorithms are needed in order to maximise network capacity without degrading QoS.

RRM can be independent RRM or joint RRM (JRRM). JRRM has an advantage over independent RRM because it offers enhanced QoS, improves radio resource utilisation and allows users to connect to any available RATs and supports handover between RATs [33]. JRRM techniques can be implemented through Joint Call Admission Control (JCAC) algorithms and load balancing algorithms amongst others.

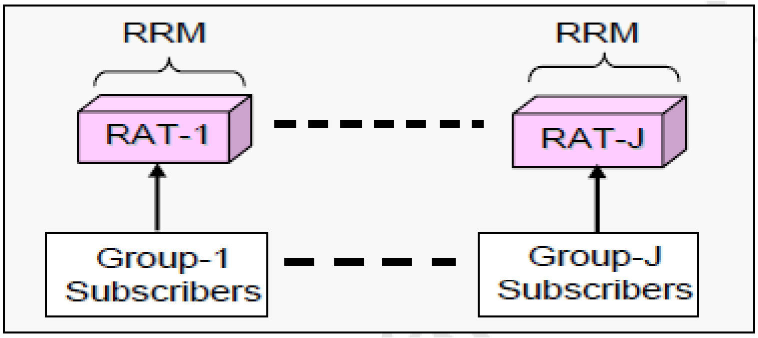

Figure 2.4.2: Independent Radio Resource Management [33]

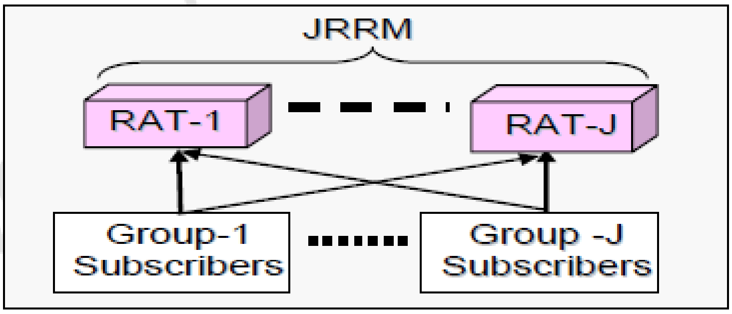

Figure 2.4.3: Joint Radio Resource Management [33]

2.4.2 Joint Call Admission Control Admission (JCAC).

Radio resources are scarce and expensive hence it is important to efficiently utilise the radio resources in wireless networks. JCAC is one of the fundamental RRM techniques used for QoS provisioning in wireless networks. The primary responsibility of JCAC is to block or drop calls if the network does not have enough resources. New or handoff calls are only accepted if the network has enough resources to meet the QoS requirements of the call without compromising the QoS of calls already admitted into the network [34].

JCAC is a vital part of wireless networks, however aspects like channel interference, unpredictable link quality, limited bandwidth and user mobility complicates the implementation of JCAC in wireless networks. Prior research has emphasised that JCAC algorithms need to be accurately designed in order to minimise false rejection and false admission [35].

- **False rejection:** This is when a call is rejected despite the network having enough capacity to satisfy the call's QoS requirements leading to poor radio resource utilisation and poor QoE.

- **False Admission:** This is when a call is admitted into a network that does not have enough capacity to satisfy the call's QoS leading to poor QoS and user's QoE.

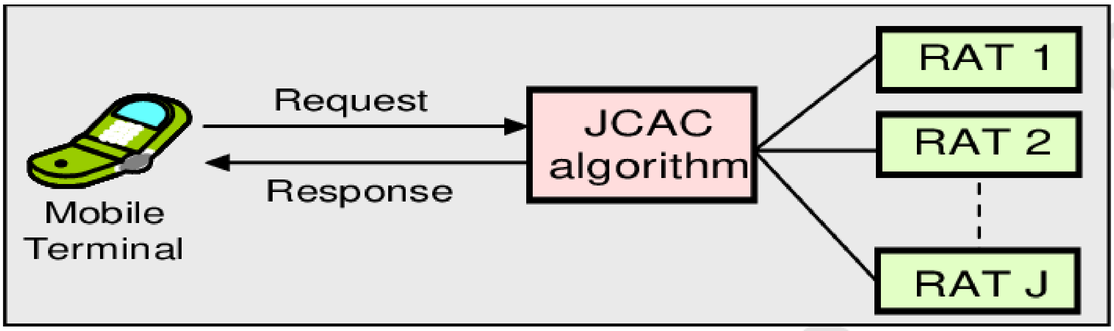

Figure 2.4.4: JCAC in Heterogenous Networks [36]

Advantages of JCAC Algorithms

- **Enhanced QoS Provisioning:** JCAC algorithms fully utilise the benefit of heterogenous networks when a new or handoff call cannot be admitted in the target RAT, the JCAC algorithm checks if there's any RAT available where the call can be admitted thereby significantly decreasing new call blocking and handoff call dropping probability [34].

- **Efficient Radio Resource Utilisation:** JCAC enhances the sharing of resources amongst different RATs whilst maintaining the required QoS levels.

- **Network Stability:** JCAC algorithms help achieve load balancing in heterogenous networks by ensuring no RAT is overloaded. JCAC monitors the load of each RAT and if RAT is overloaded or if a RAT is idle, the algorithm hands some calls from the loaded RAT to the least loaded RAT [37].

- **Service Class Prioritization:** Calls can be admitted into the network based on their class of service e.g. voice calls can be given priority over video calls because voice calls arrive in the network more frequently.

- **Guarantee Signal Quality:** JCAC algorithms guarantees signal quality by ensuring that a call is only admitted if and only if the network can maintain the signal quality allocated to ongoing calls [34].

- **Increased Revenue:** JCAC can be used to increase revenue by allowing the network to dynamically change prices depending in the traffic.

- **Enhanced User Satisfaction:** JCAC algorithms can be used to direct calls to RATs which are preferred by users in terms of price, QoS offered and services offered thereby enhancing user satisfaction [37].

2.4.3 JCAC Requirements.

For a JCAC algorithm to perform effectively, certain requirements must be satisfied. Several authors have recognised that JCAC algorithms need to be simple, scalable, adaptable, have a high execution speed, efficient and multi service.

- **Efficiency:** The efficiency of a JCAC algorithm is determined by how the algorithm guarantees resource utilisation and QoS whilst minimising the number of new calls blocked and handoff calls dropped.

- **Multi Service:** The next generation of wireless networks supports different classes of calls e.g. video, voice, video streaming, video streaming and browsing [36]. The ability of networks to support different classes of calls enhances users QoE because users are able choose from the available services, furthermore, having different services increases the networks revenue because users can subscribe to different services.

- **Simplicity:** JCAC algorithms are expected to have less computational overhead because users are less tolerant to delays, furthermore the algorithms are expected to be simple to implement and cost effective. However simple algorithms maybe less efficient hence there is a trade-off between efficiency and simplicity [37].

- **Scalability:** The demand for radio resources is expected to continue increasing hence implemented JCAC algorithms must be able to accommodate any changes in the networks capacity and seamlessly integrate with other networks [36].

- **Stability:** Instability in heterogenous networks can be defined as the situation whereby one RAT is overloaded whilst other RATs leading to frequently move users between RATs leading to poor QoS. JCAC can provide network stability by implementing load balancing [36].

- **Adaptivity:** Parameters like link quality in wireless networks are unpredictable hence it is important for JCAC algorithms to be able to adapt to changes in such parameters in the network in order to maintain high levels of resource utilisation and QoS.

- **Speed of Execution:** JCAC algorithms are expected to perform the call admission in real time because users are intolerant to delays.

2.4.4 JCAC Design Considerations

Efficient JCAC algorithms can be designed based on different approaches. Falowo in [36] describes the most used approaches when designing JCAC algorithms.

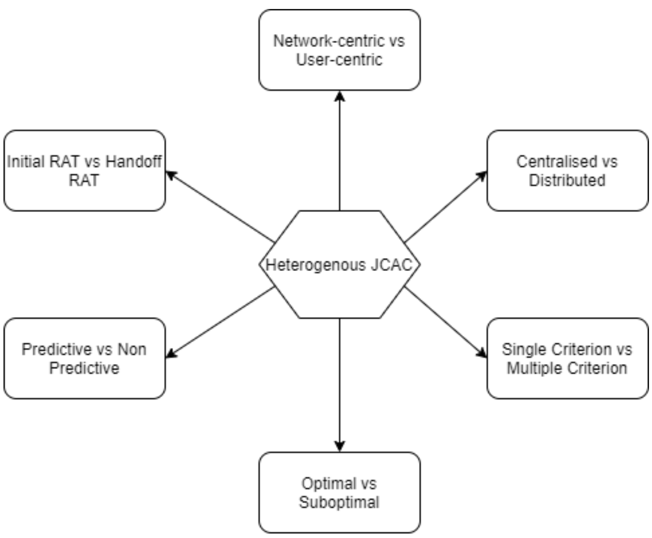

Figure 2.4.5: JCAC Design Considerations [36]

- **Network Centric vs User Centric:** In Network centric JCAC algorithms, decision to select a RAT is made based on the network's status whereas in User Centric JCAC algorithms, the decision to select a RAT is made based on the user's preference e.g. cost. Network Centric Algorithms are implemented when load balancing is the priority whereas User Centric are implemented in networks whereby user's QoE is the priority.

- **Centralised vs Distributed:** Centralised JCAC algorithms uses sensors to make RAT selection decisions. The sensors collect data and send it to the base stations for further analysis. In Distributed algorithms, the nodes are distributed across the network hence the mobile terminals do not need to communicate with centralised nodes there [36]. Distributed algorithms reduce the computation overhead because mobile terminals do not communicate with central nodes, furthermore, they are scalable hence they have an edge over centralised algorithms.

- **Single Criterion vs Multiple Criterion:** Single criterion JCAC algorithms make use of one criterion when selecting the optimal RAT whereas multiple criterion algorithms make use of multiple criterion when selecting the optimal RAT. Multiple criterion algorithms enhances user QoS and are more efficient compared to single criterion algorithms.

- **Optimal vs Sub-optimal:** Optimal JCAC algorithms implements optimising techniques like linear programming and artificial intelligence to calculate the optimal value. The main reason for implementing optimal algorithms is to either maximise or minimise network parameters e.g. radio resource utilisation [36]. Optimal algorithms are highly efficient; however, they are difficult to implement. Sub optimal algorithms on the other hand are easy to implement but they are less efficient.

- **Predictive vs Non-Predictive:** Predictive JCAC algorithms use predictive techniques like linear regression and classification machine learning algorithms to predict the type of call the user is going to request and the prediction is used to choose the best RAT. Prediction algorithms can be efficient and accurate; however, their accuracy is dependent on the accuracy of the training set data.

- **Initial Vs Handoff RAT:** Initial RAT JCAC algorithms are designed to accommodate new calls whereas handoff RAT JCAC algorithms are designed to accommodate handoff calls.

2.4.5 RAT Selection Algorithms.

The RAT selection methods implemented in JCAC can be user preference based or network operator's preference based. In user preference based JCAC algorithms, when a user makes a call request, the user also sends the preferred factors to the network. Some of the factors used in user preference-based RAT selection methods include minimum delay, maximum data rate and service cost [36]. In network operator-based RAT selection methods, the network has a set of predefined values and an incoming call is accepted if it satisfies the defined criterion. Some of the factors used in network operator-based methods includes minimisation of call blocking and dropping probabilities, revenue maximisation and load balancing.

Multiple Criteria Decision Making (MCDM) algorithms make use of preferred factors from both the user preference based and network operator-based algorithms when selecting the optimal RAT. Using MCDM algorithms guarantees that both the network and user's requirements are satisfied whereas single criterion algorithms cannot guarantee that both the network and user requirements are satisfied at the same time. However, MCDM algorithms are complex and may introduce additional computation overhead [38].

Studies of JCAC algorithms are well documented, it is well acknowledged that JCAC can be categorised into different groups based on the RAT selection criteria. In [39] Nguyen *et al* identifies the groups of RAT selection algorithms as centralised (*network controlled*), distributed (*user controlled*) and hybrid (*user and network assisted*).

- **Centralised RAT Selection:** In the centralised approach, the decisions on which RAT to connect to is made by the network. This is achieved by prompting the users to send their desired conditions to a local controller. The controller then calculates the optimal value which is based on the user's association to the RAT as well as the networks objectives. Centralised RAT selection algorithms allow the network to achieve network related objectives e.g. load balancing, throughput maximisation and user fairness enhancement [40]. However, the centralised approach require collaboration between all Base Stations (BSs) and users hence there is a significant increase in the communication overhead furthermore different network operators adopt different communication techniques hence it such levels of collaboration might not be feasible across multiple networks.

- **Distributed RAT Selection:** RAT selection algorithms are implemented at the user's side in order to mitigate the communication overhead introduced in centralised RAT selection. The algorithms are implemented as iterative game-based algorithms. The distributed game theory algorithms are classified as either partially distributed or fully distributed algorithms [41]. Partially distributed game theory algorithms are algorithms whereby the user uses information about other users to update their strategy whereas fully distributed algorithms are algorithms whereby users make their decisions without any knowledge of other users.

  The implementation of a partially distributed algorithm in the context of wireless networks can be described as the situation whereby users have complete knowledge of the network e.g. selection history of other users. The information can be used to determine the services and service parameters offered to other users. On the other hand, the implementation of a fully distributed game algorithm can be described as the situation whereby the user makes use its own selection history to select a RAT. Partially distributed algorithms are preferred over fully distributed algorithms because guarantee convergence whilst fully distributed algorithms suffer from slow convergence [41].

- **Hybrid RAT Selection:** In Hybrid RAT selection algorithms, the decision on which RAT to connect to is based on users' observations and information from the network [41]. Recent studies have proposed network structures where the global knowledge of the network can be broadcasted to users periodically and each user can use the received information in conjunction with the user's requirements to select the optimal RAT. The hybrid selection algorithm is more efficient however, the need to periodically send information to user equipment introduces a communication overhead and the information broadcasted by the Base Stations (BSs) is not always accurate therefore users would need to continuously switch between BSs in order to find the one that satisfies their needs. This leads to a significant decrease in throughput per user and a high number of exploration times [42].

2.4.6 Challenges of Designing JCAC Algorithms

Designing admission control algorithms for heterogenous networks is complicated because the RATs must support multi services. The JCAC algorithms are supposed cater for different classes of calls and calls with different QoS requirements. Furthermore, JCAC algorithms are expected to make call admission decisions for multiple RATs with different features e.g. cells size. The algorithms are also expected to make admission control decisions for vertical handover which is a feature available in heterogenous networks.

2.4.7 Bandwidth Allocation

Bandwidth management is a challenge for heterogenous networks the network has a finite number of resources available, this puts a limiting threshold on the number of users that can connect to the network at a given time. This section evaluates different bandwidth allocation schemes that can be implemented in heterogenous networks. The evaluation is based on how the bandwidth allocation schemes compare against each other.

- **Complete Sharing:** Complete sharing is one of the first bandwidth allocation strategies to be implemented. It works on a first come first serve basis with no priority given to either handoff calls or new calls. An incoming call is accepted if there are enough radio resources to accommodate it without any interruption to ongoing calls. If the network gets to maximum capacity, new incoming calls are blocked, and handoff calls are dropped.

  Advantages:

  - Good radio resource utilisation.
  - Simple to implement.

  Disadvantages:

  - Poor QoS due to high handoff call dropping probability.
  - No differentiation for calls with stringent QoS requirements.

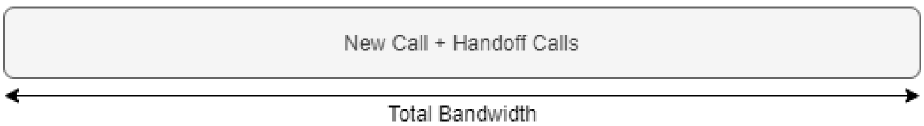

Figure 2.4.6: Complete Sharing Bandwidth Allocation Scheme

- **Complete Partitioning:** In complete partitioning, the network bandwidth is divided into two separate pools. The pools are divided such that one is dedicated to new calls and the other is dedicated to handoff calls

  Advantages:

  - Simple to implement

  Disadvantages:

- Poor radio resource utilisation, one pool might be overloaded whist the other remain idle.

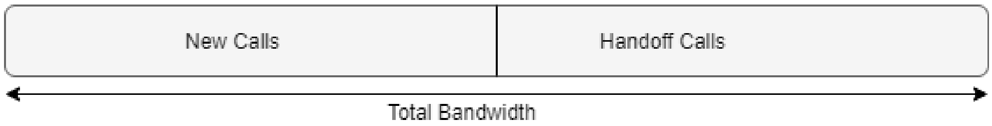

Figure 2.4.7: Complete Partitioning Bandwidth Allocation Scheme

➢ **Service Class Prioritisation:** In service class prioritisation, users are given priority based on their service classes. The decision to give users preferential treatment is based on the reasons that some calls have stringent QoS requirements and other users are willing to pay more for better QoS.
Advantages:

- Differential treatment of users based on QoS requirements.

Disadvantages:

- Difficult to implement

➢ **Handoff Prioritisation:** Due to user's mobility, it is necessary for an ongoing call to be handed over from one cell to another. However, there is no guarantee that the call handed over to another cell will get the required resources in the target cell thus the call may be dropped if the target cell does not have enough radio resources. Wireless network subscribers are known to be more intolerant towards dropped handoff calls than blocked new calls hence it is important to prioritise handoff calls over new calls. Handoff call prioritisation can be categorised into guard channel, fractional guard channel, queueing priority scheme and QoS degradation scheme.

  - **Guard Channel:** In guard channel, there are some channels known as guard channels which are reserved for handoff calls. Handoff calls can be accepted into the whenever is a channel available whereas new calls are accepted if and only if number of channels catering for ongoing calls is less than a certain threshold. For example, if the total number of channels available is C, then the number of guard channels is C-H. New calls are accepted if the total number of channels being used by ongoing calls is less than the threshold H [43]. The threshold is a value that is chosen in order to minimise handoff call dropping probability whilst maximising the number of new calls accepted into the network.

  - **Fractional Guard Channel:** In Fractional guard channel, handoff calls are given priority over new calls by accepting new calls with a certain probability that is based on the number of busy channels. The probability of blocking new calls is directly proportional to the number of busy channels i.e. if the number of busy channels is low, the probability of blocking new calls is low and when the number of busy channels increase, the probability of blocking new calls increase.

  - **Queueing Priority Scheme:** In queueing priority schemes, new and handoff calls are accepted whenever there are free channels with no prioritisation. When the network has reached full capacity, handoff calls are queued whilst new calls are blocked, alternatively new calls can be queued with a certain rearrangement in the scheme [44]. When network resources become free, one of the queued handoff calls is accepted and the remaining calls wait for resources to become available. However, storage space is limited thus calls can only be in the queue for a limited duration.

