NETWORK SLICING WITH PRIORITIZATION MECHANISM USING MARKOV CHAIN MODEL UNDER USER DEMAND UNCERTAINTY

BY

Monogren Reddy

Submitted in accordance with the requirements for the degree of

Master of Engineering

In the subject

Electrical Engineering

At the

UNIVERSITY OF SOUTH AFRICA, PRETORIA

2023

Supervisor: **Prof M Sumbwanyambe**

Declaration

I, the undersigned Monogren Reddy, hereby declare that I am the sole author of this dissertation. To the best of my knowledge this dissertation contains no material previously published by any other person except where due acknowledgements and referencing has been made. This dissertation contains no material which has been accepted as part of the requirements of any other academic degree or non-degree program, in English or in any other language. This is a true copy of the dissertation, including final revisions.

Date: 2023/07/02

Name: Monogren Reddy

Signature: M Reddy

Acknowledgements

It is virtually impossible to study any magnitude without the cooperation and assistance of people who are generous not only with their time, but who make the conscious decision of laying on the line deeply personal vestiges of themselves.

I am also sincerely grateful to the generous financial support that I received from Openserve & Telkom SA your support made it possible for me study and conduct my research.

The various processes of the research have enabled me to share and present my work at the icABCD 2022 Conference Programme.

I would like to place on record my heartfelt thanks to my supervisor, Prof M. Sumbwanyambe for reading with refreshing insight, and fine-grained the various drafts of the dissertation. Once in a lifetime, some of us have the rare privilege of being touched by a true supporter and mentor.

Special thanks Dr Ayodele Periola, for his continuous support and assistance during my research.

Finally, my endearing and enduring gratitude to my family who have remained the bedrock of my scholarly and personal adventures.

Table of Contents

Acronyms

Acronym	Definition
5G	5th Generation Mobile Network
BBU	Base Unit
bps	Bits per second
CBP	Call blocking probability
CDP	Call dropping probability
CN	Cognitive Network
CTMC	Continuous Time Markov Chain
DSL	Digital subscriber Loop
eMBB	Enhanced Mobile Broadband
FTTH	Fibre to The Home
GHz	Gigahertz
GSMA	Global System for Mobile Communications
ICT	Information Communication Technology
IoE	Internet of everything
IoT	Internet of Things
kHz	Kilohertz
LTE	Long-Term Evolution
mMTC	Massive Machine Type Communication.
NFO	Network Function Operator
NFV	Network Functions Virtualization
NFVO	Network Functions Virtualisation Orchestrator
NSE	Network Slice Engine
QoS	Quality of Service
RAN	Radio Access Network
SDN	Software Defined Network
SLA	Service Level Agreement
UHD	Ultra-High Definition
URLLC	Ultra-Reliable Low Latency Communications

VDSL	Very high-speed digital subscriber line
VLSI	Very large-scale integration

List of Figures

List of Tables

Abstract

Wireless networks are increasingly being integrated with cloud computing platforms and benefit from new paradigms. Two paradigms in this regard are software defined networking and network function virtualization. These paradigms alongside cloud integration defines the architecture of future wireless networks. In addition, it is important that future wireless networks which incorporate the software defined networking and network function virtualization provide subscribers with enhanced quality of service. In this case, an enhanced quality of service implies high throughput and ultra-low latency alongside reliable communications. These quality-of-service attributes are the network service provider and subscriber expectations from future communication networks.

In meeting the target performance expectations, future wireless networks are expected to have access to sufficient bandwidth resources. The provisioning of bandwidth resources arises in the abstraction of network slices. In the logical network system, the network slice is realized aboard a shared physical network with computing resources. The network slice is synonymous to the bandwidth and can be conceived as being the basic bandwidth unit that can be allocated to a given user context i.e., the subscriber utilizing an application. The future wireless network comprises multiple slices and aims to enable the utilization of slices in a manner that meets subscriber quality of service requirements.

The quality of service is an important concern that involves two aspects. The first aspect is that of ensuring that subscribers achieve the ideal throughput (in bits per second) and lowest latency (in seconds) from the network. The second aspect is related to the realization of call admission control. The aspect of call admission control relates to how subscribers initiating data calls or voice calls access network slices. This challenge is important for a network service provider deploying a network that incorporates the paradigms of software defined networking, network function virtualization alongside network slices. The research being presented proposes the incorporation of a buffer to improve the call admission control in relation to a multi-subscriber multi-priority call capable network that incorporates network slices. The performance evaluation is done through the approach of the Continuous Time Markov Chain and considers different network scenarios and contexts. The call admission control is investigated for the considered scenarios and contexts (seven scenarios) via the formulated metrics of call blocking probability and call dropping probability. These metrics are formulated from a state transition model comprising three call classes in a multi-subscriber scenario. In addition, the performance evaluation is executed via MATLAB numerical investigation. Performance analysis and evaluation of benefits shows that the incorporation of the buffer reduces the call blocking probability and call dropping probability by an average of (70.2–83.7) % and (51.6–83.9) %, respectively.

List of Publications

M. Reddy and S. Mbuyu, "Analysis of Buffer Influence on Network Slices Performance using Markov Chains," *2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, Durban, South Africa, 2022, pp. 1-6, doi: 10.1109/icABCD54961.2022.9856305.

The accepted, presented, and registered paper from the icABCD 2022 conference have been published in IEEE Xplore.

Titled - Analysis of Buffer Influence on Network slices performance using Markov chains.

The link is: https://ieeexplore.ieee.org/xpl/conhome/9855894/proceeding

Chapter 1 - Introduction and background to the study

1.1 Introduction

According to the Global system for mobile communications, wireless technology enables a digital society. Due to continuing technological advances, wireless technology can significantly transform sectors such as energy, automotive, logistics, manufacturing, healthcare, and finance. The current state of mobile network application arises due to the conflicting demands and preferences of businesses. 5G communication networks should be designed to deliver services that meet the requirements of different subscribers [1]. This can be realized by designing dedicated networks that can meet subscriber's differing preferences. Dedicated networks designed in this manner should enable the implementation of custom functionality for each subscriber. These networks operate on a common platform that enable the realization of network slicing. In this context, network slicing entails operating multiple logical networks on a common physical infrastructure [2].

The incorporation of network slices in this manner presents a paradigm change in 5G networks. The use of slices in a 5G network enables the realization of adaptiveness to the external environment. The use of network slicing enhances the use of digital services by business subscribers. This has the benefit of enabling a significant improvement in enterprise. In addition, mobile devices play an important role in the 5G context and enable the accessing of services by subscribers. They provide support to a significant number of services in the 5G ecosystem. The customizable network capabilities include data speed, quality, latency, reliability, security, and services. In this context, a network slice is an independent network that runs on a shared physical network, slices can provide a negotiated quality of service [3]. The technology enabling network slicing is transparent to business customers. A network slice could span across multiple parts of the network that comprises of a dedicated or shared resources. Figure 1.1 shows the application spectrum, quality of service expectations and services that are supported in 5G [4].