- **Quality of Service Degradation Scheme:** The QoS degradation scheme can be classified as either bandwidth or delay degradation. In bandwidth degradation scheme, calls are classified as degradable or non-degradable with degradable calls having flexible QoS requirements and non-degradable calls having stringent QoS requirements. When there are free channels, all calls are accepted into the network and they are allocated the maximum bandwidth according the call's requirements and as the network become congested, the bandwidth allocated to degradable calls is reduced.

2.4.8 Handover Management in Heterogenous Networks

Call handover can be defined as the procedure that transfers an ongoing call from one cell to another as the user moves from one coverage area to another [45]. The handover can be classified as either horizontal or vertical depending on the type of RATs involved in the handover. Handovers are also classified as hard or soft handovers depending on the number of connections involved. In hard handover (*break before make*), the connection with the previous Point of Attachment (PoA) is broken before a new connection is established. In Soft handover, (*make before break*), the UE maintains its connection with its old PoA whilst it's establishing a new connection with a new PoA. When changing PoAs, the UE can stay within the same domain (Intra domain handover) or migrate to another domain (inter domain handover). Figure 2.4.8: Handover Classification Figure 2.4.8 shows the handover classifications documented in literature [46]. The focus of this report is on Vertical Handover (VHO).

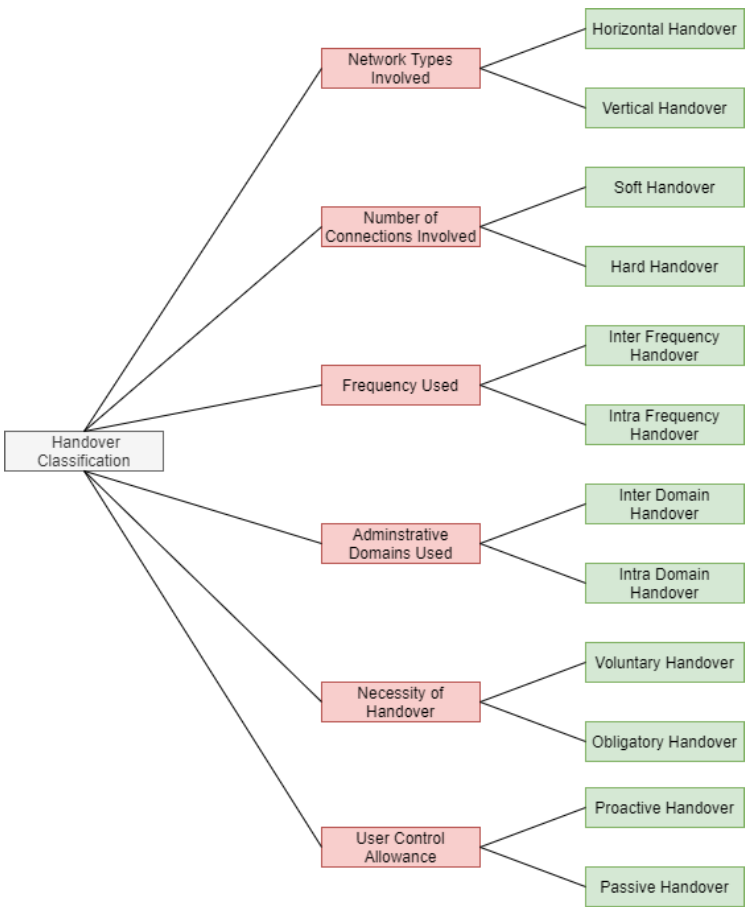

Figure 2.4.8: Handover Classification

2.5 Vertical Handover

Vertical handover occurs when between PoA supporting different network technologies. In the context of 5G network slices, it can be described as inter slice handover whereby a call attached to the eMBB slice can be handed over to the uRLLC slice. The handover decision making process can be divided into three main strategies namely: information collection, handover decision making and handover execution [47].

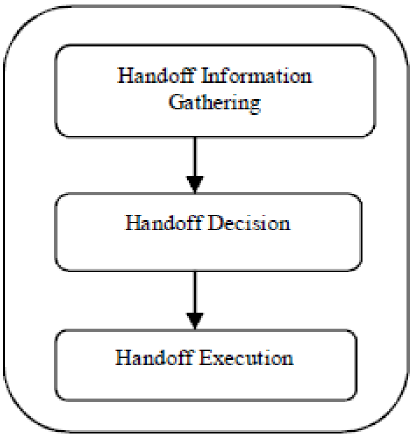

Figure 2.5.1: Handover Management Process [47]

2.5.1 Seamless handover characteristics

Seamless handover is defined as the handover that efficiently utilises radio resources whilst ensuring that the commitments made to ongoing calls are not violated that is seamless handover improves the QoS. Seamless handovers can be characterised by speed, number of handovers and multiple criteria.

- **Speed:** For a handover to be seamless, it must be executed in the shortest time possible. This can be achieved by using efficient algorithms in the handover information gathering and handover decision stages.

- **Number of handovers:** Seamless handovers are expected to maintain high levels of QoS and user QoE. This can be achieved by limiting the number of handovers executed because executing handovers is expensive in terms of time and resources like power.

- **Multiple Criteria:** For a seamless handover to select the best PoA, the algorithms implemented in the handover decision must incorporate multiple criteria decision schemes .

2.5.2 Issues with vertical handover.

Some of the issues to consider when executing vertical handover in 5G networks includes the handling of mobility and network conditions like network load, ping pong effect and maintaining high QoS. In order to implement an efficient vertical handover algorithm, the issues depicted in Figure 2.5.2 need to be addressed.

- **Mobility and Network Conditions:** 5G networks are expected to provide consistent QoS across different slices. Slices in 5G networks are designed for use cases with different requirements hence it is difficult to guarantee consistent QoS across the slices.

- **Security:** UE migrate from one slice to another during vertical handover hence it is necessary to implement robust security solutions in the slices [48]

- **Ping Pong Effect:** Users require that their ongoing sessions remain uninterrupted whist their session transition from one slice to the other. However, the session can be interrupted mainly due to handover failure. One of the main reasons for handover failure is the ping pong effect, which can be defined as the situation whereby the UE executes handover between the same slices due to the undesirable fluctuations in network resources [47]. The Ping pong effect must be avoided because it results in packet loss and increases the power consumption and connection level interference.

- **Load Balancing:** Load balancing refers to the technique of migrating UE from heavily loaded slices and connecting them to lightly loaded slices. This helps improve the QoS and resource utilisation.

- **Quality of Experience (QoE):** In order to provide users with the best QoE, it is necessary to consider user and UE preferences when designing the vertical handover algorithms.

- **Operator Control in Network Selection:** Network operators must have control over network selection because they are responsible for pricing and billing hence it is important for UE to support network initiated handovers [11].

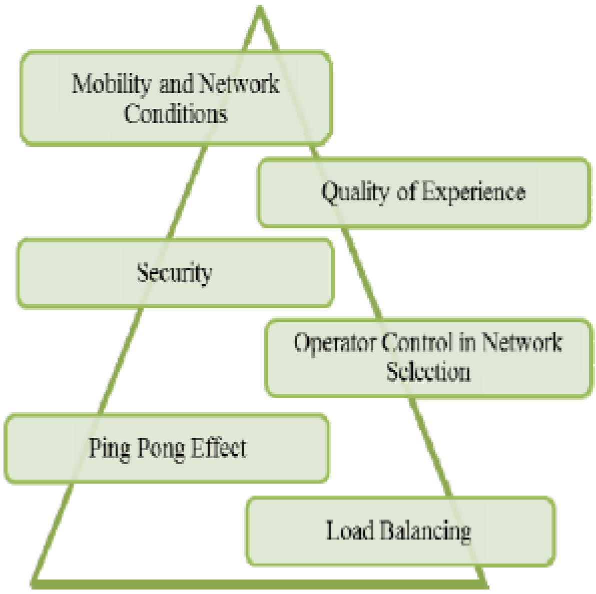

Figure 2.5.2: Factors affecting handover in 5G networks [47]

2.5.3 Vertical Handover (VHO) Decision Algorithms

Vertical handovers are carried out with the help of handover algorithms. The handover can be classified as markov chain-based models, game theory, fuzzy logic, utility theory, Multiple Attribute Decision Making Algorithms (MADM) and Machine Learning Algorithms.

- **Reflexive – Received Signal Strength (RSS) Based Algorithms:** The RSS Based strategy is VHO decision algorithm designed for dynamic environments where the RSS and available networks are continuously prompted and compared to the minimum threshold required by the UE and a decision is made based on the outcome of the comparison [49]. RSS based algorithms are easy to implement and can be implemented with other parameters like cost to enhance their efficiency.

- **Multiple Attribute Decision Making Algorithms (MADM):** The MADM algorithms make use of predefined parameters to select the best RAT from a pool of RATs after aggregating and evaluating the weights of chosen parameters based on the given criteria. MADM algorithms proposed in literature includes Simple Additive Weighting (SAW), Grey Relational Analysis (GRA), Analytic Hierarchical Process (AHP), Multiplicative Exponent Weighing (MEW), Elimination et Choix Traduisant La Realité (ELECTRE), VIseKriterijumska Optimizacija I Kaompromisno Resenje(VIKOR) [50].

- **Machine Learning Algorithms:** The rise in the number of research papers on Machine Learning can be attributed to the fact that the field has attracted a lot of attention due to the ability of neural network algorithms to learn from past data and make future predictions [51].

  In machine learning, the algorithms can be either supervised or unsupervised. In supervised learning, the algorithm is presented with a training set that consist of some input and desired outputs. In unsupervised learning, the algorithms are presented with random data and the main task of the algorithm is to identify a pattern within the data. Another section of machine learning is reinforcement learning, in reinforcement learning, the algorithm is tasked with improving on existing hypothesis functions with regards to new data and previous data e.g. geographical movements can be used to make handoff decisions.

2.6 Related Work

The increase in the number of devices requesting for services from wireless networks have led to an increase in the number of researches carried out on wireless networks in academia. This section gives an overview of the existing studies carried out on vertical handover in sliced networks and JCAC algorithms.

In [52], Yajnanarayana *et al* proposes a reinforcement learning based handover method. The method makes use of a centralised reinforcement learning agent which is used to handle network measurements coming from UEs and use the information to select the appropriate handover decision with the aim of maximising long-term utility. The proposed method is evaluated and compared against traditional handover algorithms and the results show an increase in the efficiency of making the right handover decision by at least 20%. However, the implemented algorithm follows the hybrid RAT section algorithm introduced in 2.4.5 and this method comes with an increase in computation overhead and a decrease in per user throughput.

In [37], Falowo *et al* presents a JCAC algorithm which makes call admission decisions based on user preferences. The proposed algorithm was simulated for instances where users incurred the least cost for their preferred services. The algorithm was simulated with a Markov Chain model in order to determine how the algorithm can effectively reduce the service cost and the results proved that the service cost can be decreased significantly by applying the algorithm.

Chen *et al* in [53], proposes a threshold-based call admission control scheme which is based on the Markov chain model and sensitivity analysis. The sensitivity analysis is introduced to limit the number of re-computations performed by the algorithm and the markov chain is used to derive the optimal policy. The algorithm was simulated in an event driven system with three party classes and the algorithm outperformed the guarded channel and complete sharing algorithm.

In [54] Alfoudi *et al* proposes a mobility management architecture which follows a modular approach whereby each slice has a module which handles it's mobility requirements and enforces the slice mobility management policies. The proposed architecture was implemented with a shared abstraction which is used by the slices to share resources. Each user is assigned a slice ID to aid with seamless connectivity and when the user is handed over to a different RAN, it uses its slice ID to determine which slice would offer it the best service. The proposed architecture results in high user satisfaction and efficient resource utilisation.

Falowo *et al* [55] proposed an Adaptive Bandwidth Management and joint call admission control (AJCAC) which can be used to enhance system utilisation, reduce call blocking and dropping probabilities and guarantee QoS requirements of all admitted calls are satisfied. The algorithm was simulated, and the results were compared against the results from a non-adaptive JCAC and an improvement by up to 20% in system utilisation was observed.

In [56] Lee *et al* adopted a distributed reinforcement learning algorithm which allows each device to learn the policies for efficient multi-connectivity configuration and select the appropriate RAT. The Authors proposes a Q-learning based approach because of its ability to learn the best policy without having prior knowledge of its environment. Reinforcement learning algorithms considers instant costs and cumulative costs for the future, when implemented with learning, the algorithm is envisaged to select the optimal policy which minimises total cost. During the learning process the learning rate ρ indicates the level to which the learnt value will override the old one with $\rho=0$ indicating nothing was learnt and $\rho=1$ indicating new knowledge which is the only data considered. Users in the reinforcement learning monitors the level of reference signal received power (RSRP) and selects the Base station with the strongest highest power. The performance of the algorithm is evaluated via simulations and its performance is compared against a single connectivity RAT selection algorithm and a dual connectivity RAT selection algorithm. The results showed that the proposed reinforcement learning algorithm achieves better reliability performance and better QoS relative to the single and dual connectivity algorithms, However the despite its advantages, the difficulty of implementation is the algorithms most notable disadvantage.

3. System Design

This chapter focuses on the design of a handover decision algorithm for sliced 5G wireless networks. The sliced network architecture allows the creation of logically independent networks which run on common shared infrastructure thereby allowing each slice to cater for specific use cases. The designed algorithm follows the formulas and notations used by Falowo and Chan in [57] and a numerical approach based on the Markov decision process is used to evaluate the handover decision algorithm.

3.1 Network Model

The concept of network slicing enables the building of multiple self-contained networks on top of a common shared infrastructure. The owners of the shared infrastructure are known as Infrastructure Providers (InPs) and they are responsible for providing resources to the tenants. Mobile Virtual Network Operators (MVNOs) are example of tenants that will receive resources from InPs. Resource allocation in sliced networks is achieved through a hierarchical scheme which involves both the InPs and MVNOs. The InPs and MVNOs are placed at different levels of the hierarchy and perform different resource allocation tasks. InPs allocates resources to the MVNOs which in turn allocates the resources to different slices.

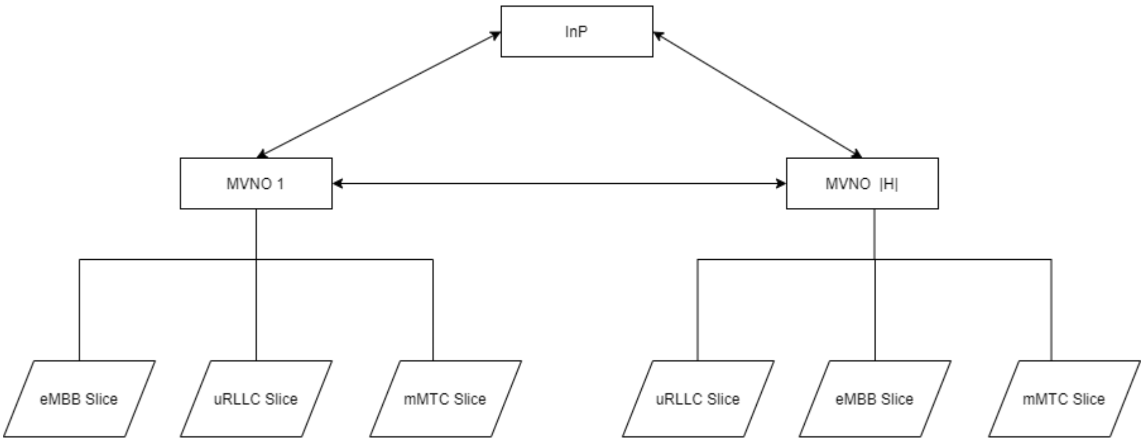

Figure 3.1.1: Hierarchical Network Model.

As identified in the literature review in section 2.3.1, a sliced network has distinct slices namely eMBB, mMTC, uRLLC with each slice having different requirements. The slices have resources for their specific use cases allocated to them, however the slices are not responsible for accepting any calls. In this report, a call refers to any device connected to the network ore requesting for resources from the network. The MVNO is responsible for accepting calls from different users and the CAC algorithm is implemented at the MVNO level of the hierarchy. The bidirectional arrow between the MVNOs in

Figure 3.1.1 illustrates the possibility of intra and inter slice handover i.e. when a call is admitted to the eMBB slice in MVNO 1 is handed over to the eMBB slice in MVNO 2 or the uRLLC slice in MVNO 2.

3.1.1 Network Model Assumptions

The following assumptions are made during the implementation of the network model.

- Slices are given equal priority during resource allocation by the MVNO in order to achieve good radio resource utilisation.

- Users at each slice are prioritised based on their arrival time.
- mMTC slice does not cater for handoff calls because it's UEs are not mobile.
- Handoff calls are prioritised over new calls in eMBB and uRLLC.
- Each user has different requirements as illustrated in Table 3.1.1
- Users are uniformly distributed in a slice.

Use Case	Use Case Call Type	Requirements
Enhanced Mobile Broadband (eMBB)	Video Streaming	Availability: Available in Normal Conditions Reliability: Normal reliability requirements. Latency: 5-10ms
Massive Machine Type Communication (mMTC)	Smart Home Devices	Availability: High Availability requirements Reliability: Normal – High reliability requirements Latency: 10ms (*more tolerant to delays*)
Ultra-Reliable Low Latency Communication	Autonomous Cars	Availability: Very High availability Reliability: Very High reliability Latency: 1ms (less tolerant to delays)

Table 3.1.1: User Requirements

3.2 Call Admission Control (CAC) Model

The proposed CAC algorithm for sliced 5G networks is designed with the following objectives;

- Guarantee QoS requirements
- Give priority to handoff calls.
- Minimise Call blocking and dropping probabilities.

Consider an InP *i* offering infrastructure services to a set of MVNOs.

$$H = \{1,2, \dots, |H|\} \qquad \textbf{3.2.1}$$

Each MVNO has a total of

$$X^h = \{1,2, \dots, |X^h|\} \qquad \textbf{3.2.2}$$

slices, with each slice isolated from other slices in the same network. Each slice in the virtual network is characterised by a use case and capacity. The capacity of the slice refers to the maximum amount of radio resources that can be offered by the slice. Several studies suggest that radio resources are offered in terms of timeslots, frequency channels or code sequences depending on the multiplexing technique implemented, however for this purpose of this study, the capacity is offered in terms of Basic Bandwidth Units (BBU). The use case refers to a group of services with similar requirements e.g. eMBB. A call requesting for services can be accepted by the MVNO and is assigned to a slice depending on its requirements. For example, a video streaming device can make a request for services and it is admitted to the eMBB slice by the MVNO. During the call admission, the MVNO considers the slice conditions in terms of the load. If the eMBB slice is currently overloaded, the call can be accommodated into the next slice which can satisfy the call's requirements depicted in

Table 3.1.1. When a call is handed over to a slice supporting different use cases e.g. eMBB call handed over to uRLLC slice, this is referred to as inter slice handover and when a call is handed over to a slice supporting the same use case e.g. mMTC call handed over to mMTC slice in , this is called intra slice handover. Calls for uRLLC slice have more stringent requirements hence they cannot be admitted in any of the slices. Calls for eMBB have requirements that can be satisfied by the uRLLC slice hence they can be handed over to the uRLLC slice. Calls for mMTC slice have flexible requirements and can be satisfied by both the eMBB and uRLLC slices hence they can be handed over to both slices. The calls are admitted into a slice if and only if the slice has enough radio resource to satisfy the requirements of the call without affecting other admitted calls otherwise the call is handed over to another slice or blocked or dropped.