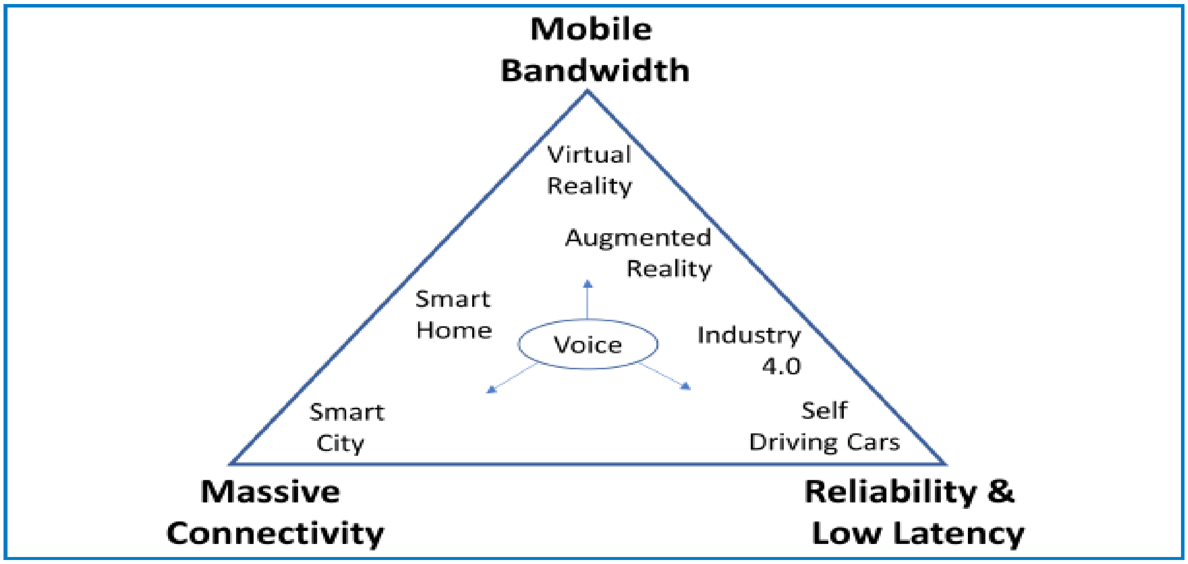


Figure 1.1: 5G mobile network evolution [4].

Background

1.2 Overview of Network Slicing

Conventionally, telecommunication networks were built with physical nodes comprising of a monolithic network. This network enabled the delivery of services from the operator to the users. In this case, the network was statically configured and manually managed within a “single tenant environment”. However, telecommunication networks need to evolve to support emerging business needs and incorporate new technologies [5]. The incorporation of these technologies enables the realization of evolved networks. These evolved networks (logical networks) can be designed, instantiated, and operated in a dynamic manner. These logical networks can meet the unique needs of subscribers while accommodating different business models and achieving resource self-management [6].

These logical networks comprise of multiple network slices. A network slice hosts multiple network functions, resource ownerships and relationships. In addition, a network slice can span multiple domains. Each domain can comprise of network-based functionalities and entities. These entities comprise of terminal, access network, transport network, core network, data centre domain that hosts third-party applications, as well as network management system [7].

In addition, network slices are realized via an end-to-end logical network hosted by a physical infrastructure. The physical infrastructure can provide quality of service at a specified level. Subscribers utilizing network slices via 5G technology are oblivious of the underlying technology [8].

Network slices have paths that traverse different aspects of a network. These multiple paths could span across multiple operators. From a computational perspective, network slices utilize dedicated or shared resources such as processing power, storage, and network bandwidth. In addition, a network slice has its own identity and is self-contained in its own resources. Furthermore, types of slices can be defined from a functional or behavioural viewpoint, hence, service provider's i.e., mobile network operators can deploy different types of slices for varying applications. In addition, multiple network slices (varying types) can be packaged as a single product. This product is now utilized by business customers that have different service requirements [9-10].

In [2], the GSMA observes that different capacities should be envisaged for network slices. Hence different types of network slices should be designed for use in future networks with the aim of meeting the needs of different user demand contexts. This is presented in Figure 1.2. Furthermore, network slices should exhibit a high degree of flexibility and agility. This is shown in Figure 1.3. The incorporation of flexible and agile capabilities implies that network slicing enables adaptive capabilities in 5G networks. In 5G networks, this enables the realization of the features of automation, orchestration, and sophisticated service creation. These capabilities enable the use of network slices which improves quality of service requirements for different services (verticals) that benefit from networking in 5G. In access networks, stacks spanning the lower level to the higher level can be sliced and customized to realize its performance requirements. The protocol stack realized in this manner can be designed with the goal of obtaining ultra-reliable low latency communications [11-12].

In a transport network, slice isolation is realizable via resource sharing and virtualization of a dedicated resource. Core network functionality provides customized network functions based on demands received from vertical industries [13].

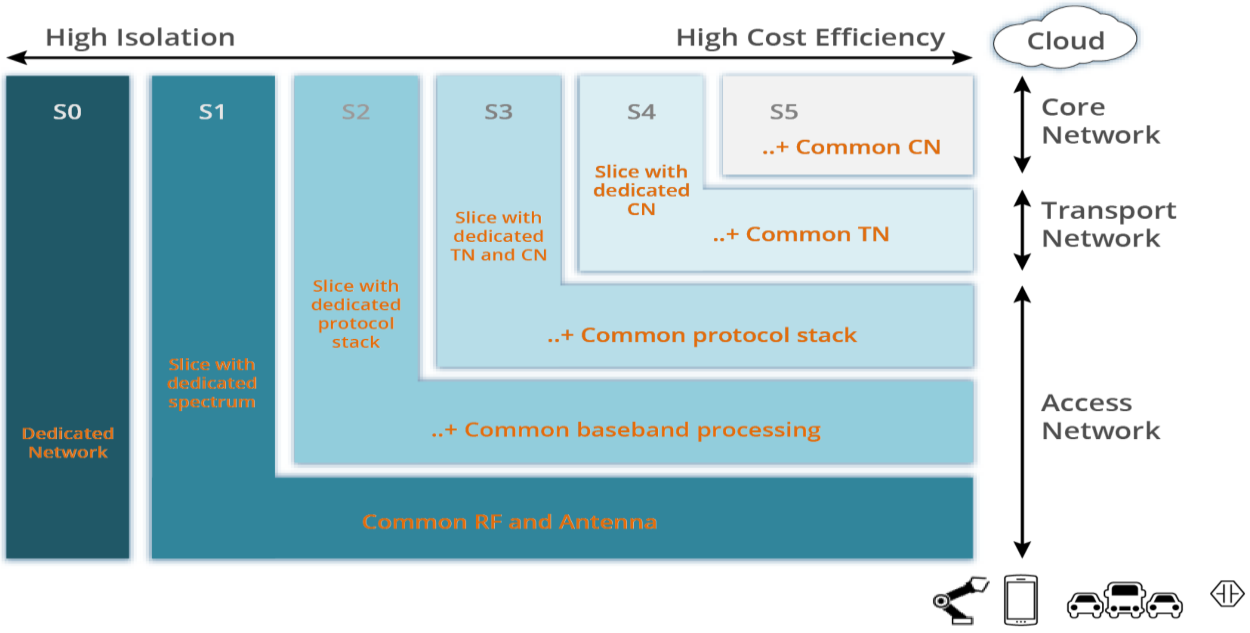


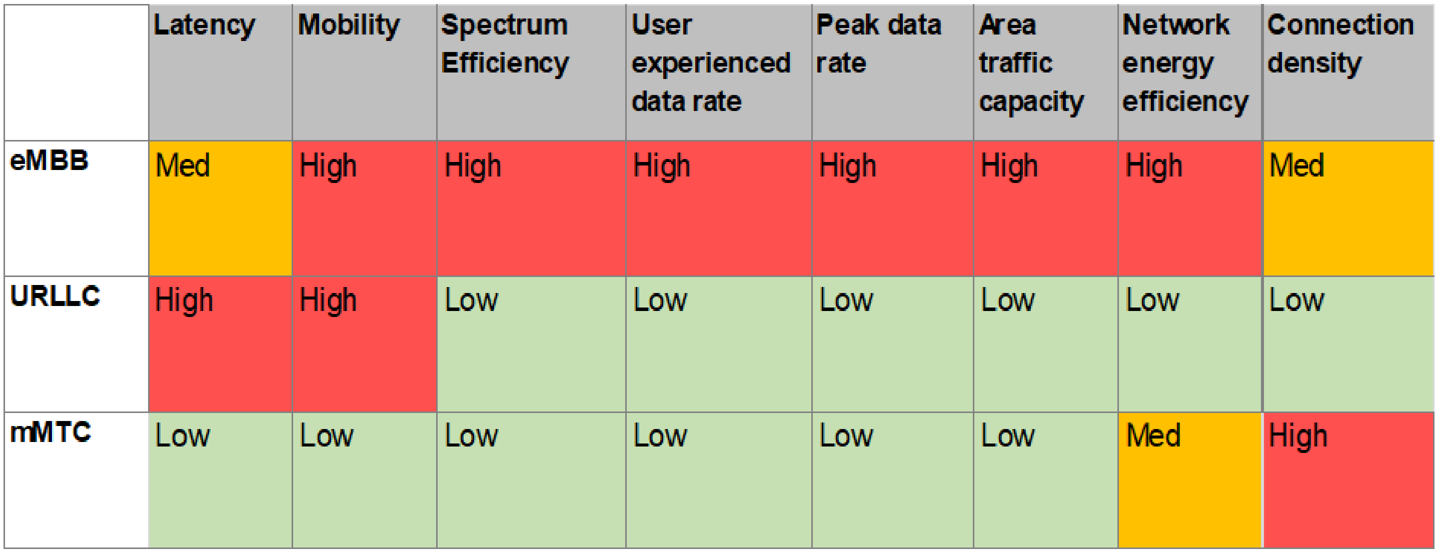
Figure 1.2: Network slicing different solution spaces to handle diverse vertical requirements [2].

In Figure 1.3 below where:

eMBB-enhanced mobile broadband

URLLC-ultra reliable low latency communications

mMTC -massive machine type communications



	Latency	Mobility	Spectrum Efficiency	User experienced data rate	Peak data rate	Area traffic capacity	Network energy efficiency	Connection density
eMBB	Med	High	High	High	High	High	High	Med
URLLC	High	High	Low	Low	Low	Low	Low	Low
mMTC	Low	Low	Low	Low	Low	Low	Med	High

Figure 1.3: 5G Requirements from different vertical industries [9].

1.3 The 5G technology Concept

5G networks should be capable of meeting the needs of different mixes of supported services. They should incorporate improved quality of service management capability in an end-to-end cloud environment [4]. The emergence of 5G networks should also consider ultra-dense heterogeneous networks. These networks can provide high data transfer and low latency capabilities as they use the 3GHz spectrum. The enhanced quality of service translates to improved movie download and uploads on mobile devices [14]. These capabilities should be provided over a wide area and support many internet–connected devices such as wearable health sensors and other internet of things-based technologies. In addition, 5G is diffusive and will influence new applications across multiple sectors as shown in Figure 1.4.