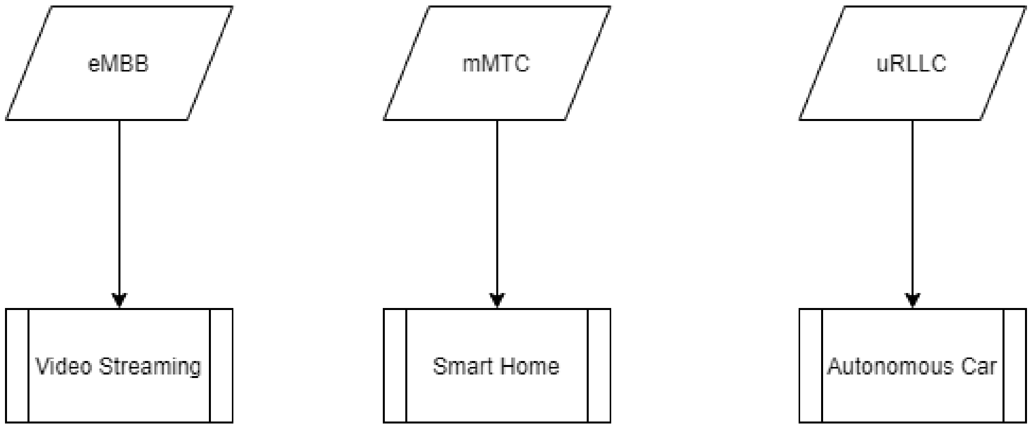

Figure 3.2.1: Use Cases

Figure 3.2.1 shows the slice and use-case combination considered in this project. The slices can be referred to as eMBB, mMTC, uRLLC or slice 1, slice 2 and slice 3 respectively hence the terms will be used interchangeably for the rest of the report.

Each slice $x^h \in X^h$ which belongs to a MVNO h has a capacity denoted by C_x hence the total capacity of the MVNO is

$$C_h = \sum_{1}^{X} C_x \qquad \textbf{3.2.3}$$

Each slice has a set of users U_x^h . The users are strictly dependent on the slice and they are denoted by

$$U_x^h = \{1,2,\dots,|U_x^h|\} \qquad \textbf{3.2.4}$$

When admitted into the slice, each user $u \in U$ demand a bandwidth of B_u and the slice allocates a fraction F_x of its total bandwidth to the user. Following the assumption that users are uniformly distributed, the probability of a user $u \in U$ belonging to a slice $x^h \in X^h$ which belongs to MVNO h is

$$p_b = \frac{1}{X^h} \qquad \textbf{3.2.5}$$

In slices that support handoff calls (eMBB and uRLLC), handoff calls are given priority over new calls because users are less tolerant to dropped handoff calls hence a bandwidth reservation policy is implemented.

3.2.1 Bandwidth Reservation Policy

In order to meet the QoS requirements of users, a bandwidth reservation policy that prioritises handoff calls is adopted. The adopted bandwidth reservation policy follows the guard channel approach described in section 2.4.7. Users are more sensitive to dropped handoff calls than blocked new calls thus handoff calls are given higher priority over new calls [53]. For a slice x with total capacity C_x, thresholds are set for both new and handoff calls. Let T_{ux} denote the threshold for rejecting new calls from user u in slice x and L_x denote the load at each slice. The bandwidth available for new calls is denoted by

$$N_x = T_{ux} - L_x \qquad \textbf{3.2.6}$$

The bandwidth available for handoff calls is denoted by

$$H_x = C_x - L_x \qquad \textbf{3.2.7}$$

For the mMTC slice which doesn't support handoff calls as mentioned in the assumptions, the bandwidth available for new calls is denoted by

$$N_{xm} = C_x - L_x \qquad \textbf{3.2.8}$$

Intra Slice handover Flow Chart.

The intra slice handover flow chart follows the assumptions introduced in section 3.1.1. Furthermore, only three types of users make request for calls to the use cases defined.

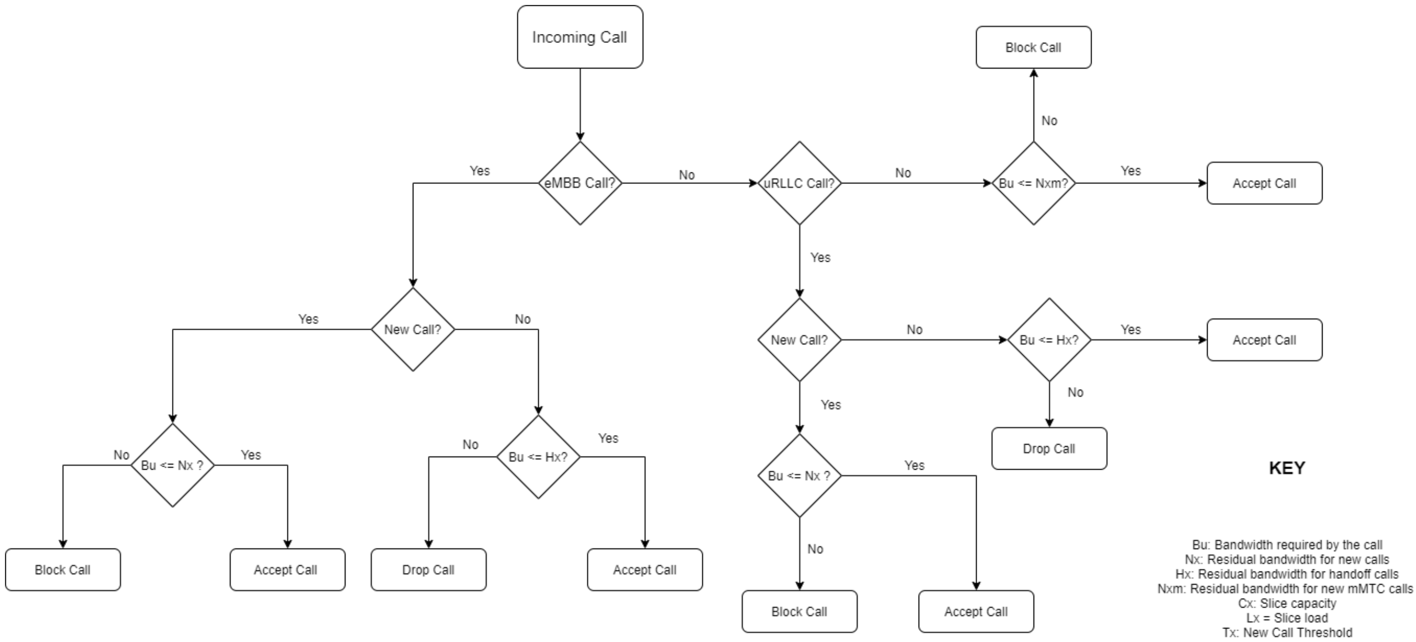

Figure 3.2.2: Intra Slice CAC flow chart

Inter Slice CAC flow chart.

The Inter slice handover CAC flow chart follows the assumptions mentioned in section 3.1.1. Furthermore, mMTC slice calls can also be admitted into the uRLLC and eMBB slices because the QoS requirements of mMTC slice calls can be satisfied in both the eMBB and uRLLC slice. eMBB slice calls can be admitted in the uRLLC slice because its QoS requirements can be satisfied by the uRLLC slice and no slice can satisfy the QoS requirements of the uRLLC slice hence uRLLC slice calls cannot be handed over to any slice.

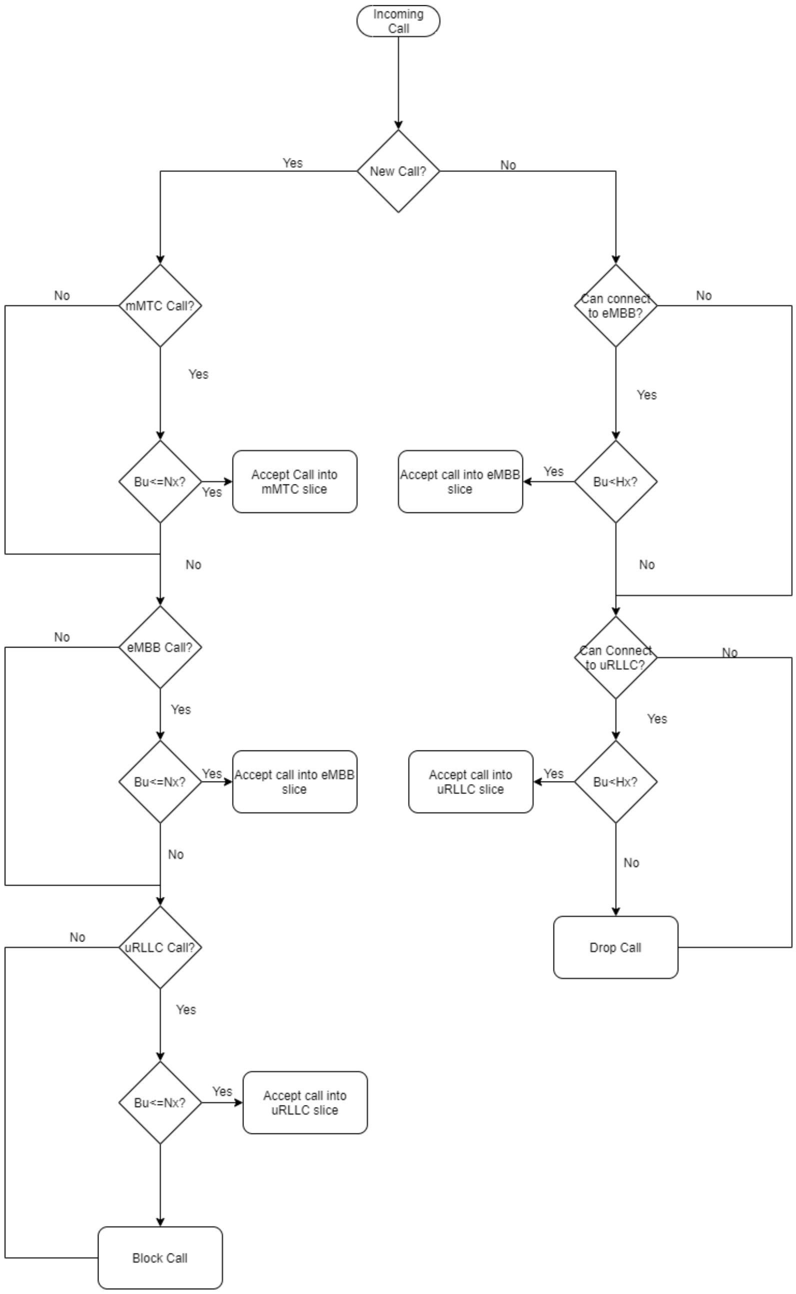

Figure 3.2.3: Inter Slice CAC flow chart

3.2.2 Slice Parameters

- **Threshold –** Wireless network subscribers are more intolerant to dropped handoff calls than blocked new calls hence there is need to give higher priority to handoff calls. In order to meet the QoS requirements of users, a threshold is used to prioritise handoff calls over new calls. The threshold value can be a fixed value, or it can be set dynamically in a slice. Threshold values are set in uRLLC and eMBB slices because they support handoff calls.

  In this report, a fixed threshold is used. The threshold value is passed as a parameter when the slice is initialised during simulation.

- **Capacity –** Slice capacity refers to the maximum amount of radio resources that can be offered by the slice. The total capacity of the MVNO is the sum of all 3 slices as illustrated by 3.2.3. The slice capacity varies from slice to slice depending on the amount of resources required by the calls.

- **Basic Bandwidth Unit (BBU) –** This describes the bandwidth requirements of calls. Different calls requesting for services from slices have different bbu requirements because they require different services. The total bandwidth offered by the slice is equal to the capacity of the slice hence new and handoff calls are only accepted if there is enough bandwidth to accommodate the call.
- **Mean Channel Holding Time -** This refers to the amount of time a call spends in the channel. Channel holding time follows an exponential distribution.
- **Handoff Rate –** This is an exponentially distributed variable $\frac{1}{h}\frac{min}{call}$.
- **Arrival Rate -** The call arrival rate is the average number of calls that are arriving in the network. Arrival rate follows the poison distribution.
- **Departure rate –** This refers to the average rate at which users end calls.

3.3 Markov Model

The markov model can be used for performance evaluation and reliability analysis [58]. Markov decision models are used to design models which are state space based because it consists of some states and the markov decision model can clearly show the transition between different states . The arrival of calls in wireless networks follows a poison distribution hence it can be modelled an M/M/m/m queueing model which is more suitable for the JCAC. In [59], Zhang *et al* identified that a multi-dimensional markov model is more suitable for modelling heterogenous networks because the model can cater for different types of calls, different slices and both new and hand off calls.

In Figure 3.3.1, the two-dimensional markov model illustrates the parameters for the slice. C is the total capacity of the slice; H is the bandwidth reserved for handoff calls; λ_n, λ_h represent the call arrival rate for new calls and handoff calls respectively and μ_n, μ_h represent the departure rate of new calls and handoff calls respectively.

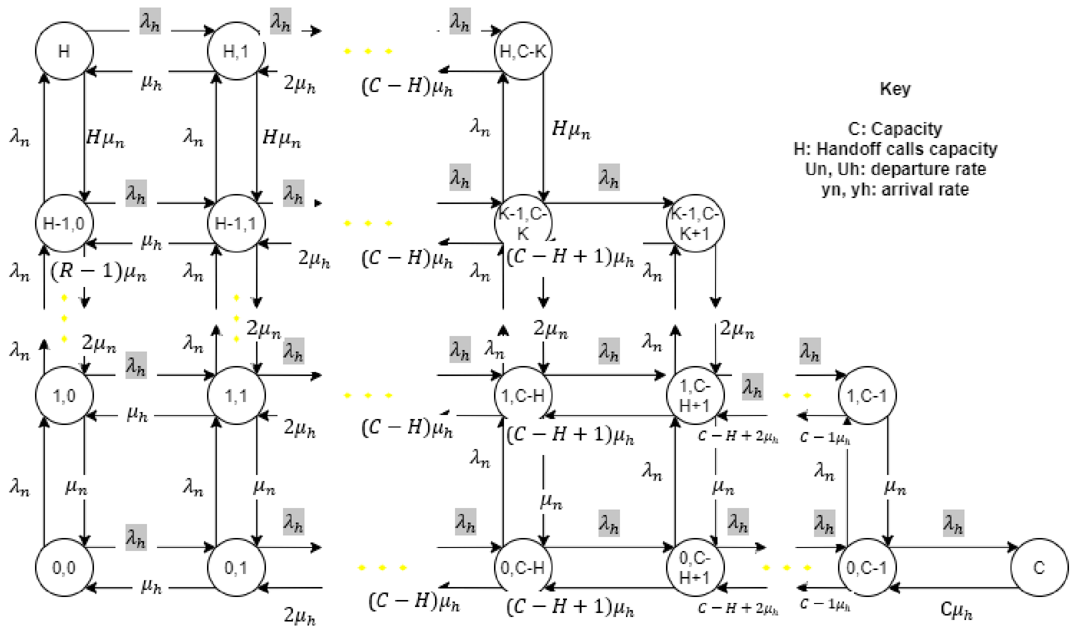

Figure 3.3.1: 2D Markov Chain Model [60]

The CAC scheme described in section 3.2 is a multi-dimensional state chain and can be modelled by a markov chain similar to that in Figure 3.3.1 with the current state being given by:

$$\Omega = (n_{u,x}, h_{u,x}, u = 1, \dots, U, x = 1, \dots, X) \qquad \textbf{3.3.1}$$

with $n_{u,x}$ and $h_{u,x}$ denoting the number of ongoing new and handoff calls in the slices.
The admissible state for the network can be represented by a vector S. The admissible states refer to a set of all possible number of new and handoff calls that can be admitted into the network at the same time.

Following 3.3.1, the set of all possible states S is given by:

$$S = \left\{ \sum_{u=1}^{U} \sum_{x=1}^{X} (b_u h_{u,x} + b_u n_{u,x} \le C_x) \wedge \sum_{u=1}^{U} \sum_{x=1}^{X} (b_u h_{u,x} \le T_h) \wedge \sum_{u=1}^{U} \sum_{x}^{X} (b_u n_{u,x} \le T_n) \right. \qquad \textbf{3.3.2}$$

with b_u, C_x, T_n, T_h denoting required bandwidth, slice capacity, new call threshold and threshold for handoff calls respectively. The probability of *P(s)* of the system being in the state $s(s \in \Omega)$ is expressed by:

$$P(s) = \frac{1}{G} \left(\prod_{u=1}^{U} \prod_{x=1}^{X} \left(\frac{(\rho n_{u,x})^{n_{u,x}}}{n_{u,x}!} \right) \left(\frac{(\rho h_{u,x})^{h_{u,x}}}{h_{u,x}\,!} \right) \right) \qquad \textbf{3.3.3}$$

With $\rho n_{u,x}$ and $\rho h_{u,x}$ denoting the loads generated at the slices and they are given by:

$$p n_{u,x} = \frac{\lambda^n(u,x)}{\mu^n(u,x)} \ \forall\, u, x \qquad \textbf{3.3.4}$$

$$p h_{u,x} = \frac{\lambda^h(u,x)}{\mu^h(u,x)} \ \forall\, u, x \qquad \textbf{3.3.5}$$

The variable G is the normalisation constant denoted by:

$$G = \sum_{(s \in \Omega)} \prod_{u=1}^{U} \prod_{x=1}^{X} \left(\frac{(\rho n_{u,x})^{n_{u,x}}}{n_{u,x}!} \right) \left(\frac{(\rho h_{u,x})^{h_{u,x}}}{h_{u,x}\,!} \right) \qquad \textbf{3.3.6}$$

3.3.1 Action Spaces

The CAC algorithm is responsible for deciding if a call is accepted or rejected. A call is accepted into a resource constrained network if and only if the slice has enough bandwidth to accommodate the call without violating the commitments made to ongoing calls on the network.

$$A = \{a = \{a_1^n, \dots, a_u^n, a_1^h, \dots, a_u^h\}\} \qquad \textbf{3.3.7}$$

$$a_u^n, a_u^h \in (0,1, \dots, (u+1), \dots, U) \qquad \textbf{3.3.8}$$

With a_u^n being the action taken on arrival of new calls and a_u^h being the action taken on arrival of handoff calls.

$$a_u = \begin{cases} 0; reject\ call \\ 1; Accept\ call\ into\ slice\ 1 \\ U; Accept\ call\ into\ slice\ U \end{cases}$$

3.4 Evaluation Metrics

The performance of the CAC algorithm is evaluated by simulating the algorithms in Python. The performance of each algorithm *(inter and intra slice)* is analysed using the Handoff Call Dropping Probability (HCDP) and Newcall Blocking Probability (NCBP). For better QoS, the evaluation metrics are kept as low as possible.