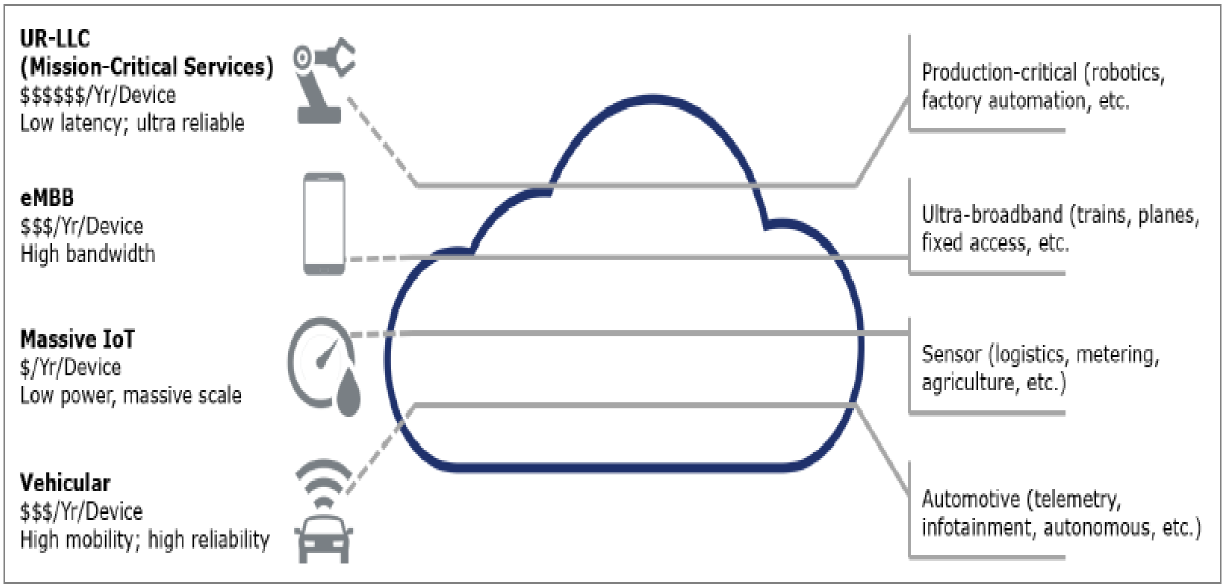

UR-LLC
(Mission-Critical Services)
$$$$$$/Yr/Device
Low latency; ultra reliable
eMBB
$$$/Yr/Device
High bandwidth
Massive IoT
$/Yr/Device
Low power, massive scale
Vehicular
$$$/Yr/Device
High mobility; high reliability
Production-critical (robotics, factory automation, etc.
Ultra-broadband (trains, planes, fixed access, etc.
Sensor (logistics, metering, agriculture, etc.)
Automotive (telemetry, infotainment, autonomous, etc.)

Figure 1.4: Network slicing concept [4].

1.4 Efficiency Requirements

Mobile broadband-based services and internet of things constitute the main drivers and provide basis for the prospect of future of mobile communications. In addition, mobile broadband based services provide a network support in an affordable way for the deployed 5G system [6]. The adoption of the 5G network should result in making significant improvements. These include spectrum efficiency (3 – 5) times, energy efficiency (100+ times) and cost efficiency (100 + times).

The definitions of these efficiency indicators are listed in Table 1.1

Table 1.1: 5G Key efficiency indicators [6].

Efficiency indicators	Definition
Spectrum efficiency (bps/Hz/cell or bps/Hz/km²	The data throughput per unit of spectrum resource per cell (or per unit area)
Energy efficiency (bit/J)	The number of bits that can be transmitted per joule of energy.
Cost efficiency (bit/Y)	The number of bits that can be transmitted per unit cost.

1.5 Capabilities

Liu et al. [6] state that 5G system and efficiencies must outperform previous communication systems in the following aspects:

i. connection density ($1 \text{ x } 106$ connections per km^2).
ii. traffic density volume density increased data rate (tens of Gbps per km^2).
iii. reduced end to end latency of the order of milliseconds.
iv. support for mobility speeds exceeding 50km/hr.
v. support for more increased connected devices.
vi. enhanced spectrum efficiency.
vii. increased traffic support capacity.

In addition, there is lower energy consumption in terms of communications required to execute, end to end RAN and CN connections, transportation, orchestration for heterogeneous network resource management, network slices expansion with increased network functionality and robust network slice functioning.

The expected key capabilities in 5G showing the projected performance benefits are presented in Figure 1.5.

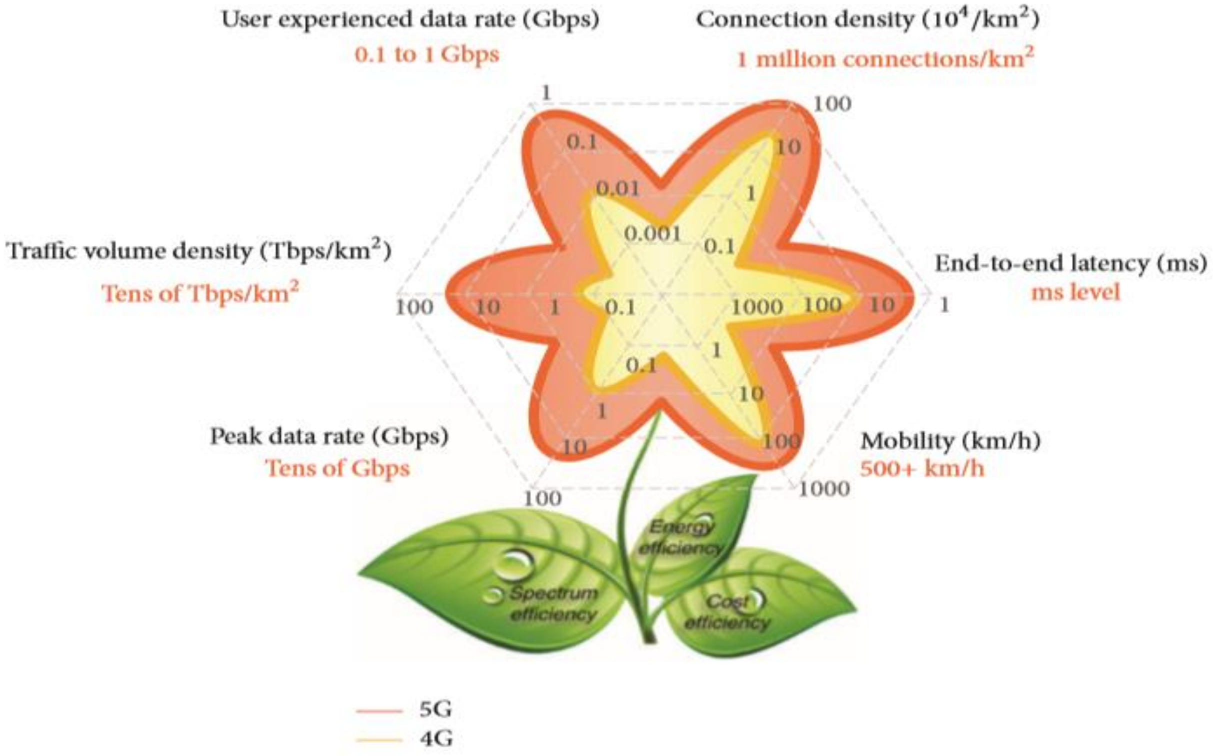


Figure 1.5: 5G key capabilities [6].

1.6 Software-Defined Networking (SDN) and Network Function Virtualization (NFV)

Network slices play an important role in providing resources enabling the delivery of services in 5G networks. In this regard, slices are crucial in the technologies of software defined networking (SDN) and network function virtualization (NFV). SDN and NFV jointly deliver the features of programmability, flexibility, and modularity [15]. These features enable the creation of multiple logical networks. These logical networks are custom designed for a given service that are supported on a common network [9]. Network slicing incorporated in SDN and NFV enable the realization of network softwarization. Lucena et al. [7] observes that network softwarization transforms network via the increased use of software–based solutions. The combination of SDN and NFV in the realization of network softwarization enables service delivery for multiple users with different use cases on a common network. This is shown in Figure 1.6.