3.4.1 New Call Blocking Probability (NCBP)

NCBP is the probability that new call is not admitted (blocked). A new call is blocked if the slice configured to admit the call does not have enough radio resources to admit the call i.e. when maximum capacity is exceeded. Let $S_{bu} \subset \Omega$, denote the state in which new calls can be blocked, the blocking condition can be expressed as:

$$S_{bu} = \{s \in S: \left(\left(b_u + \sum_{u=1}^{U} n_{u,x} b_u > T_n\right) \vee \left(b_u + \sum_{u=1}^{U} (n_{u,x} + h_{u,x}) \times b_u > C_x\right)\right) \forall x \qquad \textbf{3.4.1}$$

Hence the new call blocking probability is given by:

$$P_b(s) = \sum_{s \in S_{bu}} P(s) \qquad \textbf{3.4.2}$$

3.4.2 Handoff Call Blocking Probability (HCBP)

HCBP is the probability that a handoff call is not accepted (dropped). A handoff call is dropped if the slice configured to admit the handoff call does not have enough bandwidth to admit the call i.e. when maximum capacity is exceeded. $S_{du} \subset \Omega$, denote the state in which handoff calls can be dropped, the dropping condition can be expressed as:

$$S_{du} = \left\{ s \in S: \left(\left(b_u + \sum_{u=1}^{U} b_u h_{u,x} > T_h \right) \vee \left(b_u + \sum_{u=1}^{U} b_u \times (h_{u,x} + n_{u,x} > C_x \right) \right) \forall x \right. \quad \textbf{3.4.3}$$

Hence the handoff call dropping probability is given by:

$$P_d(s) = \sum_{s \in S_{du}} P(s) \quad \textbf{3.4.4}$$

4. Simulation Design

In chapter 3 the design of the network model and various evaluation metrics were outlined. This chapter focuses on the implementation of the algorithm.

4.1 Programming Language

Python programming language was used to implemented and simulate the algorithm designed in chapter 3. The programming language was chosen because of its simplicity and diverse applications, some of the language's advantages include productivity, speed, presence of third-party modules, availability of support libraries and a rich internet resource bank. In the context of productivity and speed, Python provides process control capabilities and a clean object-oriented design. Being an object-oriented programming language, Python allows breaking down of the program and incorporates methods like inheritance and polymorphism. Inheritance increases code reusability and development speed by allowing new (base) classes to inherit methods from existing (super) classes and polymorphism is the ability of to process objects differently based on data types [61]. Furthermore, Python allows building of modular programs. Modularity makes it easy to troubleshoot programs by allowing developers to change other classes without affecting other classes. This can be useful in sliced networks where there is need to change properties of the physical infrastructure without changing the slices.

4.2 Simulation Environment

The PyCharm Community Integrated Development Environment (IDE) was used to simulate the algorithm. PyCharm IDE was chosen because it provides useful plugins like auto complete and other productive shortcuts. Furthermore, PyCharm comes with a built in terminal and virtual environments making it easy to install packages. PyCharm is easy to install and runs on any computer as long as the python interpreter is installed.

The program was developed on a Dell Inspiron 15 3000 with the following specifications

System Specifications	
Operating system	Windows 10 Home (2019)
processor	Intel(R) Core (TM) i5-7200 CPU @2.50GHz2.70GHz
Installed memory(RAM)	8GB (7.87GB usable)
System type	64 Bit Operating System, x64-based processor

Table 4.2.1: System Specifications

4.3 Algorithm Implementation

This subsection describes the classes developed in order to meet the requirements of the handover algorithm introduced in section 3. The implemented algorithm comprises of 6 classes namely Simulator, UseCases, Slice, Slices, IntraCalculations and InterCalculations. Each class has methods associated with that class and the slices class and InterCalculations classes inherit and override methods from the Slice and IntraCalculations classes respectively. Inheriting methods from the super classes improved code usability and overriding methods from the super classes allowed the child classes to have their own model without changing the methods in the parent class.

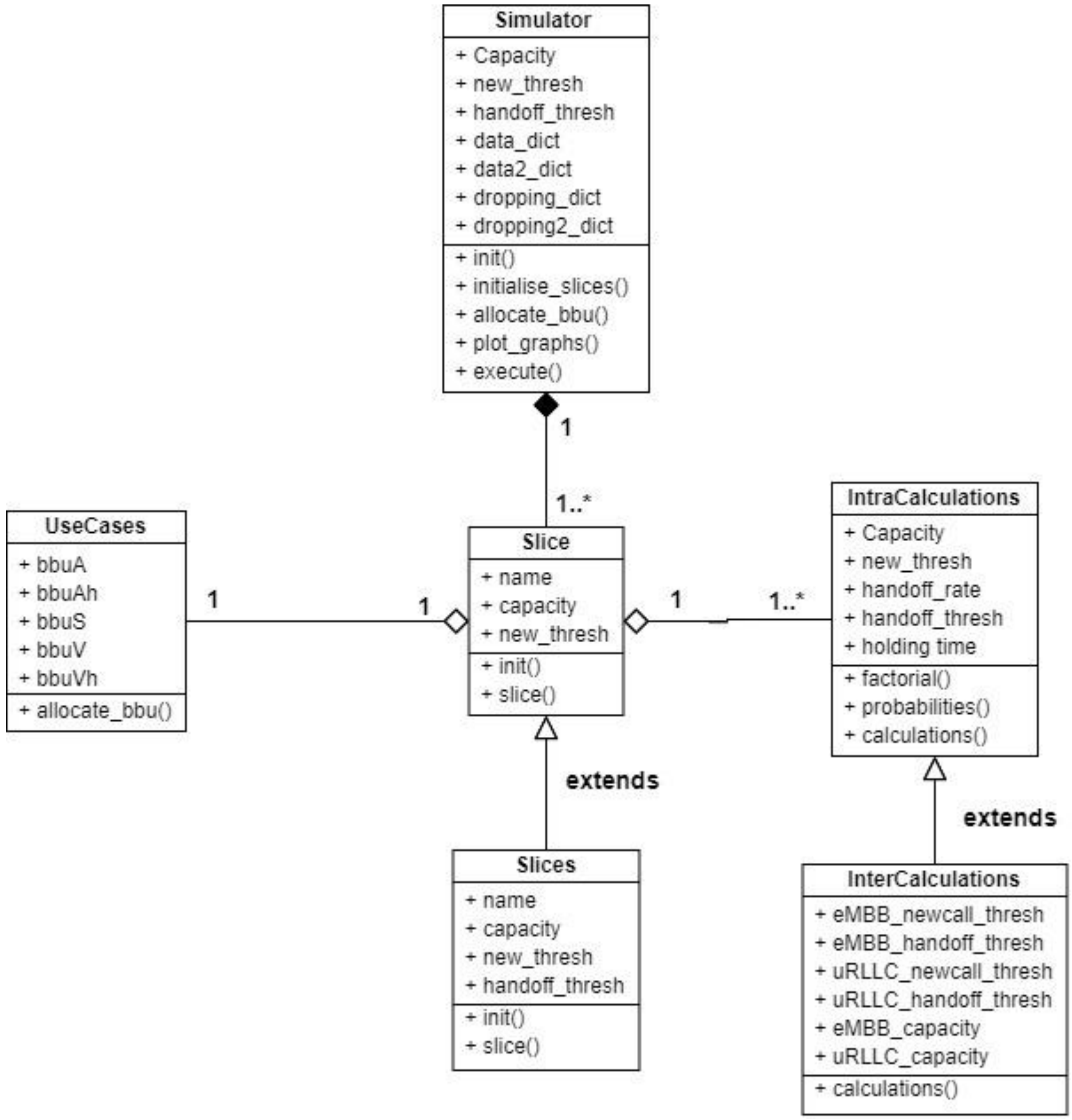

Figure 4.3.1: Intra and Inter Slice handover UML Class Diagram

4.3.1 Simulator Class

The Simulator class is the main class of the program which is used to run the program. It is made up of five methods namely init, allocate_bbu, initialise_slices, plot_graphs and execute. The init method is a reserved method in Python and it is called whenever a simulator class object is created, and it allows the class to initialise the attributes of the class. The initialise_slices method is responsible for creating slice objects. Given name, capacity, new call threshold and handoff call threshold as parameters, the method creates a slice object. eMBB and uRLLC slices take in all four defined parameters because they support handoff calls, whereas the mMTC slice only support new calls hence it excludes the handoff threshold parameter.

```
@property
    def initialise_slices(self):
        """Function that initialises the slices"""
        slices = ['eMBB', 'uRLLC', 'mMTC']
        for slice_name in slices:
            if slice_name is 'eMBB' or slice_name is 'uRLLC':
                # create Slices object
                initialise = Slices(slice_name, self.capacity, self.new_thresh, self.handoff_thresh).slice

            else:
                # create a Slice object
                initialise = Slice(slice_name, self.mMTC_cap, self.mMTC_thresh).slice

        # return a dict with slice name as the key and an array of parameters as values
        return initialise
```

Listing 4.3.1: initialise_slices method

The allocate_bbu method takes in slice name as a parameter and returns the bandwidth required by calls for that specific slice.

```
@staticmethod
    def allocates_bbu(slice_name):

        useCase = UseCases()
        bbu = useCase.allocate_bbu(slice_name)
        return bbu
```

Listing 4.3.2: allocate_bbu method

When the algorithm is executed, the results are presented in text form and graphical forms. The graphs are plotted using the plot_graphs method in code Listing 4.3.3. The method takes in the algorithm name and a data frame containing the results of the simulation as parameters.

```
@staticmethod
    def plot_graphs(algorithm, df):

        print(f'Plotting {algorithm} Slice Handover Algorithms\n')
        fig1 = plt.figure(1)
        plt.plot(df['index'], df['eMBB_BP'])
        plt.xlabel('Call Arrival Rate')
        plt.ylabel('Call Blocking Probabilty')
        plt.title(f'Effect of Arrival Rate on New Call Blocking Probability')
        plt.grid()
        plt.show()
```

Listing 4.3.3: plot graphs method

The execute method is the main method of the program, it is responsible for displaying the calculated probabilities and making calls to the appropriate methods in the Simulator class. The method creates objects of the Calculations class with the appropriate slice parameters.

```
@property
    def execute(self):
        slice_type = self.initialise_slices

        # Choose Algorithm to Simulate.
        print('Choose the type of handover you want to simulate: ')
        print('1. Intra Slice handover')
        print('2. Inter Slice handover')

         choice = eval(input("Enter your choice: "))
         if choice == 1:
             print('Simulating Intra Slice handover\n')
             # Create the IntraSlice object
             calc = IntraCalculations()
             algorithm = 'Intra'
             for key in slice_type:
                 if key == 'eMBB' or key == 'uRLLC':
                     # pass slice parameters
                     probabilities = calc.calculations(key, slice_type[key])
                     # display probabilities
                     print(f'The blocking probability for {probabilities[0]} slice is: {probabilities[1]} \n')
                     print(f'The dropping probability for {probabilities[0]} slice is: {probabilities[2]} \n')

                     # update data dictionaries
                     self.data_dict.update({key: probabilities[1]})
                     self.dropping_dict.update({key: probabilities[2]})
                 else:
                     probabilities = calc.calculations(key, slice_type[key])
                     print(f'The blocking probability for {probabilities[0]} slice is: {probabilities[1]} \n')
                     self.data_dict.update({key: probabilities[1]})
```

Listing 4.3.4: execute method

4.3.2 Slice and Slices Class

The slice and slices class are put together because the slices class inherits from the slice class as illustrated by Figure 4.3.1 The slice class is the superclass and contains a slice method which specifies the slice in terms of its name, capacity and new call threshold. The slices class is the subclass and it contains a slice method which overrides the slice method in the superclass. The slices class inherits the name, capacity and new call threshold parameters from the superclass and adds the handoff call threshold parameter. The classes were separated for ease of initialising the slices because the mMTC slice has three parameters and the uRLLC and eMBB slice has four parameters hence separating the classes makes the program more robust. The init method in both methods is responsible for initialising the attributes.

```python
class Slice:
    slice_dict = {}

    def __init__(self, name, capacity, new_thresh):
        # initialise attributes
        self.name = name
        self.capacity = capacity
        self.new_thresh = new_thresh
        # self.

     @property
     def slice(self):
         # Pass slice parameters
         self.slice_dict.update({self.name: [self.capacity,
self.new_thresh]})
         return self.slice_dict
```

Listing 4.3.5: Slice Class

```python
class Slices(Slice):
    def __init__(self, name, capacity, new_thresh, handoff_thresh):
        # inherit from parent class
        super().__init__(name, capacity, new_thresh)
        self.handoff_thresh = handoff_thresh

    @property
    def slice(self):
        # Pass slice parameters
         self.slice_dict.update({self.name: [self.capacity,
self.new_thresh, self.handoff_thresh]})
         return self.slice_dict
```

Listing 4.3.6: Slices Class

4.3.3 UseCases Class

The UseCases class specifies the basic bandwidth units (bbu) required by calls at each slice. The bbu values are specified through the allocate_bbu method. In a slice that supports handoff calls, the bbu demanded by new calls is assumed to be equal to the bbu demanded by handoff calls because the calls are originating from similar UEs and require the same resources. However, calls for different slices have different bbu requirements because they require different resources. Different bbu values are allocated depending on the slice name.

```
class UseCases:
    bbuV = 3 # new call bbu for eMBB
    bbuS = 1 # new call bbu for mMTC
    bbuA = 2 # new call bbu for uRLLC
    bbuAh = 2 # handoff call bbu for uRLLC
    bbuVh = 3 #handoff call bbu for eMBB

    @classmethod
    def allocate_bbu(cls, names):
         if 'mMTC' in names:
             return cls.bbuS
         elif 'eMBB' in names:
             return cls.bbuV, cls.bbuVh
         elif 'uRLLC' in names:
             return cls.bbuA, cls.bbuAh
         else:
             return None
```

Listing 4.3.7: UseCases Class

4.3.4 IntraCalculations Class

The IntraCalculations class is responsible for all the intra slice handover computations. The class is made up of four methods namely __init__, factorial, probabilities and calculations. The __init__ method is responsible for indicating to the user that the calculations are in progress i.e. it runs when the IntraCalculations object is created in the execute method described in section 4.3.1. The factorial method is used to calculate the factorial of a given number.

```
@classmethod
    def factorial(cls, num):
        """Function that calculates the Factorial of a given number"""
        if num == 1 or num == 0:
            return 1
        else:
            return num * cls.factorial(num - 1)
```

Listing 4.3.8: Factorial Method

The probabilities method is used to calculate the new call blocking and handoff call dropping probabilities.

```
@classmethod
    def probabilities(cls, numerator: object, denominator: object) -> object:
        """Function that calculates the probabilities"""
        probability = numerator / denominator
        return probability
```

Listing 4.3.9: Probabilities Method

The ultimate method in the IntraCalculations class is the calculations method. The method is responsible for calculating the required probabilities, it takes in the slice name and slice parameters as its method arguments. The method initialises NumPy arrays for the blocking probabilities, dropping probabilities, normalisation factor, new and handoff call arrival rates and network load amongst others. The handoff and new call arrival rate refer to the rate at which the calls arrive at the slices. mMTC slice calls are assumed to arrive at a rate two times greater than eMBB and uRLLC calls. The assumption is supported by the fact that “massive” part of the mMTC name meaning a lot of calls arrive at the network because the devices requesting for services are usually densely populated.

The method extracts the slice parameters based on the slice name received.

```
        if slice_name == 'eMBB':
            required_bbu = UseCases()
            bbuN = required_bbu.allocate_bbu(slice_name)[0]  # eMBB new call bbU
            bbuH = required_bbu.allocate_bbu(slice_name)[1]  # eMBB handoff call bbu
            bbuNU = required_bbu.allocate_bbu('uRLLC')[0] # uRLLC new call bbu
            bbuHU = required_bbu.allocate_bbu('uRLLC')[1] # uRLLC handoff call bbu
            bbuNM = required_bbu.allocate_bbu('mMTC') # mMTC bbu

            cls.capacity = params[0]  # slice capacity
             cls.new_thresh = params[1]  # new call threshold
             cls.handoff_thresh = params[2]  # handoff call threshold
```

Listing 4.3.10: Extracting slice parameters

After extracting the parameters for all the slices, the method uses the parameters to determine the networks admissible state. The admissible state follows equation 3.3.2. If the condition for the admissible state are satisfied, the call is deemed to be in the admissible state.

```
# Admissible states
    if (bbuN * (newC + handC) <= cls.capacity) and \
          (bbuNM * newCM <= cls.mMTC_capacity) and \
          (bbuNU * (newCU + handCU) <= cls.uRLLC_capacity) and \
          (bbuN * newC <= cls.new_thresh) and \
          (bbuNU * newCU <= cls.uRLLC_newcall_thresh):
```

Listing 4.3.11: Intra Slice Network Admissible State

When the call is in the admissible state, the normalisation constant for admissible state is calculated and the algorithm checks for the conditions to block new calls and drop handoff calls in the slices. The conditions for blocking new calls and dropping new calls follow equations 3.4.1 and 3.4.3 respectively.

```
# Condition for blocking new calls
    if (bbuN + bbuN * (newC + handC) > cls.capacity) or \
            (bbuN + bbuN * newC > cls.new_thresh):
             G1[j] = G1[j] + qn[j] + qh[j] + qh_1[j] + qn_1[j] + qn_2[j]
```

Listing 4.3.12: Intra Slice eMBB new call blocking condition

```
# Condition for dropping handoff calls
        if (bbuN + bbuN * (newC + handC) > cls.capacity) or \
             bbuN + bbuN * handC > cls.handoff_thresh:
             G2[j] = G2[j] + qn[j] + qh[j] + qh_1[j] + qn_1[j] + qn_2[j]
```

Listing 4.3.13: Intra Slice eMBB handoff call dropping condition

```
# Condition for dropping new calls
    if bbuNM + bbuNM * newCM > cls.mMTC_capacity:
            G1[j] = G1[j] + qn_2[j] + qn_1[j] + qn[j] + qh[j] + qh_1[j]
```

Listing 4.3.14: Intra Slice mMTC new call dropping condition

```
# Condition for blocking new calls
    if (bbuN + bbuN * (newC + handC) > cls.capacity) or \
             (bbuN + bbuN * newC > cls.new_thresh):
             G1[j] = G1[j] + qn[j] + qh[j] + qh_1[j] + qn_1[j] + qn_2[j]
```

Listing 4.3.15: Intra Slice uRLLC new call blocking condition

```
# Condition for dropping handoff calls
   if (bbuN + bbuN * (newC + handC) > cls.capacity) or \
            bbuN + bbuN * handC > cls.handoff_thresh:
            G2[j] = G2[j] + qn[j] + qh[j] + qh_1[j] + qn_1[j] + qn_2[j]
```

Listing 4.3.16: Intra Slice uRLLC handoff call dropping condition

4.3.5 InterCalculations Class

The InterCalculations class is responsible for the inter slice handover computations in the program. The class inherits the factorial, __init__ and probabilities methods from the IntraCalculations (super) class. The calculations method in InterCalculations class overrides the method in the superclass because the method adds some parameters to the calculations. The method initialises the NumPy arrays and extracts values in the same way as the method in the superclass (Listing 4.3.10). The call arrival of mMTC slice follows the same assumption described in 4.3.4.