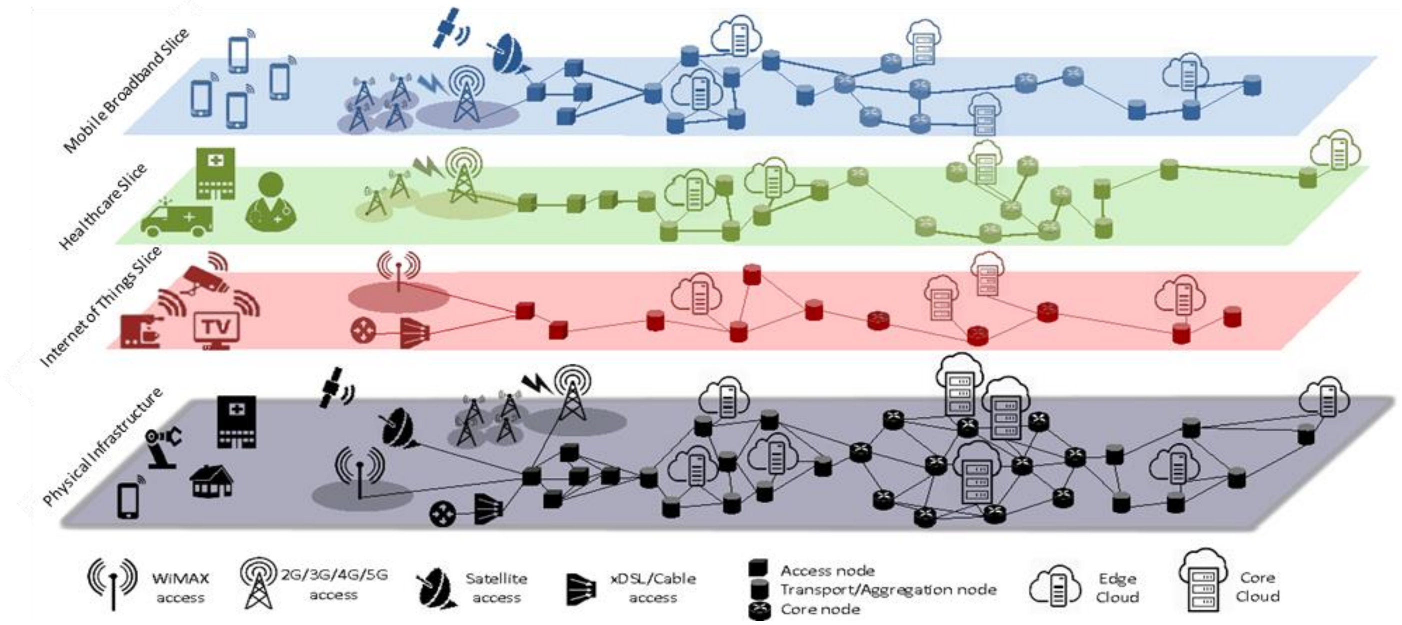


Figure 1.6: 5G network slices running on a common underlying multi-vendor and multi-access network. Each slice is independently managed and addresses a particular use case [7].

An NFV based platform is presented by Lucena et al. [7] where the proposed platform incorporates the multi datacentre service ChalN emulator for network services. The platform supports the system to deploy virtual resources for services in multi-domain infrastructure. The presented platform demonstrates that virtualization plays an important role in resource abstraction. It is observed in Nguyen et al. [8] that NFV enables the transition of functions from hardware based standalone boxes to the software-based cloud environment. In this case, the cloud environment comprises commodity servers as shown in Figure 1.7. In this case, each NFV based function is executed aboard a virtual machine. In addition, it is observed that SDN and NFV have complementary features and capabilities. This is because they enhance the innovation of new services towards software-based ecosystem. In this regard, NFV enables SDN realization. This is done by virtualizing elements such as the: SDN controller and SDN data forwarding entities. The data forwarding entities are executed in the cloud with support for the dynamic migration of these components to their optimal locations. In this way, SDN enables the realization of the goals of NFV by enabling programmable network connectivity. This enables the realization of optimized network traffic management [8]. Costa-Requena, et al. [10] notes that a candidate 5G technology should be backward compatible with existing network technologies. In addition, the use of SDN and NFV provides a seamless migration based on the needs of the service provider. Furthermore, SDN and NFV incorporation supports incremental updates of network elements while maintaining backward compatibility. Collicutt [9] states that the network slice engine executes quality of service configuration as defined by NFVO descriptors. In this case, the network slice engine is implemented via the Java io framework. This is executed as an external component that communicates with the NFVO via open baton SDN. In addition, the network slice engine (NSE) utilizes the plugin mechanism to support driver setup.

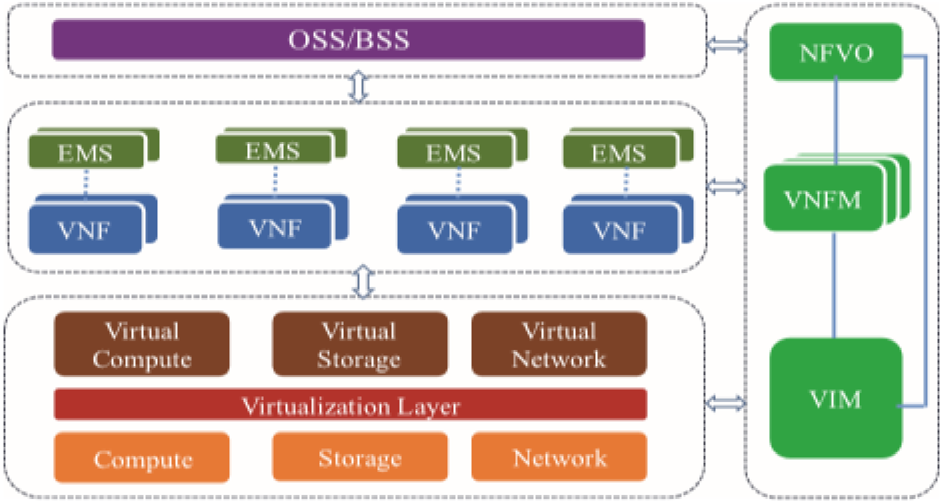


Figure 1.7: NFV architectural framework [8].

Cisco forecasting identifies that an increase in the number of devices being supported by mobile networks will exceed global population and reach 11.6 billion with 30.6 exabytes of monthly traffic. 5G networks will be expected to support a wide variety of use cases due to the emergence of internet of everything [4]. This will support ultra-high-definition video, augmented reality, and high-capacity communications, additional applications being supported comprise mission critical internet of things and low latency requiring autonomous vehicles. These services enable the realization of ultra-reliable services alongside delivering efficient network slicing. This is done via the execution of real-time dynamics of network resources and uncertain customer demands [16].

Network slices provide network services in an end-to-end fashion that provide connectivity and transport capabilities. This enables subscribers to control and manage these resources while providing improved quality of service [16]. Network slicing provides a middle ground approach to enable the support of multiple services on a shared platform with defined QoS in each logical partition as shown in Figure 1.8.

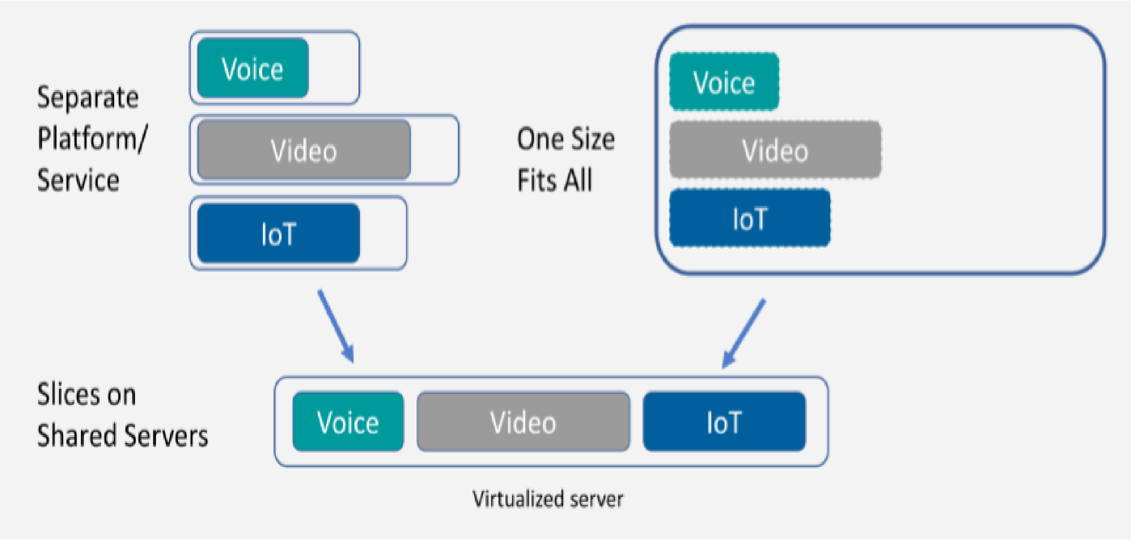


Figure 1.8: With slicing, networks can be adapted to customers and applications [16].

Network slicing provides a model for operating virtualized and programmable cloud-based infrastructure. This enables the features of automatic and granular orchestration which are crucial to manage the logical slices on the underlying infrastructure. Hence, it is important for network operators to possess processes and systems enabling slice creation efficiently, its capability to deliver end to end services, slices are decoupled and independent of one another. This results in networks with slices having more versatility than existing networks via the incorporation of enhanced transport functionality. These features provide the capabilities of catalysing the change required to realize industry growth as shown in figure 1.9 [13].

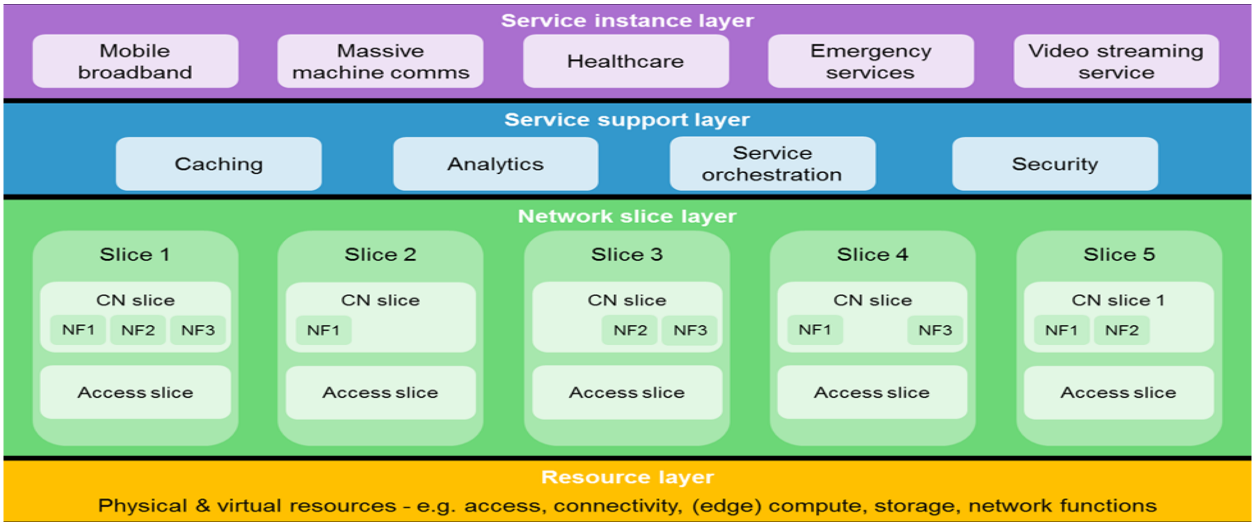


Figure 1.9: Diagram of slicing [13].