After extracting the parameters, the method determines the networks admissible state. The network admissible state follows equation 3.3.2. If the admissible state equations are satisfied, the network is deemed to be in the admissible state.

```
# Admissible state condition
      if (bbuNU * (newCU + handCU) + bbuNM * newCM + bbuN * (
                 newCE + handCE) <= cls.uRLLC_capacity) and (
                 bbuN * (newCE + handCE) + bbuNM * newCM <=
cls.eMBB_capacity) and (
                 bbuNM * newCM <= cls.capacity) and (
                 bbuN * newCE + bbuNU * newCU + bbuNM * newCM <=
cls.uRLLC_newcall_thresh) and (
                 bbuNM * newCM + bbuN * newCE <= cls.eMBB_newcall_thresh):
```

Listing 4.3.17: Inter Slice Network Admissible State

If the conditions for admissible state are satisfied, the normalisation constant for the admissible state is calculated and the algorithm determines the conditions to block new calls and drop handoff calls. The conditions for blocking new calls and dropping handoff calls follow equation 3.4.1 and 3.4.3 respectively.

```
# Condition for blocking new calls
    if ((bbuN + bbuN * (newCE + handCE) + bbuNM * newCM > cls.capacity) or
             (bbuN + bbuN * newCE + bbuNM * newCM > cls.new_thresh)) and \
             ((bbuN + bbuNU * (newCU + handCU) + bbuN * (newCE + handCE) +
              bbuNM * newCM > cls.uRLLC_capacity) or
              (bbuN + bbuNU * newCU + bbuN * newCE +
              bbuNM * newCM > cls.uRLLC_newcall_thresh)):
```

Listing 4.3.18: Inter Slice eMBB new call blocking condition

```
# Condition for dropping handoff calls
    if (bbuN + bbuNU * (newCU + handCU) + bbuN * (newCE + handCE) +
          bbuNM * newCM > cls.uRLLC_capacity) and \
         (bbuN + bbuN * (newCE + handCE) + bbuNM * newCM > cls.capacity):
```

Listing 4.3.19: Inter Slice eMBB handoff call dropping condition

```
# Condition for blocking new calls
    if (bbuNM + bbuNM * newCM > cls.capacity) and \
         ((bbuNM + bbuNU * (newCU + handCU) + bbuN * (newCE + handCE) +
         bbuNM * newCM cls.uRLLC_capacity) or
         (bbuNM + bbuNU * newCU + bbuN * newCE +
         bbuNM * newCM > cls.uRLLC_newcall_thresh)) and \
         ((bbuNM + bbuN * (
         newCE + handCE) + bbuNM * newCM > cls.eMBB_capacity)
         or (
         bbuNM + bbuN * newCE + bbuNM * newCM > cls.eMBB_newcall_thresh)):
```

Listing 4.3.20: Inter Slice mMTC new call blocking condition

```
# condition for blocking new calls
   if (bbuNU + bbuNU * (newCU + handCU) + bbuNE * (newCE + handCE) +
        bbuNM * newCM > cls.capacity) or \
        (
        bbuNU + bbuNU * newCU + bbuNM * newCM +
        bbuNE * newCE > cls.new_thresh):
```

Listing 4.3.21: Inter Slice uRLLC new call blocking condition

```
    # condition for dropping handoff calls
   if (bbuNU + bbuNU * (newCU + handCU) + bbuNE * (newCE + handCE) +
         bbuNM * newCM > cls.capacity):
```

Listing 4.3.22: Inter Slice uRLLC handoff call dropping condition

The full implementation of the algorithm can be found in Appendix A: Simulation Program.

5. Results and Discussion

This Chapter presents the results obtained from simulating the algorithm developed in Chapter 4. The performance of the algorithm under different conditions which are determined by the network parameters is evaluated using new call blocking and handoff call dropping probabilities. The algorithm was simulated under the conditions detailed in sections 4.2 and the results obtained are illustrated on graphs generated using the matplotlib library in Python.

5.1 Scenario 1: Effect of Call Arrival Rate

In this scenario, the call arrival rates were varied whilst the parameters in Table 5.1.1 were kept constant. The values of the call arrival rate are incremented by 0.5 from 3 up to 8 thus there are 10 different call arrival rate values. The increase in call arrival rate suggests that users are initiating more calls and the demand for radio resources from networks is increasing. The new call arrival rate is the same for eMBB and uRLLC slices and twice as much in mMTC slice. Handoff call arrival rate is assumed to be equal to new call arrival rate and the departure rate for the slices is equal and kept constant.

Slice Parameters			
	eMBB	mMTC	uRLLC
Capacity	30	15	30
New Call Threshold	50%	-	50%
Handoff Call Threshold	100%	-	100%
Required BBU	3	1	2

Table 5.1.1: Scenario 1 Slice Parameters

5.1.1 Effect of Call Arrival Rate on Intra Slice Handover.

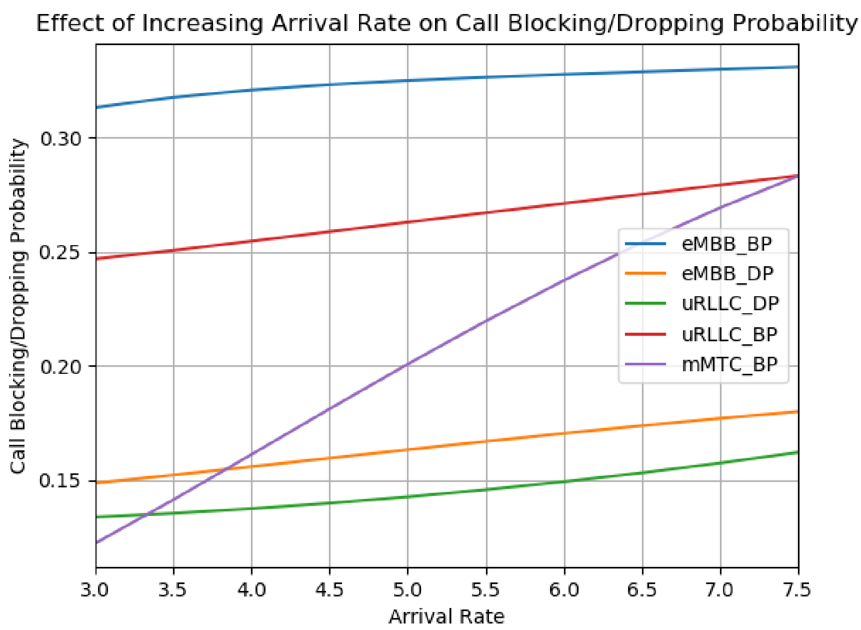

Figure 5.1.1: Effect of Increasing Call Arrival Rate on Intra Slice Call Blocking/Dropping Probability

Figure 5.1.1 illustrates the effect of increasing call arrival rate on the new call blocking and handoff call dropping probability in intra slice handover. The graphs show that the call blocking and dropping probabilities increase as the number of incoming calls increase. This is because when users initiate more calls, the network struggle to admit all the calls requesting for services because the network has finite radio resources. From the graphs. eMBB slice has the highest new call blocking probability amongst all slices because calls requesting for services in the slice has the highest bbu requirement. Despite having a higher call arrival rate, the mMTC slice has the lowest new call dropping probability because calls requesting for services in the slice have the lowest bbu requirements and the slice has no threshold for new calls. The intra slice algorithm is implemented with bandwidth reservation for handoff calls hence handoff call dropping probabilities are less than new call dropping probabilities. uRLLC has the lowest handoff call dropping probability, this can be attributed to the fact that calls for the uRLLC slice require less bbu compare to calls for the eMBB slice.

5.1.2 Effect of Call Arrival Rate on Inter Slice handover

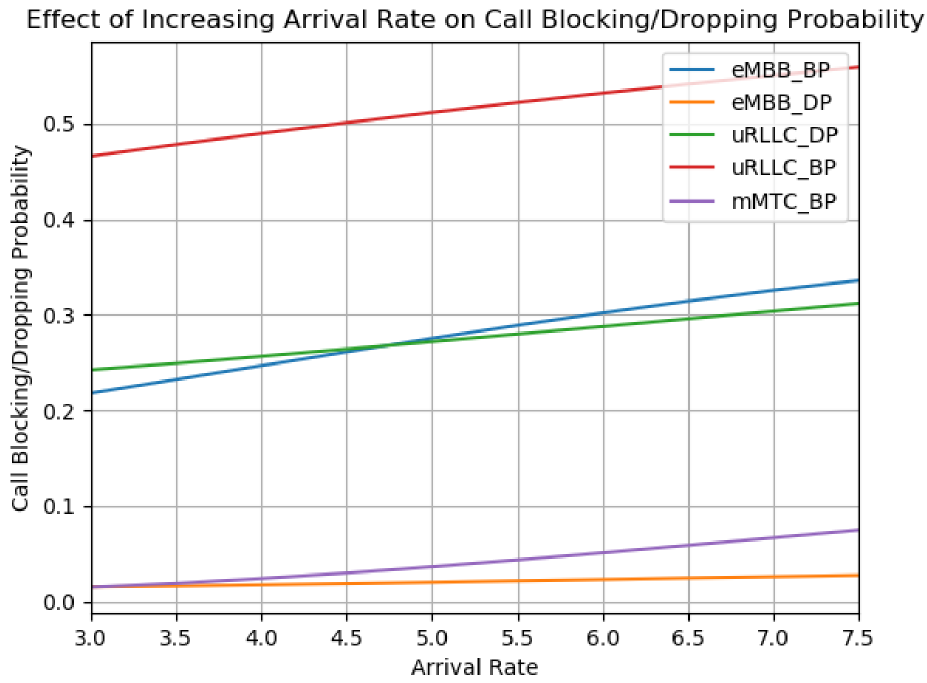

Figure 5.1.2: Effect of Increasing Call Arrival Rate on Inter Slice Call Blocking/Dropping Probability

Figure 5.1.2 illustrates the effect of increasing call arrival rate on the new call blocking and handoff call dropping probability in inter slice handover. The graphs show that the call blocking and dropping probabilities increase as the number of calls increase. This can be attributed to the network having finite resources hence when the number of incoming calls increase, more calls are blocked and dropped because the network cannot admit all calls. uRLLC has the largest call blocking probability because in inter slice, uRLLC slice can offer services to calls for mMTC and eMBB slice hence more calls arrive at uRLLC slice. mMTC has the lowest call blocking probability because the slice has no threshold for new calls and the calls require the least amount of bbu. Amongst the slices with bandwidth reservation, eMBB has the lowest call dropping probability despite having calls that demand the most bbu. This can be attributed to the fact that calls for the eMBB slice can be admitted into the uRLLC slice thus before being dropped from the eMBB slice, the call can be admitted in the uRLLC slice thus reducing the call dropping probability.

5.2 Scenario 2: Effect of Threshold

In this scenario, the threshold for new calls is varied whilst the parameters in Table 5.2.1 are kept constant. The threshold values are increased by 10% in every iteration with the minimum threshold being 6 and the maximum threshold being 30 (maximum capacity). The increase in threshold suggest that more new calls will be admitted hence the new call blocking probability should decrease whilst the handoff call blocking probability should increase.

Slice Parameters			
	eMBB	mMTC	uRLLC
Capacity	30	15	30
Handoff Call Threshold	100%	-	100%
Required BBU	3	1	2
Call Arrival Rate	3	3	3

Table 5.2.1:Scenario 2 Slice Parameters

5.2.1 Effect of Threshold on Intra Slice Handover.

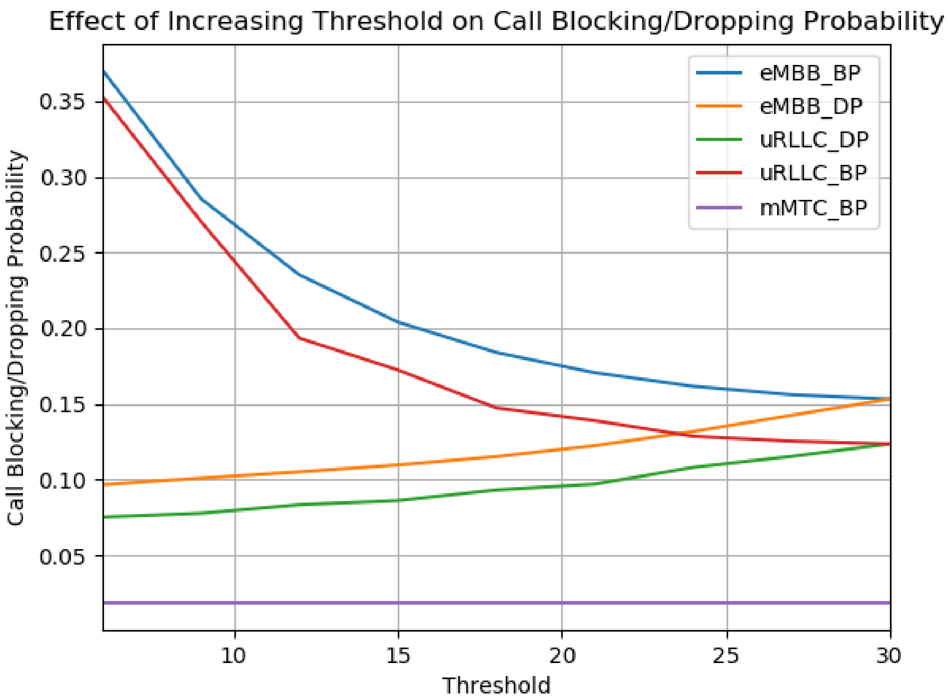

Figure 5.2.1: Effect of Increasing Threshold on Intra Slice Call Blocking/ Dropping Probability

Figure 5.2.1 illustrates the effect of increasing Threshold on the new call blocking and handoff call dropping probabilities. The graphs show that the new call blocking probabilities decrease as the Threshold increase, this is because as the threshold for new calls increase, the network can admit more new calls. From the results it should be noted that there is a slight increase in the handoff call dropping probability, this is because as the threshold increases, the bandwidth reserved for handoff calls decreases hence more handoff calls are dropped. The blocking probability of mMTC slice is constant because the slice has no new call threshold. At 30 (*maximum capacity*) the probability of blocking new calls is equal to the probability of dropping handoff calls, this is because the when the threshold is equal to the slice capacity, there is no bandwidth reservation hence no priority is given to handoff calls.

5.2.2 Effect of Threshold on Inter Slice Handover

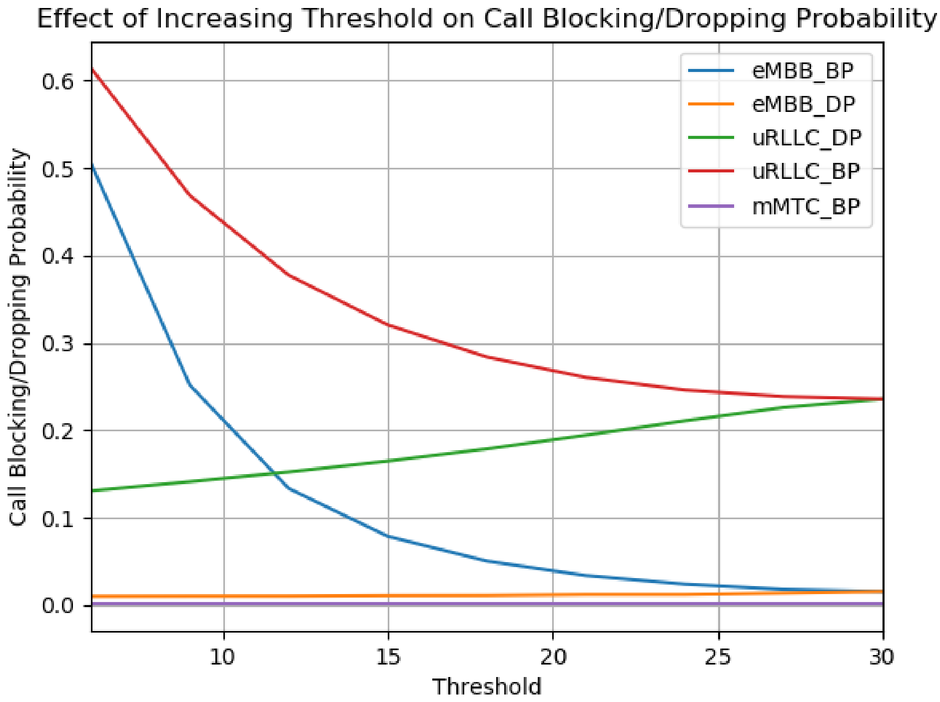

Figure 5.2.2: Effect of Increasing Threshold on Inter Slice Call Blocking/Dropping Probability

Figure 5.2.2 illustrates the effect of increasing Threshold on the new call blocking and handoff call dropping probability. The graphs show that the new call blocking probability decrease as the Threshold increase, this can be attributed to the fact that an increase in new call threshold means more new calls are accepted in the network. However, as the new call blocking probability decrease, handoff call dropping probability increases because the bandwidth reserved for handoff calls decreases. uRLLC slice has the highest call blocking and dropping probabilities because the slice caters for calls from other slices hence the slice receives more calls. mMTC has a constant call blocking probability because the slice has no new call threshold. When the threshold reaches 30 (*100% of the capacity*) new call blocking probability is equal to handoff call dropping probability because there is no handoff call priority when threshold is equal to capacity.

5.3 Scenario 3: Effect of Capacity

In this scenario, the capacity of each slice is varied whilst the threshold for new calls in slices that support handoff calls is kept as a percentage of the capacity and the other parameters in are kept constant. The values of capacity are increased by 10 bbu each time New call threshold is assigned as 50% of the slice capacity to ensure better radio resource utilisation of resources and enhance QoS by reserving enough bandwidth for handoff calls. An increase in capacity suggest that the network can admit more calls hence the call blocking and dropping probabilities should decrease.

Slice Parameters			
	eMBB	mMTC	uRLLC
New Call Threshold	50%	100%	50%
Handoff Call Threshold	100%	-	100%
Required BBU	3	1	2
Call Arrival Rate	3	3	3

Table 5.3.1: Scenario 3 Slice Parameters

5.3.1 Effect of Capacity on Intra Slice Handover

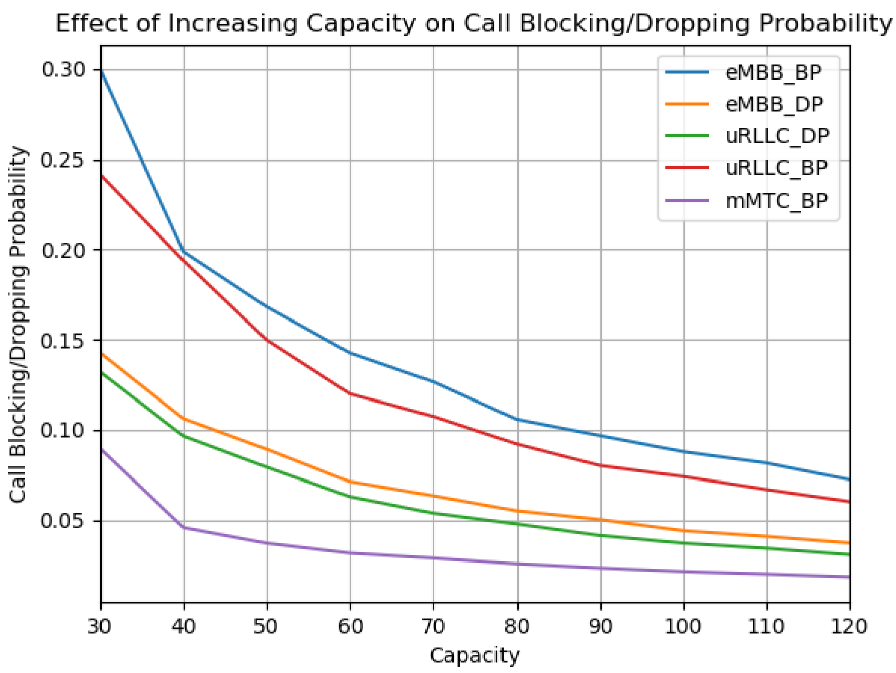

Figure 5.3.1: Effect of Increasing Capacity on Intra Slice Call Blocking/Dropping Probability

Figure 5.3.1 illustrates the effect of increasing capacity on new call blocking and handoff call dropping probability. Form the graph, it is evident that increasing the network capacity has a significant effect on call blocking and dropping probabilities. Increasing the capacity results in improved QoS because of the decrease in the blocking and dropping probabilities. The probabilities decrease because the network has more bandwidth to accept more calls. eMBB slice has the highest blocking and dropping probabilities because calls for the slice demand the most bandwidth. Despite having the highest call arrival rate, mMTC slice has the lowest new call blocking probability because the calls for the slice demand less bbu and the slice has no bandwidth reservation.

5.3.2 Effect of Capacity on Inter Slice Handover

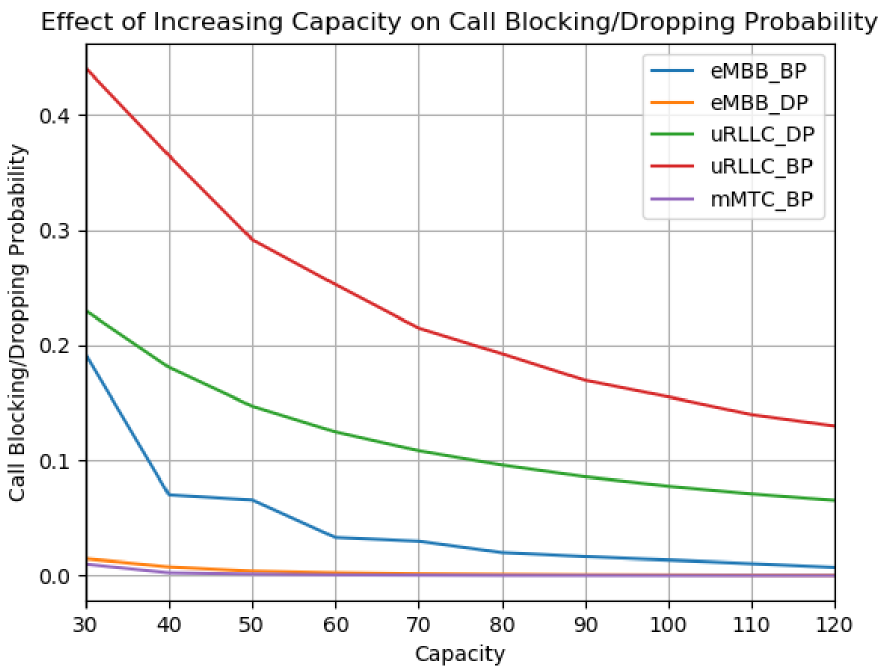

Figure 5.3.2: Effect of Increasing Capacity on Inter Slice Call Blocking/Dropping Probability

Figure 5.3.2 illustrates the effect of increasing slice capacity of new call blocking and handoff call dropping probability. The graphs show a decrease in the call blocking and dropping probabilities as the capacity increases. This can be attributed to the fact that as the capacity increases, the network can accept more calls hence less calls are blocked or dropped. uRLLC slice has the highest blocking and dropping probability because the slice caters for calls from the other slices hence more calls arrive at the slice. eMBB calls demand the most bbu but they do not have the highest blocking or dropping probability because the calls can also be admitted into the uRLLC slice hence less calls can be blocked or dropped. mMTC slice calls has the least blocking probability because the calls demand the least bbu, furthermore, the calls can also be admitted in other slices hence less calls are blocked.

5.4 Scenario 4: Effect of Basic Bandwidth Unit

In this scenario, the bbu required by the calls for each slice is increased by a unit of 1 in each iteration whilst the parameters in Table 5.4.1 are kept constant. Calls are said to demand resources which are measured by the bbu hence the values of bbu are incremented by 1 bbu each time. Increasing the demanded bbu suggest that the calls will demand more resources from the network hence the call blocking and dropping probabilities should increase.

Slice Parameters			
	eMBB	mMTC	uRLLC
Capacity	30	15	30
New Call Threshold	50%	100%	50%
Handoff Call Threshold	100%	-	100%
Arrival Rate	3	3	3

Table 5.4.1 : Scenario 4: Parameters

5.4.1 Effect of bbu on Intra Slice Handover

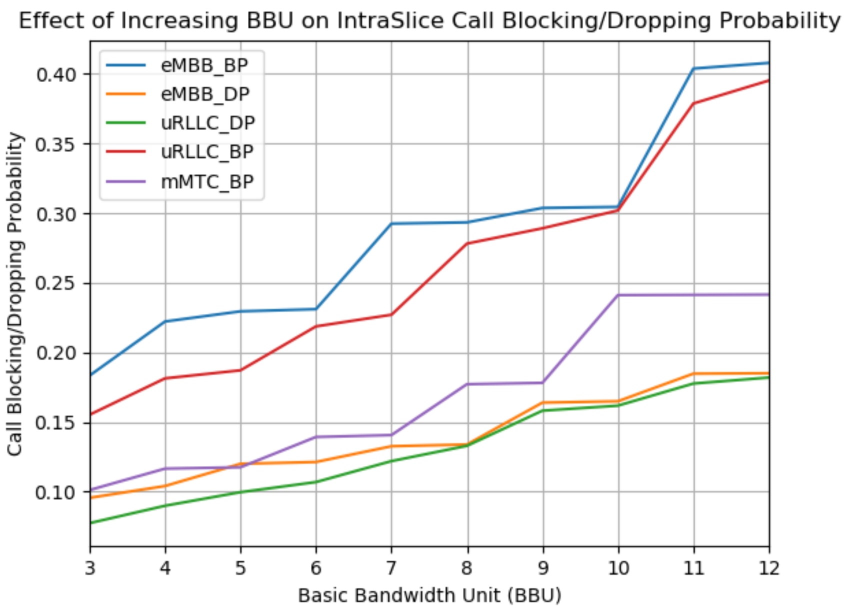

Figure 5.4.1: Effect of Increasing bbu on Intra Slice call blocking/dropping Probabilities

Figure 5.4.1 illustrates the effect of increasing required bbu on new call blocking and handoff call dropping probabilities. From the graph, it is evident that increasing the bbu demanded by calls result in a significant decrease in the QoS of the network as the number of new calls blocked and handoff calls dropped is increased. The blocking/dropping probabilities are increased because the network has finite resources hence as the bandwidth requested by calls increase, the network admit less calls in order to maintain acceptable levels of QoS. eMBB calls has the highest required bbu hence it has the highest new call blocking probability and handoff call dropping probability. mMTC has the least bbu requirements hence it has the least new call dropping probability.

5.4.2 Effect of bbu on Inter Slice Handover

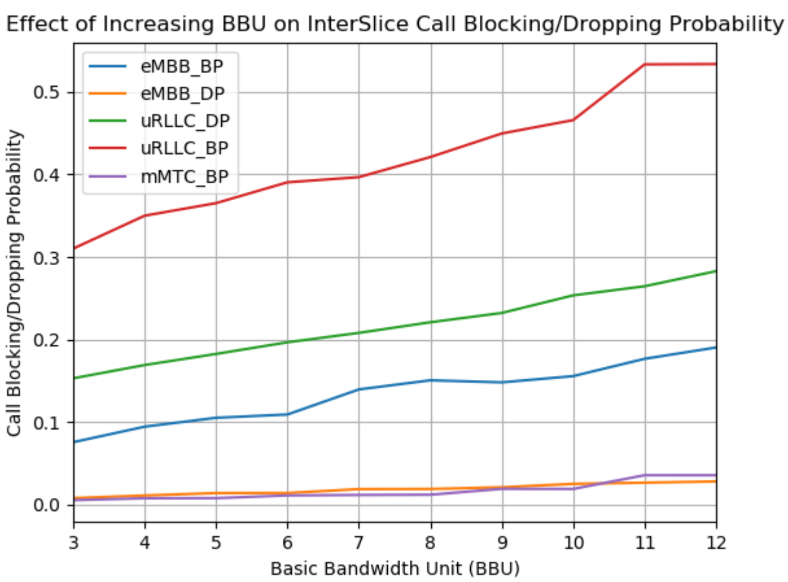

Figure 5.4.2: Effect of Increasing bbu on Inter Slice Call Blocking/Dropping Probabilities

Figure 5.4.2 illustrates the effect of increasing the bbu required by calls on new call blocking and handoff call dropping probability. From the graph it is evident that increasing the required bbu results in an increase in the blocking and dropping probabilities. This is because an increase in the demanded bbu results in the network accepting less calls because the network has limited resources hence can only provide services to a limited number of calls. uRLLC has the highest call blocking and dropping probabilities because it caters for calls from other slices thus it has less bandwidth to cater for its calls. mMTC has the least blocking probability because when the slice is out of bandwidth for its calls, the calls can also be admitted in the eMBB or uRLLC slice hence less mMTC calls are blocked.

5.5 Scenario 5: Effect of Departure Rate

In this scenario, the call departure rates for calls are varied whilst the parameters in
Table 5.5.1 are kept constant. The values of call departure rate are increased by 0.5 from 0.5 to 5 thus there are 10 different departure rates. The departure rate of new calls is assumed to be equal to the departure rate of handoff calls. Increase call departure rates with constant arrival rate suggest that more calls are leaving the network hence there should be a decrease in new call blocking and handoff call dropping probability.

Slice Parameters			
	eMBB	mMTC	uRLLC
Capacity	30	15	30
New Call Threshold	50%	100%	50%
Handoff Call Threshold	100%	-	100%
Arrival Rate	3	3	3

Table 5.5.1: Scenario 5 Parameters

5.5.1 Effect of Call Departure rate on Intra Slice Handover

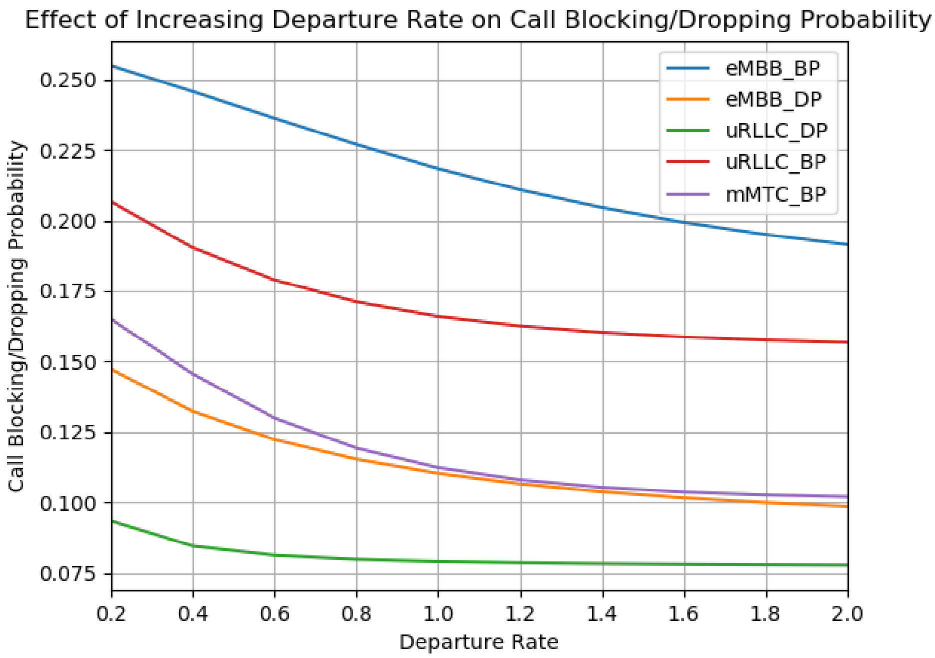

Figure 5.5.1: Effect of Increasing Departure Rate on Intra Slice Call Blocking/Dropping Probability

Figure 5.5.1 illustrates the effects of increasing call departure rate on new call blocking and handoff call dropping probability. The graph shows that the new probabilities decrease as the call departure rate decrease. This is because an increase in the departure rate suggests that calls are leaving the network faster hence a reduction in the call blocking/dropping probabilities. This means there is good connection within the network, however, it is important to note that calls will not arrive at a constant rate. mMTC slice calls has the lowest call blocking probability because the calls required the least bbu.

5.5.2 Effect of Call Departure rate on Inter Slice Handover

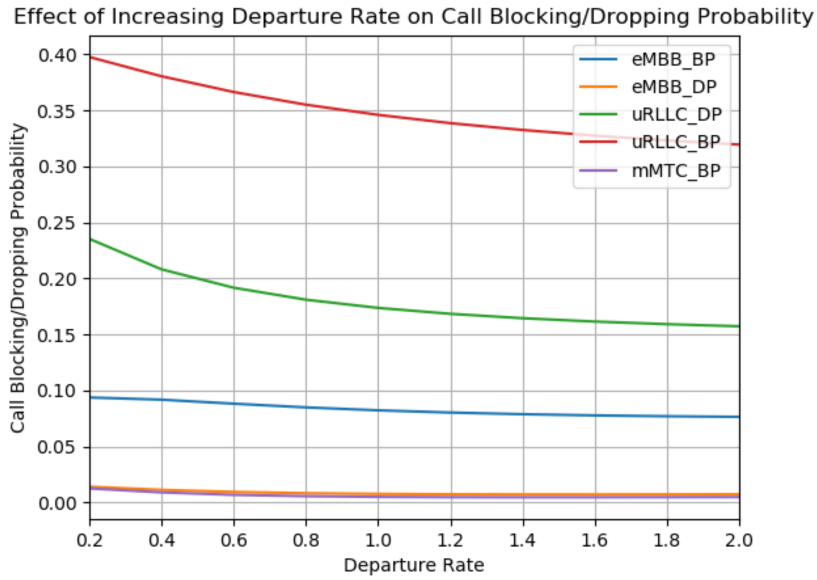

Figure 5.5.2: Effect of Increasing Departure Rate on Inter Slice Call Blocking/Dropping Probability

Figure 5.5.2 illustrates the effects of increasing call departure rate on new call blocking and handoff call dropping probability. From the graphs, it is evident that the call blocking/dropping probabilities decrease with increase in call departure rate. This is because an increase in departure rate means that calls are leaving the network faster hence less calls are blocked or dropped. The blocking/dropping probabilities are greater between 0 and 1 after that, the probabilities fall to a number close to zero in mMTC and eMBB. This is because calls for the two slices can be admitted in other slices with calls for mMTC slice being accommodated in eMBB and uRLLC and calls for eMBB slice being accommodated in uRLLC slice. Furthermore, this suggest good performance in the network. uRLLC slice has the highest call blocking/dropping probabilities because it accommodates calls from other slices hence there is less bandwidth available for uRLLC type calls.

5.6 Scenario 6: Comparative Analysis

In this scenario, the performance of the two algorithms implemented is compared through the call blocking/dropping probabilities for each slice. The results obtained in previous scenarios are consistent thus comparison between the two algorithms is made based on the results obtained in Scenario 3: Effect of Capacity.