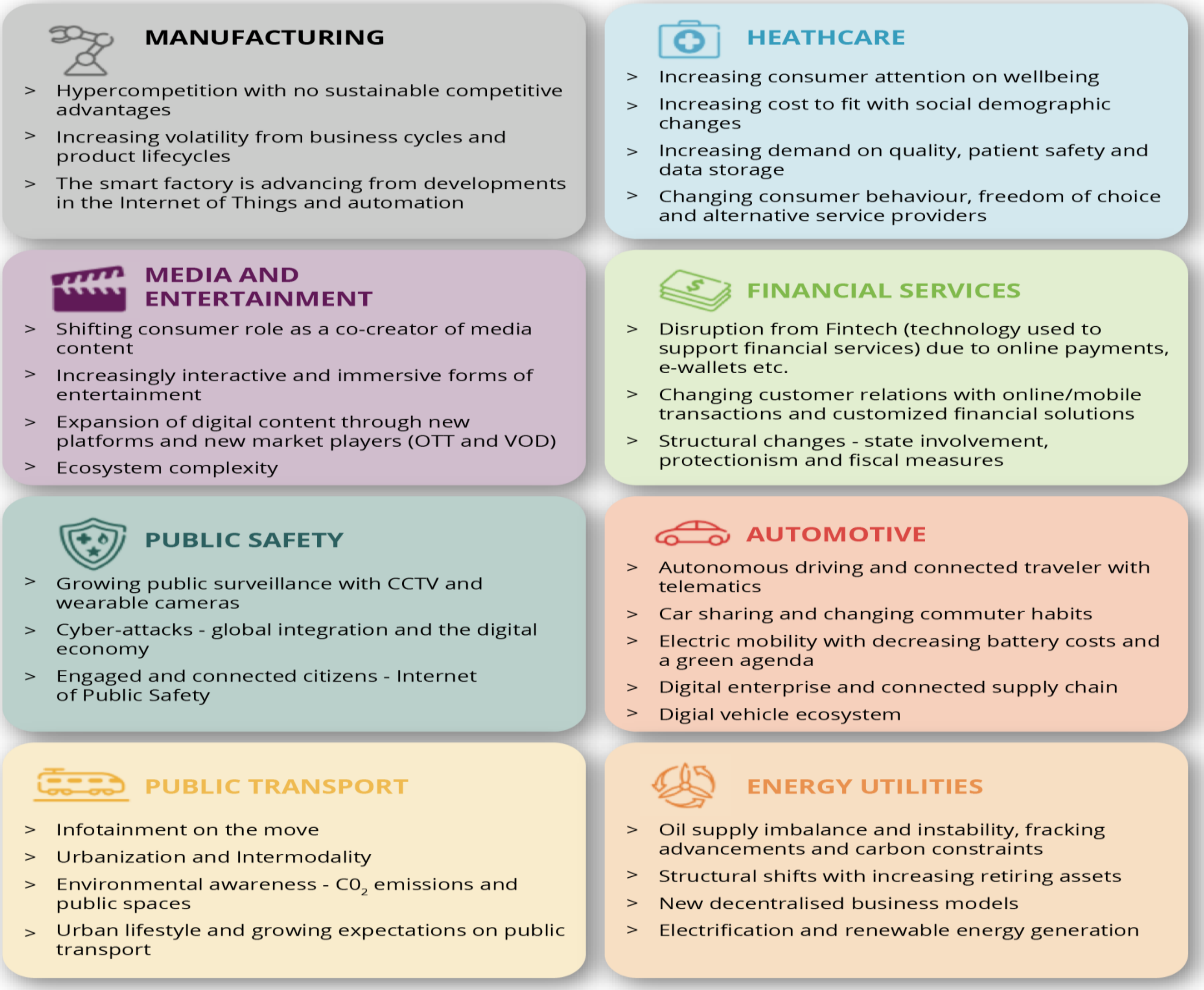
1.7 Problem Statement

The emergence of cloud computing systems plays an important role in future generation communication networks. The concerned communication networks are those that incorporate the software defined network (SDN) and network function virtualization (NFV) paradigms. The paradigms of SDN and NFV enable the provision of services in cloud driven radio access networks. The future generation networks incorporating SDN and NFV capabilities are expected to provide users with high speed and ultra-low latency services and access. In addition, it is also important to ensure that network slices in these networks are utilized in a manner that meets the service preferences of different subscribers associated with a given network service provider. Therefore, it is important to design mechanisms recognizing that network slices are finite, which ensures that subscribers do not experience degrading call support capabilities. The problem being addressed in this regard considers the design of network slice resource allocation mechanisms for a network service provider (incorporating SDN and NFV) that provides services to multi-class subscribers with voice and data calls having varying resource access priorities. In addition, the problem should be addressed in a manner that evaluates the performance improvement derivable from the use of a mechanism that enhances network slice resource usage that meets subscriber preferences.

Trends show that the evolution of networks such as 256k dial up modems, DSL, VDSL, FTTH, and LTE enable the realization of 5G. The exponential growth in data demand underlies the realization of an increasingly digital and global connected world in supporting more services. In addition, there is an explosion of services requiring high quality of service performance. Furthermore, future 5G networks will also be expected to support increased traffic arising from emerging bandwidth intensive applications. These are expectations for emerging ICT–driven transformation.

The need for 5G to incorporate support for internet of everything and provisioning of sensor connectivity alongside the provisioning of connected robots for industrial applications. The realization of this goal requires supporting vertical industries while maintaining efficiency and low costs. Industries that benefit from vertical integration and 5G support are shown in Figure 1.10.

This research proposes the use of buffering for enhancing call quality in future networks incorporating network slicing where slices host the resources enabling data and voice execution. The contribution of this research is the design of a mechanism considering different service priorities in 5G networks comprising network slices. The proposal of a network buffering mechanism for resource access in the context of multi-subscriber priorities in 5G networks. The presented research develops an analytic model for evaluating the role of a buffer in a 5G network comprising network slices which also executes voice calls and data calls having varying resource usage preferences (i.e., priorities). The role of a buffer is examined in this case and important metrics are formulated. These metrics are the call blocking probability and call dropping probability. The identified metrics are investigated in the absence (existing mechanism) and in the presence of a buffer (proposed mechanism). The call blocking probability and call dropping probability are evaluated considering different call arrival rates, service rates and buffering rates. Furthermore, numerical evaluation of the call blocking probability and call dropping probability is executed in the case of existing and proposed mechanisms. In Figure 1.10 each identified industry has connectivity and communication needs requiring specific solution [5].



MANUFACTURING

> Hypercompetition with no sustainable competitive advantages

> Increasing volatility from business cycles and product lifecycles

> The smart factory is advancing from developments in the Internet of Things and automation

HEATHCARE

> Increasing consumer attention on wellbeing

> Increasing cost to fit with social demographic changes

> Increasing demand on quality, patient safety and data storage

> Changing consumer behaviour, freedom of choice and alternative service providers

MEDIA AND ENTERTAINMENT

> Shifting consumer role as a co-creator of media content

> Increasingly interactive and immersive forms of entertainment

> Expansion of digital content through new platforms and new market players (OTT and VOD)

> Ecosystem complexity

FINANCIAL SERVICES

> Disruption from Fintech (technology used to support financial services) due to online payments, e-wallets etc.

> Changing customer relations with online/mobile transactions and customized financial solutions

> Structural changes - state involvement, protectionism and fiscal measures

PUBLIC SAFETY

> Growing public surveillance with CCTV and wearable cameras

> Cyber-attacks - global integration and the digital economy

> Engaged and connected citizens - Internet of Public Safety

AUTOMOTIVE

> Autonomous driving and connected traveler with telematics

> Car sharing and changing commuter habits

> Electric mobility with decreasing battery costs and a green agenda

> Digital enterprise and connected supply chain

> Digial vehicle ecosystem

PUBLIC TRANSPORT

> Infotainment on the move

> Urbanization and Intermodality

> Environmental awareness - $C0_2$ emissions and public spaces

> Urban lifestyle and growing expectations on public transport

ENERGY UTILITIES

> Oil supply imbalance and instability, fracking advancements and carbon constraints

> Structural shifts with increasing retiring assets

> New decentralised business models

> Electrification and renewable energy generation

Figure 1.10:10 Key Industries and their challenges [5].