5.6.1 Effect of Increasing Capacity on eMBB call Blocking/Dropping Probability

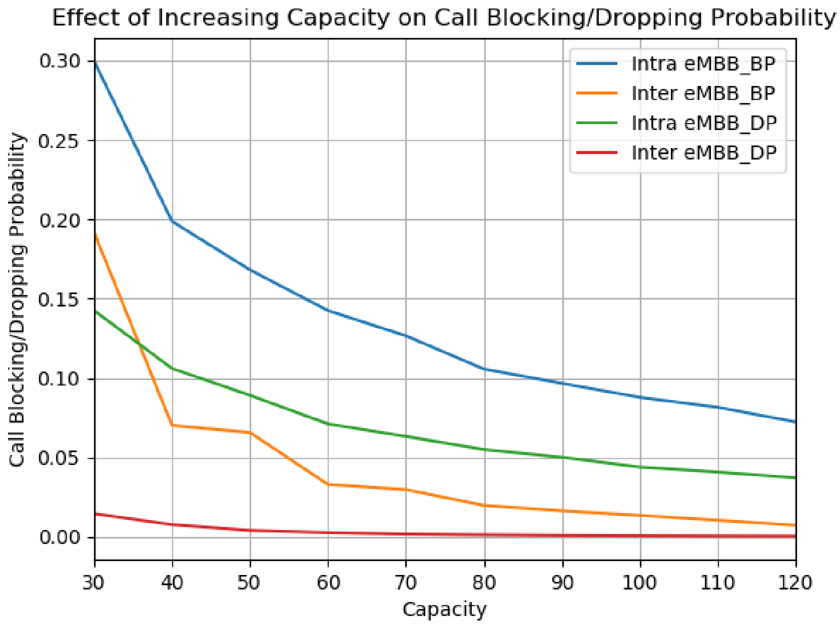

Figure 5.6.1: Effect of Increasing Capacity on (eMBB) Intra and Inter Slice Call Blocking/Dropping Probability

Figure 5.6.1 illustrates the effect of increasing capacity on new call blocking and handoff call dropping probability in intra slice and inter slice handover. From the graph, it is evident that Inter slice provides the best QoS because it maintains the lowest call blocking and dropping probabilities with the inter slice new call blocking probability being less than the intra slice new call blocking probability by at least 36% and the inter slice handoff call blocking probability being less than the intra slice handoff call blocking probability by at least 89%. This is because in inter slice handover, the calls for eMBB slice can also be accommodated in the uRLLC slice thus when a call cannot be admitted in the eMBB slice it is handed over to the uRLLC slice before it is blocked or dropped hence less calls are blocked or dropped in inter slice handover whilst in intra slice handover, the calls cannot be handed over to another slice

5.6.2 Effect of Increasing Capacity on mMTC Call Blocking Probability

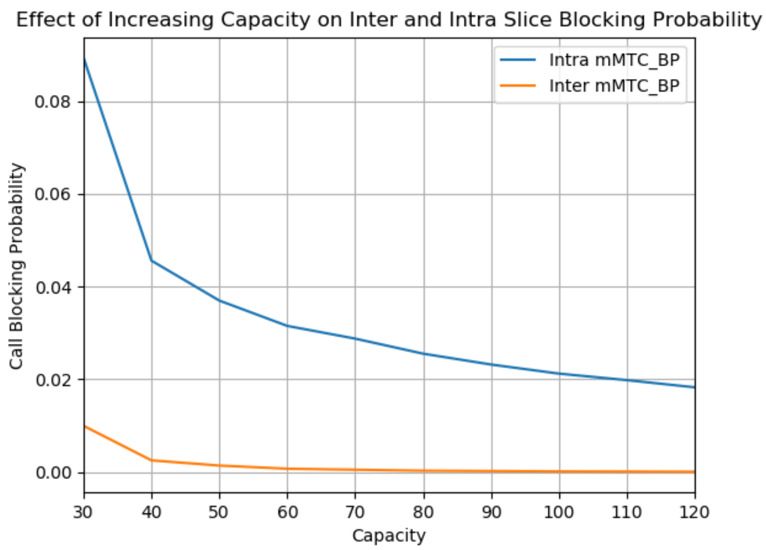

Figure 5.6.2: : Effect of Increasing Capacity on (mMTC) Intra and Inter Slice Call Blocking Probability

Figure 5.6.2 illustrates the effect of increasing Capacity on new call blocking probability in Intra and Inter slice handover. The figure shows that the blocking probability in inter slice handover is consistently lower than that in intra slice by at least 88%. This is because calls for mMTC slice can also be admitted in the eMBB slice and uRLLC slice thus when a call cannot be admitted in the mMTC slice it is handed over to one of the slices with enough space to accommodate it before it can be blocked hence less calls are blocked in inter slice compared to intra slice handover where calls cannot be handed over to other slices.

5.6.3 Effect of Increasing Capacity on uRLLC Call Blocking/Dropping Probability

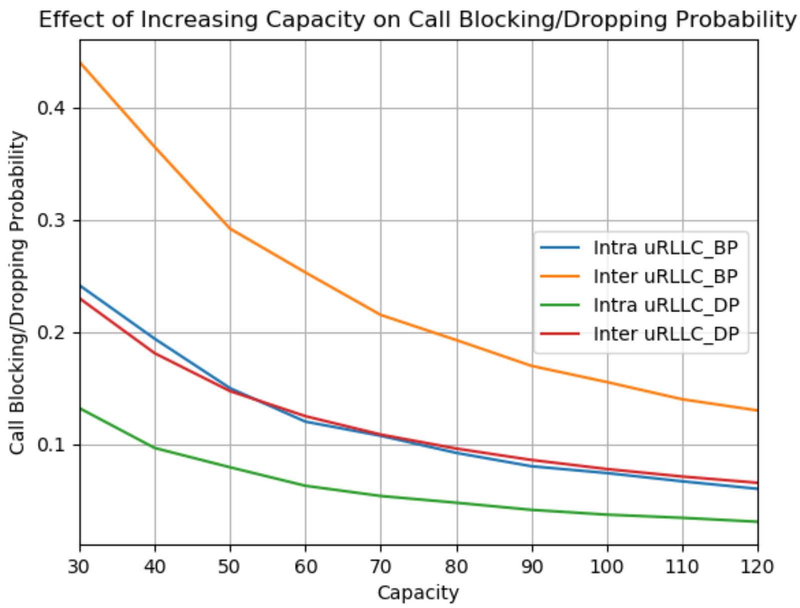

Figure 5.6.3: : Effect of Increasing Capacity on (uRLLC) Intra and Inter Slice Call Blocking/Dropping Probability

Figure 5.6.3 illustrates the effect of increasing Capacity on new call blocking and handoff call dropping probability in intra slice and inter slice handover. Contrary to what was observed in sections 5.6.1 and 5.6.2 the call blocking and dropping probabilities in intra slice handover are less than those inter slice handover by at least 45% in call blocking and 42% in call dropping. This is because the uRLLC slice in inter slice handover accommodates calls from other slices hence the slice does not have enough bandwidth reserved for uRLLC calls thus more calls are blocked and dropped in inter slice handover.

6. Conclusions

The aim of this research project was to design an algorithm for making handover decisions in sliced 5G wireless networks and evaluate its performance using an analytical model. In sliced 5G wireless networks, the physical network is sliced into multiple virtual networks, each designed to cater for specific services.

By considering different slice handover conditions, it was possible to implement two different algorithms. In intra slice handover algorithm, there is no handover between different slices and in Inter slice handover there is handover between slices. The analytical results obtained from the simulations showed that in intra slice handover the call blocking and dropping probabilities increase with an increase in the call arrival rate. The rate of increase of the probability with call arrival rate means that the network would not be able to provide good QoS to network subscribers. Consequently, the inter slice handover algorithm which enables calls to be handed over to different slices was simulated. In Inter slice handover, the call blocking and call dropping probabilities were reduced significantly under the same call arrival rates showing that a network with inter slice handover can provide better QoS to 5G network subscribers. However, inter slice handover means calls for the uRLLC slice experiences higher call blocking and call dropping probabilities because the slice caters for calls from other slices thus less space is available for calls for the uRLLC slice.

The effect of increasing slice capacity was also used to evaluate the network performance and the results showed a reduction in the call blocking and dropping probability in all the slices in both intra slice and inter slice handover. This conveys that increasing network resources enables the network to cater for more calls, however network resources remain challenge hence there is a limit to how much the capacity can be increased by.

Furthermore, the effect of call departure rate was simulated, and the results obtained showed an improvement in QoS by decreasing the call blocking and call dropping probabilities in all the slices. The challenge remains predicting how long users will take on a call however, some percentage of users are expected to make short calls thereby improving QoS.

In overall the implemented algorithms show that the handoff call dropping probability will always be less than new call blocking probability due to the integration of the guard channel bandwidth reservation policy in the network. In addition, inter slice handover provides better QoS than intra slice handover algorithm in eMBB and mMTC slices whilst the Intra slice handover algorithm provides better QoS in the uRLLC slice.

It can thus be concluded that the implemented handover algorithms are sufficient to make handover decisions in sliced 5G networks whilst ensuring continuous connection.

7. Recommendations

The following recommendations are made for future work:

- This thesis assumes that the network has three slices, future work should investigate a network with more slices in order to mimic real life networks.

- This research assumes that in inter slice handover, both new calls and handoff calls can be handover to other slices. Future work should consider making handover to other slices available for handover calls only.

- Future work should consider the potential effects of dynamic resource allocation in slices, for example in inter slice handover in order to improve the QoS of the uRLLC slice more resources can be allocated to the slice when it is being overwhelmed by calls.

- Further investigations should be done on the mobility of users. The impact of user mobility on the network performance should be investigated.

- Future work should consider implementing the inter slice handover algorithm with multi-thresholds. With one threshold catering for calls from other slices and the other threshold catering for uRLLC slice type calls only.

- In order to get better insight on the performance of the network, more performance metrics can be used. Performance metrics like handoff delay, signalling overhead and call blocking ratio can be used to evaluate the efficiency of the algorithms.

8. References

[1] A. U. Gawas, “An Overview on Evolution of Mobile Wireless Communication Networks: 1G-6G,” *International Journal on Recent and Innovation Trends in Computing and Communication,* vol. 3, no. 5, pp. 3130-3133, 2015.

[2] A. A.-M. Bulbul, S. Biswas, M. B. Hossain and S. Biswas, “Past, Present and Future of Mobile Wireless Communication,” *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE),* vol. 12, no. 5, pp. 55-88, 2017.

[3] N. Bhandari, S. Devra and K. Singh, “Evolution of Cellular Network: From 1G to 5G,” *Internation Journal of Engineering and Techniques,* vol. 3, no. 5, pp. 98-105, 2017.

[4] S. S. Kolahi, “Comparison of FDMA. TDMA and CDMA system capacities,” *10th WSEAS International Conference on COMMUNICATIONS,* pp. 325-328, 2006.

[5] P. Manhas, S. Thakral and C. Kumar, “Performance Analysis of GPRS/EDGE,” *International Journal of Electronic Engineering Research,* vol. 2, no. 5, pp. 655-658, 2010.

[6] I. u. Haq, Z. U. Rahman, S. Ali and E. M. Faisal, “GSM Technology: Architecture, Security and Future Challenges,” *International Journal of Science Engineering and Advance Technology,* vol. 5, no. 1, pp. 138-139, 2017.

[7] M. M. u. i. Mir and D. S. Kumar, “Evolution of Mobile Wireless Technology from 0G to 5G.,” *(IJCSIT) International Journal of Computer Science and Information Technologies,* vol. 6, no. 3, pp. 2545-2551, 2015.

[8] A. Kumar1, D. Y. Liu2, D. J. Sengupta3 and Divya4, “Evolution of Mobile Wireless Communication Networks: 1G to 4G,” *The International Journal of Electronics & Communication Technology (IJECT),* vol. 1, no. 1, 2010.

[9] V. Gazis, N. Housos, A. Alonistioti and L. Merakos, “GENERIC SYSTEM ARCHITECTURE FOR 4G MOBILE COMMUNICATIONS,” *Vehicular Technology Conference, 1988, IEEE 38th,* vol. 3, no. 3, pp. 1512 - 1516, 2003.

[10] S. O. Oladejo and O. E. Falowo, “5G Network Slicing: A Multi-Tenancy Scenario,” *Global Wireless Summit (GWS),* pp. 88-92, 2017.

[11] H. Zhang, N. Lui, X. Chu and K. Long, “Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges,” *IEEE Communications Magazine,* vol. 55, no. 8, pp. 138 - 145, 2017.

[12] A. GUPTA and R. K. JHA, “A Survey of 5G Network: Architecture and Emerging Technologies,” *IEEE Access,* vol. 3, pp. 1206-1232, 2015.

[13] P. Paudel and A. Bhattarai, “5G Telecommunication Technology: History,Overview, Requirements and Use Case Scenarioin Context of Nepal,” in *ITD4*, Nepal, 2018.

[14] GSMA, *Energy Efficiency: An Overview,* GSM Association, 2019.

[15] S. Xiao and W. Chen, “Dynamic Allocation of 5G Transport Network Slice Bandwidth Based on LSTM Traffic Prediction,” *IEEE 9th International Conference on Software Engineering and Service Science (ICSESS),* pp. 735-739, 2018.

[16] F. Al-Ogaili and R. M. Shubair, “Millimeter-Wave Mobile Communications for 5G: Challenges and Opportunities,” in *2016 IEEE International Symposium on Antennas and Propagation (APSURSI)*, Fajardo, Puerto Rico, 2016.

[17] A. Nordrum and K. Clark, “IEEE Spectrum,” 27 January 2017. [Online]. Available: https://spectrum.ieee.org/video/telecom/wireless/everything-you-need-to-know-about-5g. [Accessed 17 April 2020].

[18] P. Paudel, “5G Telecommunication Technology: History, Overview, Requirements and Use Case Scenario in Context of Nepal,” Kathmandu, Nepal, May 2018.

[19] J. Hoadley and P. Maveddat, “Enabling small cell deployment with HetNet,” *IEEE Wireless Communications,* vol. 19, no. 2, pp. 4-5, 2012.

[20] U. kar and D. K. Sanyal, “An overview of device-to-device communication in cellular networks,” *Korean Intiitute of Communications and Information Sciences,* vol. 4, no. 4, pp. 203-208, 2017.

[21] A.-F. Naser and A. Omar, “Technologies for 5G Networks: Challenges and Opportunities,” *IT Profesional,* vol. 19, no. 1, pp. 12-20, 2017.

[22] B. Sanou, “Setting the Scene for 5G: Opportunities & Challenges,” in *Telecommunication Development Bureau Place des Nations*, Geneva, 2018.

[23] M. Jiang, M. Condoluci and T. Mahmoodi, “Network slicing management & prioritization in 5G Mobile Systems,” *European Wireless,* pp. 1-6, 2016.

[24] Infradata, “Infradata,” Infradata, 2020. [Online]. Available: https://www.infradata.com/resources/what-is-network-slicing/. [Accessed 12 May 2020].

[25] K. Liolis, S. Watts and P. S. Khodashenas, “Use cases and scenarios of 5G integrated satellite-terrestrial networks for enhanced mobile broadband: The SaT5G approach,” *International Journal of Satellite Communications and Networking ·,* vol. 37, pp. 91-112, 2019.

[26] A. A. Barakabitze, A. Ahmad, A. Hines and R. Mijumbi, “5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges,” *Computer Networks,* p. 106984, 2019.

[27] M. Smith, M. Dvorkin, Y. Laribi, V. Pandey, P. Garg and N. Weidenbache, “OpFlex Control Protocol,” Internet Engineering Task Forc, 2014.

[28] A. Lara, A. Kolasani and B. Ramamurthy, “Network Innovation using OpenFlow: A Survey,” *IEEE Communications Surveys & Tutorials,* vol. 16, no. 1, pp. 493-512, 2014.

[29] G. A. Carella, M. Pauls, T. Magedanz, M. Cilloni, P. Bellavista and L. Foschini, ““Prototyping NFV-based multiaccess edge computing in 5G ready networks with open baton,” *IEEE Conference on Network Softwarization (NetSoft),* pp. 1-4, 2017.

[30] J. S. J.-L. G. N. B. F. T. S. M. R. Mijumbi and R. Boutaba, “Network Function Virtualisation: State of the Art and research challenges,” *IEEE Communications Surveys & Tutorials,* vol. 18, no. 1, pp. 239-279, 2016.

[31] C.-X. Wang, F. Haider, X. Gao, X.-H. You, Y. Yang, D. Yuan, H. M. Aggoune, H. Haas, S. Fletcher and E. Hepsaydir, “Cellular architecture and key technologies for 5G wireless communication networks,” *5G WIRELESS COMMUNICATION SYSTEMS:PROSPECTS AND CHALLENGES,* vol. 52, no. 2, pp. 122-130, 2014.

[32] Open Air Interface, “Open Air Interface,” 2019. [Online]. Available: https://www.openairinterface.org/?page_id=458. [Accessed 18 May 2020].

[33] S. Tarapiah, K. Aziz and S. Atalla, “Common Radio Resource Management Algorithms in Heterogeneous Wireless Networks with KPI Analysis,” *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 6, no. 10, pp. 53-58, 2015.

[34] M. H. Ahmed, “Call admission control in wireless networks: A comprehensive survey,” *IEEE Communications Surveys & Tutorials,* vol. 7, no. 1, pp. 49-68, 2005.

[35] E. Z. Tragos, G. Tsiropoulos, G. T. Karetsos and S. A. Kyriazakos, “Admission control for QoS Support in Heterogeneous 4G Wireless Networks,” *IEEE Network,* vol. 22, no. 3, pp. 30-37, 2008.

[36] O. E. Falowo, “Efficient Joint Call Admission Control and Bandwidth Management Schemes for QoS Provisioning in Heterogeneous Wireless Networks,” Cape Town, 2008.

[37] O. E. Falowo and A. H. Chan, “Joint Call Admission Control Algorithms: Requirements, Approaches, and Design Considerations,” *Computer Communications,* vol. 31, no. 6, pp. 1200-1217, 2008.

[38] R. B. H.S., G. Shankar and S. P.S., “Call Admission Control in Beyond 3G Networks Using Multi Criteria Decision Making,” *2009 First International Conference on Computational Intelligence, Communication Systems and Networks,* pp. 492-496, 2009.

[39] D. D. Nguyen, H. X. Nguyen and L. B. White, “Evaluating Performance of RAT Selection Algorithms for 5G HetNets,” *IEEE Access,* vol. 6, no. 9, pp. 61212-61222, 2018.

[40] W. Li, S. Wang, Y. Cui, X. Cheng, R. Xin, M. A. Al-Rodhaan and A. Al-Dhelaan, “AP Association for Proportional Fairness in Multirate WLANs,” *IEEE/ACM Trans. on Networking,,* vol. 22, no. 1, pp. 192-201, 2014.

[41] Z. Han, D. Niyato, W. Saad and T. Başar, “Game Theoryin Wireless and Communication Networks: Theory, Models, and Applica-tions.,” Cambridge University Press, Vancouver, 2011.

[42] D. D. Nguyen, H. X. Nguyen and L. B. White, “Reinforcement LearningWith Network-Assisted Feedback for Heterogeneous RAT Selection,,” *IEEE Transactions on Wireless Communications,* vol. 6, no. 9, pp. 6062 - 6076, 27 June 2017.

[43] A. Sgora and D. D. Vergados, “Handoff Prioritization and Decision Schemes in Wireless Cellular Networks: a Survey,” *IEEE COMMUNICATIONS SURVEYS & TUTORIALS,* vol. 11, no. 4, pp. 57-77, 2009.

[44] S. Y. Diaba and E. Affume, “Performance Analysis of Queuing Priority Schemes in cellular communication,” *International Journal of Advanced Research in Computer and Communication Engineering(IJARCCE),* vol. 4, no. 1, pp. 232-236, January 2015.

[45] J. khan, “Handover management in GSM cellular system,” *International Journal of Computer Applications,* vol. 2, no. 2, 2014.

[46] H. Bhute, P. P. Karde and V. M. Thakare, “Vertical Handover Decision Strategies in Heterogeneous Wireless Networks,” in *Recent Trends in Information, Telecommunication and Computing, ITC* , Bengaluru, 2014.

[47] X. Yan, Y. A. Sekerciogluy and S. Narayananz, “A Survey of Vertical Handover Decision Algorithms in Fourth Generation Heterogeneous Wireless Networks,” *Computer Networks,* vol. 54, no. 11, pp. 1848-1863, 2010.

[48] M. Axente-Stan and E. Borcoc, “Performance evaluation of handover policies in mobile heterogenous networks,” *MESH 2011 - 4th International Conference on Advances in Mesh Networjs,* pp. 20-25, 2011.

[49] A. Bijwe and C. G. Dethe, “RSS based Vertical Handoff algorithms for Heterogeneous wireless networks - A Review,” *(IJACSA) International Journal of Advanced Computer Science and Applications,* vol. 1, no. 2, 2011.

[50] M. Kassar, B. Kervella and G. Pujolle, “Vertical Handover Decision Strategies in Heterogeneous Wireless Networks,” *Recent Trends in Information, Telecommunication and Computing, ITC,* vol. 31, no. 10, pp. 2607-2620, 2008.

[51] S. Horsmanheimo, N. Maskey, H. Kokkoniemi-Tarkkanen, L. Tuomim¨aki and P. Savolainen, "Learning Based Proactive Handovers in Heterogeneous Networks," *Vuorimiehentie 3, Espoo, 02044 VTT,* vol. 125, pp. 57-68, 2018.

[52] V. Yajnanarayana, H. Ryd´en, L. ´. H´evizi, A. Jauhari and M. Cirkic, "5G Handover using Reinforcement Learning," *IEEE 5G WF 2020,* vol. 1, no. 3, 2019.

[53] H. Chen, C.-C. Cheng and H.-H. Yeh, ""Guard-Channel-Based Incremental and Dynamic Optimisation on Call Admission Control for Next-Generation QoS-Aware Heterogeneous Systems," *EEE Transactions on Vehicular Technology,* vol. 57, no. 5, pp. 3064-3081, February 2008.

[54] D. A. S. Alfoudi, D. Mohammed, O. Abayomi, P. Rubem and L. G. Myoung, "Mobility Management Architecture in Different RATs Based Network Slicing," *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA),* vol. 19, no. 2, pp. 270-274, 2018.

[55] O. E. Falowo and H. A. Chan, ""Adaptive Bandwidth Management and Joint Call Admission Control to Enhance System Utilisation and QoS in Heterogeneous Wireless Networks," *EURASIP Journal on Wireless Communications and Networking,* vol. 1, 2007.

[56] L. Haeyoung, S. Vahid and K. Moessner, "Machine Learning based RATs Selection supporting MultiConnectivity for Reliability," *Institute for Communication Systems (ICS),* pp. 31-41, 2019.

[57] O. E. Falowo and H. A. Chan, "Join call admission control algorithm to enhance connection-level QoS in heterogeneous cellular networks," *AFRICON 2007,* pp. 1-7, 2007.

[58] G. Bolch, S. Greiner, H. d. Meer and K. S. Trivedi, Queueing Networks and Markov Chains: Modeling and Performance Evaluation With Computer Science Applications, Second Edition, vol. 95, John Wiley, 2006, pp. 869-878.

[59] H. Zeng and I. Chlamtac, "Adaptive Guard Channel Allocation and Blocking Probability Estimation in PCS Networks," *Computer Networks,* vol. 43, no. 2, pp. 163-176, 2003.

[60] G. Bolch, S. Greiner, H. d. Meer and K. S. Trivedi, Queueing Networks and Markov Chains: Modeling and Performance Evaluation With Computer Science Applications, Second Edition, John Wiley & Sons, 2006, pp. 869-878.

[61] D. Phillips, Python 3 Object-oriented Programming, Birmingham, 26 July 2010.

[62] Q. T. A. Pham, K. Piamrat and C. Viho, "Resource Management in Wireless Access Networks: Alayer-based classification - Version 1.0," *Internes de l'IRISA,* 2014.

9. Appendices

9.1 Appendix A: Simulation Program

All the program used in the simulations can be found in the Algorithm for slice handover GitHub repository.

9.2 Appendix B: Extracted Simulation Results

Effect of Arrival Rate on Intra Slice Handover

Arrival Rate	eMBB_BP	uRLLC_BP	mMTC_BP	eMBB_DP	uRLLC_DP
3	0,313213	0,246894	0,122357	0,148738	0,133863
3,5	0,31763	0,250619	0,141427	0,152263	0,135504
4	0,320838	0,254642	0,161243	0,155925	0,137533
4,5	0,323189	0,258788	0,181187	0,159641	0,139932
5	0,324977	0,26295	0,200764	0,163342	0,142693
5,5	0,326419	0,267074	0,219592	0,16697	0,14582
6	0,327664	0,27115	0,237397	0,170484	0,149319
6,5	0,328803	0,275195	0,253986	0,173847	0,153202
7	0,32989	0,27924	0,269234	0,177033	0,157478
7,5	0,330949	0,283328	0,28307	0,18002	0,162155

Effect of Arrival Rate on Inter Slice Handover

Arrival Rate	eMBB_BP	uRLLC_BP	mMTC_BP	eMBB_DP	uRLLC_DP
3	0,218179	0,466163	0,015092	0,015468	0,242471
3,5	0,232434	0,478292	0,01918	0,016401	0,249443
4	0,246915	0,489947	0,024196	0,017556	0,256752
4,5	0,261347	0,501142	0,030036	0,018869	0,264324
5	0,275492	0,511888	0,036573	0,020283	0,272088
5,5	0,289151	0,522197	0,043669	0,021754	0,279982
6	0,302162	0,532077	0,051185	0,023244	0,287956
6,5	0,314393	0,541542	0,058984	0,024723	0,295969
7	0,325747	0,550601	0,066938	0,026166	0,303992
7,5	0,336153	0,55927	0,07493	0,027554	0,312003

Effect of Capacity on Intra Slice Handover

Capacity	eMBB_BP	uRLLC_BP	mMTC_BP	eMBB_DP	uRLLC_DP
30	0,29957	0,241377	0,08937	0,142501	0,131907
40	0,198648	0,193692	0,045604	0,106013	0,096385
50	0,16811	0,149659	0,036987	0,089116	0,079212
60	0,14249	0,119874	0,031531	0,07109	0,06273
70	0,126573	0,107263	0,028788	0,063247	0,053631
80	0,105686	0,092075	0,025544	0,054947	0,047666
90	0,096672	0,080132	0,023204	0,050091	0,041295
100	0,087832	0,07407	0,021244	0,043916	0,037035
110	0,081549	0,066712	0,019835	0,040775	0,034203
120	0,072344	0,060037	0,018288	0,037148	0,030703

Effect of Capacity on Inter Slice Handover

Capacity	eMBB_BP	uRLLC_BP	mMTC_BP	eMBB_DP	uRLLC_DP
30	0,191659	0,440393	0,009983	0,014579	0,229837
40	0,070166	0,364521	0,002532	0,007607	0,180792
50	0,065678	0,291776	0,001413	0,00399	0,14708
60	0,032991	0,252787	0,000732	0,002576	0,124747
70	0,029709	0,214915	0,00051	0,00164	0,108348
80	0,019765	0,192601	0,000303	0,001188	0,095893
90	0,016341	0,169566	0,000232	0,00084	0,085733
100	0,013478	0,155144	0,000149	0,000652	0,077658
110	0,010395	0,139813	0,000121	0,000493	0,070977
120	0,007248	0,12986	8,19E-05	0,0004	0,065389

Effect of Threshold on Intra Slice Handover

Threshold	**eMBB_BP**	**uRLLC_BP**	**mMTC_BP**	**eMBB_DP**	**uRLLC_DP**
6	0,370378	0,352901	0,018221	0,096591	0,075274
9	0,285569	0,270492	0,018221	0,100999	0,077825
12	0,235238	0,193437	0,018221	0,105136	0,083423
15	0,204037	0,172462	0,018221	0,109693	0,086151
18	0,183971	0,147412	0,018221	0,115204	0,093177
21	0,170623	0,138981	0,018221	0,122433	0,096969
24	0,161688	0,128654	0,018221	0,13175	0,108113
27	0,156099	0,125425	0,018221	0,142503	0,115328
30	0,153217	0,123593	0,018221	0,153217	0,123593

Effect of Threshold on Inter Slice Handover

Threshold	eMBB_BP	uRLLC_BP	mMTC_BP	eMBB_DP	uRLLC_DP
6	0,506044	0,614671	0,001003	0,010189	0,130828
9	0,251837	0,468784	0,001003	0,010392	0,141302
12	0,133604	0,377473	0,001003	0,010451	0,152373
15	0,078765	0,320632	0,001003	0,011088	0,164859
18	0,050544	0,284006	0,001003	0,011222	0,178797
21	0,033952	0,26059	0,001003	0,012273	0,194491
24	0,024147	0,246149	0,001003	0,012274	0,210931
27	0,018109	0,238645	0,001003	0,013892	0,226281
30	0,015707	0,235897	0,001003	0,015707	0,235897

Effect of Departure Rate on Intra Slice Handover

eparture Ra	eMBB_BP	uRLLC_BP	mMTC_BP	eMBB_DP	uRLLC_DP
0,2	0,254939	0,206683	0,165044	0,147259	0,093439
0,4	0,24575	0,190332	0,145533	0,132259	0,084493
0,6	0,236268	0,178871	0,129893	0,122248	0,081205
0,8	0,226942	0,171113	0,119196	0,115287	0,079743
1	0,21837	0,165958	0,112302	0,110259	0,078981
1,2	0,210881	0,162532	0,107968	0,106529	0,078528
1,4	0,204549	0,160227	0,105269	0,103708	0,078227
1,6	0,199295	0,158648	0,103596	0,101541	0,07801
1,8	0,194978	0,157541	0,102564	0,099852	0,077844
2	0,191445	0,156747	0,101931	0,098519	0,07771

Effect of Departure Rate on Inter Slice Handover

eparture Ra	eMBB_BP	uRLLC_BP	mMTC_BP	eMBB_DP	uRLLC_DP
0,2	0,093761	0,397743	0,012586	0,014064	0,235408
0,4	0,091792	0,380537	0,009014	0,011296	0,208175
0,6	0,088213	0,366467	0,00682	0,009442	0,191741
0,8	0,084919	0,355169	0,005621	0,008316	0,181058
1	0,082292	0,346076	0,00501	0,007677	0,173716
1,2	0,080303	0,338681	0,004732	0,007349	0,168448
1,4	0,078832	0,33259	0,004643	0,007215	0,164534
1,6	0,07776	0,327508	0,004664	0,007203	0,161541
1,8	0,07699	0,32322	0,004747	0,007265	0,159198
2	0,076447	0,319562	0,004868	0,007373	0,157324

Effect of Basic Bandwidth Unit on Intra Slice handover

bbu	eMBB_BP	uRLLC_BP	mMTC_BP	eMBB_DP	uRLLC_DP
3	0,183363	0,155239	0,10117	0,095577	0,077392
4	0,222213	0,18138	0,116552	0,104061	0,089892
5	0,229513	0,187159	0,117529	0,120027	0,09965
6	0,231074	0,218704	0,139292	0,121317	0,106897
7	0,292502	0,227066	0,140647	0,132627	0,121967
8	0,293406	0,278151	0,177199	0,13389	0,132928
9	0,303749	0,289152	0,178139	0,163957	0,158179
10	0,304532	0,301946	0,24119	0,164926	0,161781
11	0,40385	0,378777	0,241363	0,184768	0,177657
12	0,407931	0,395267	0,24154	0,185081	0,181961

Effect of Basic Bandwidth Unit on Inter Slice Handover

bbu	eMBB_BP	uRLLC_BP	mMTC_BP	eMBB_DP	uRLLC_DP
3	0,07564	0,310197	0,005404	0,007893	0,153025
4	0,094426	0,350085	0,007675	0,011074	0,169116
5	0,105204	0,365298	0,00785	0,014015	0,182582
6	0,109213	0,390577	0,011125	0,014078	0,19669
7	0,139597	0,396777	0,011716	0,018806	0,208092
8	0,150608	0,421146	0,011995	0,018962	0,220961
9	0,148152	0,449684	0,019132	0,020987	0,23223
10	0,155684	0,465898	0,018991	0,025168	0,253569
11	0,176681	0,533447	0,035634	0,026578	0,264574
12	0,190378	0,533831	0,035634	0,028109	0,282871

10. EBE Faculty: Assessment of Ethics in Research Projects

Any person planning to undertake research in the Faculty of Engineering and the Built Environment at the University of Cape Town is required to complete this form before collecting or analysing data. When completed it should be submitted to the supervisor (where applicable) and from there to the Head of Department. If any of the questions below have been answered YES, and the applicant is NOT a fourth year student, the Head should forward this form for approval by the Faculty EIR committee: submit to Ms Zulpha Geyer (Zulpha.Geyer@uct.ac.za; Chem Eng Building, Ph 021 650 4791).Students must include a copy of the completed form with the final year project when it is submitted for examination.

Name of Principal Researcher/Student: Tawanda Muzanenhamo **Department:** ELECTRICAL ENGINEERING

If a Student: YES **Degree:** Electrical and Computer Engineering **Supervisor:** Prof O.E Falowo

If a Research Contract indicate source of funding/sponsorship: No

Research Project Title: Design of an Algorithm for Slice Handover in 5G Networks

Overview of ethics issues in your research project:

Question 1: Is there a possibility that your research could cause harm to a third party (i.e. a person not involved in your project)?		NO
Question 2: Is your research making use of human subjects as sources of data? If your answer is YES, please complete Addendum 2.		NO
Question 3: Does your research involve the participation of or provision of services to communities? If your answer is YES, please complete Addendum 3.		NO
Question 4: If your research is sponsored, is there any potential for conflicts of interest? If your answer is YES, please complete Addendum 4.		NO

If you have answered YES to any of the above questions, please append a copy of your research proposal, as well as any interview schedules or questionnaires (Addendum 1) and please complete further addenda as appropriate.

I hereby undertake to carry out my research in such a way that

- there is no apparent legal objection to the nature or the method of research; and
- the research will not compromise staff or students or the other responsibilities of the University;
- the stated objective will be achieved, and the findings will have a high degree of validity;
- limitations and alternative interpretations will be considered;
- the findings could be subject to peer review and publicly available; and
- I will comply with the conventions of copyright and avoid any practice that would constitute plagiarism.

Signed by:

	Full name and signature	Date
Principal Researcher/Student:	**Tawanda Muzanenhamo**	05 July 2020

This application is approved by:

Supervisor (if applicable):	**O Falowo**	05 July 2020
HOD (or delegated nominee): Final authority for all assessments with NO to all questions and for all undergraduate research.	**Janine Buxey**	05 July 2020
Chair: Faculty EIR Committee For applicants other than undergraduate students who have answered YES to any of the above		

ADDENDUM 2: To be completed if you answered YES to Question 2:

It is assumed that you have read the UCT Code for Research involving Human Subjects (available at http://web.uct.ac.za/depts/educate/download/uctcodeforresearchinvolvinghumansubjects.pdf) in order to be able to answer the questions in this addendum.

2.1 Does the research discriminate against participation by individuals, or differentiate between participants, on the grounds of gender, race or ethnic group, age range, religion, income, handicap, illness or any similar classification?		NO
2.2 Does the research require the participation of socially or physically vulnerable people (children, aged, disabled, etc) or legally restricted groups?		NO
2.3 Will you not be able to secure the informed consent of all participants in the research? (In the case of children, will you not be able to obtain the consent of their guardians or parents?)		NO
2.4 Will any confidential data be collected, or will identifiable records of individuals be kept?		NO
2.5 In reporting on this research is there any possibility that you will not be able to keep the identities of the individuals involved anonymous?		NO
2.6 Are there any foreseeable risks of physical, psychological or social harm to participants that might occur in the course of the research?		NO
2.7 Does the research include making payments or giving gifts to any participants?		NO

If you have answered YES to any of these questions, please describe below how you plan to address these issues:

ADDENDUM 3: To be completed if you answered YES to Question 3:

3.1 Is the community expected to make decisions for, during or based on the research?		NO
3.2 At the end of the research will any economic or social process be terminated or left unsupported, or equipment or facilities used in the research be recovered from the participants or community?		NO
3.3 Will any service be provided at a level below the generally accepted standards?		NO

If you have answered YES to any of these questions, please describe below how you plan to address these issues:

ADDENDUM 4: To be completed if you answered YES to Question 4

4.1 Is there any existing or potential conflict of interest between a research sponsor, academic supervisor, other researchers or participants?		NO
4.2 Will information that reveals the identity of participants be supplied to a research sponsor, other than with the permission of the individuals?		NO
4.3 Does the proposed research potentially conflict with the research of any other individual or group within the University?		NO

If you have answered YES to any of these questions, please describe below how you plan to address these issues:

Application for Approval of Ethics in Research (EIR) Projects
Faculty of Engineering and the Built Environment, University of Cape Town

ETHICS APPLICATION FORM

Please Note:
Any person planning to undertake research in the Faculty of Engineering and the Built Environment (EBE) at the University of Cape Town is required to complete this form **before** collecting or analysing data. The objective of submitting this application *prior* to embarking on research is to ensure that the highest ethical standards in research, conducted under the auspices of the EBE Faculty, are met. Please ensure that you have read, and understood the **EBE Ethics in Research Handbook** (available from the UCT EBE, Research Ethics website) prior to completing this application form: http://www.ebe.uct.ac.za/ebe/research/ethics1

APPLICANT'S DETAILS		
Name of principal researcher, student or external applicant		TAWANDA MUZANENHAMO
Department		EBE
Preferred email address of applicant		MZNTAW001@myuct.ac.za
If Student	Your Degree: e.g., MSc, PhD, etc.	BSc
	Credit Value of Research: e.g., 60/120/180/360 etc.	40
	Name of Supervisor (if supervised):	Olabisi Falowo
If this is a researchcontract, indicate the source of funding/sponsorship		
Project Title		Design of an Algorithm for Slice hardware in 5G Networks

I hereby undertake to carry out my research in such a way that:

- there is no apparent legal objection to the nature or the method of research; and
- the research will not compromise staff or students or the other responsibilities of the University;
- the stated objective will be achieved, and the findings will have a high degree of validity;
- limitations and alternative interpretations will be considered;
- the findings could be subject to peer review and publicly available; and
- I will comply with the conventions of copyright and avoid any practice that would constitute plagiarism.

APPLICATION BY	Full name	Signature	Date
Principal Researcher/ Student/External applicant	TAWANDA MUZANENHAMO	[signature]	24/02/2020
SUPPORTED BY	Full name	Signature	Date
Supervisor (where applicable)	Olabisi Falowo	[signature]	24/02/2020

APPROVED BY	Full name	Signature	Date
HOD (or delegated nominee) Final authority for all applicants who have answered NO to all questions in Section 1; and for all Undergraduate research (Including Honours).			
Chair: Faculty EIR Committee For applicants other than undergraduate students who have answered YES to any of the questions in Section 1.			