1.8 Purpose of the study

The purpose of the study is:

i. To evaluate and determine the need and benefits of 5G Network Slicing, while analysing user demand uncertainty.
ii. To investigate and simulate a bandwidth prioritization mechanism using Markov Chain model, while characterizing the users into three different priority classes, class A high priority, class B medium priority and class C lowest priority.

1.9 Objectives of the study

The objectives of the study are:

i. To develop analytic models for analysing dynamics in a communication network having calls with different bandwidth resource access priorities.
ii. To examine the role of a buffer in ensuring the execution of voice calls and data calls considering the paucity of network slices i.e., bandwidth resources.
iii. To formulate the call blocking probability and call dropping probability before and after the introduction of a buffer in the proposed network.
iv. To evaluate the performance of the call blocking probability and call dropping probability before and after the use of the proposed buffer considering different call arrival rates, service rates and buffering rates.
v. To execute performance analysis to determine reduction in the call blocking probability and call dropping probability due to the use of the buffer.

1.10 Research questions

Improvement in Call Performance: What is the reduction in call blocking probability and call dropping probability for a multi-subscriber and multi-priority call network using network slices for due to the introduction of a call buffer?

i. What is the variation between, the call arrival and departure rates, call dropping probability and call blocking probability for multi priority call communication systems?
ii. What is the obtainable call dropping probability and call blocking probability for communication systems supporting calls with varying bandwidth access priorities?

iii. How does the use of a buffer reduce the call dropping probability and call blocking probability for systems supporting dynamic priority calls?

1.11 Research methodology

The research investigates the improvement in call performance for the slice based multi-subscriber and multi-priority call network using mathematical model. The mathematical model enables the formulation of the call blocking probability and call dropping probability metrics. The further investigation of the performance of these metrics in each network is done via numerical simulation. The numerical simulation tool that has been used in this regard is MATLAB due to the ability to realize the different network contexts. In this case, the network scenario context is influenced by call arrival rates (descriptive of user demand for network) and call server rates (indicative of bandwidth available to meet subscriber demands**).**

The operational framework enabling the realization of the research objectives is described in the research methodology. The research methodology describes the mechanism for collecting data related to the study. This is done via constructivism paradigm and assumes that events and experiences are derived by individuals. Hence, individuals enable the construction of realities in which individuals participate by deciding the most suitable method for investigating the problem under consideration.

1.12 Research approach

The study adopts a multi-method approach and gives more weight to data and simulations analysis in the presented study. This is triangulated at different levels to accommodate a quantitative approach and analysis. The presented research incorporates qualitative and quantitative via triangulation.

1.13 Significance of the study

The study recognizes the increasing number of subscriptions to mobile technology with increasing annual growth rate. In addition, the increasing subscriptions to 5G are projected to exceed 849 million. This makes it necessary to design solutions enabling industries to positively participate and experience ICT driven transformation. In addition, industries should support business trends such as hyper competition, new customer power and sophistication. These are features enabling industries to adopt the fast-paced capabilities emerging from the adoption of new technologies. This prompts an investigation to evaluate and analyse the service quality alongside user benefits for services enabled via network slicing in 5G networks [1].

1.14 Structure of the research

Chapter one: Presented the introduction and background to the study.

Chapter two: Literature review.

Chapter three: Presents the research, methodology and simulations.

Chapter four: Results, data analysis and interpretation.

Chapter five: Summary, conclusions, and recommendation.

Chapter 2 - Literature Review and Contributions

2.1 Introduction

Wireless communications have become an integral part of our lives. In this chapter we examine a spectrum of topics in the world of communication networks by investigating the theoretical concepts and practical applications that are incorporated into these networks. Telecommunication has become a backbone of our interconnected world and it allows data exchange and communication across vast distances.

This chapter will commence with a discussion on theoretical alternatives and an introduction to wireless propagation to understand the fundamental principles of wireless networks. It will then explore critical topics such as traffic analysis, network functions virtualization (NFV), and software-defined networks (SDN) that have critical roles in improving the network. This study will further investigate network performance and quality of service (QoS) consideration that include resource management in cellular networks. To understand the network dynamics, Markov chains and their applications in the context of network slicing are introduced. Finally, the critical aspects of bandwidth management and prioritization that are essential for optimizing network resources and ensuring seamless communication are presented.

Through these discussions, this chapter aims to provide a comprehensive foundation for understanding the intricate operation of mobile systems and network management.

2.2 Theoretical alternatives in Telecommunication systems

Telecommunication networks enable the delivery of information through media such as wire, radio, optical, or other electromagnetic systems. This enables the exchange of information between entities in a communication session. In a simple defined way, telecommunications will occur between two stations or nodes that engage in information transmission. These stations constitute the transmitting stations and receiving stations that are involved in exchanging data. This can, either be in wireline systems or wireless cellular networks.

As a matter of fact, Simon in [19] describes the processes of information transmission from the source to a specified destination, in the following manner. He states that the process of signal transmission could involve several processes such as signal formatting and signal modification. By and large it is important that the receiver should be able to overcome the unexpected modifications on the signal caused by the channel. These processes are applicable to all communication systems whether wireless or otherwise.

For all intents and purposes wireless communications are networks that do not rely on physical connections for connectivity. Wireless networks use radio or electromagnetic waves to connect to the devices. According to [16] these networks allow devices such as laptops, smartphones, and tablets to connect to the internet without a physical connection. Usually, the networks are operated by mobile operators that use a system base station to provide network coverage to their subscribers. As a matter of fact, wireless or mobile communication has become a norm and a leading form of technology (examples will include, cellular communication, Wi-Fi, LiFi and many more wireless technologies) when it comes to the transfers of information from one node to another. Mobile communication has changed the way people interact and work. By and large, wireless communication has a compelling influence on businesses and industries, such as health care, logistics and retail to improve their operation and customer services [20].

2.3 Introduction to wireless propagation

The propagation channel has the biggest influence on the design of a transmitter and wireless receiver. As the signal travels from the transmitter to the receiver through the wireless channel, power is usually lost. Between the transmitter and the receiver, several propagation pathways with various delays are created by reflections, diffractions, and scattering. In conclusion, multipath caused by wireless propagation causes frequency selectivity in the channel and results in a decrease of received signal power [19].

Architecture of a generic communication system is illustrated in Figure 2.1. The wireless channel is an essential part of operation, design, and analysis of any system.

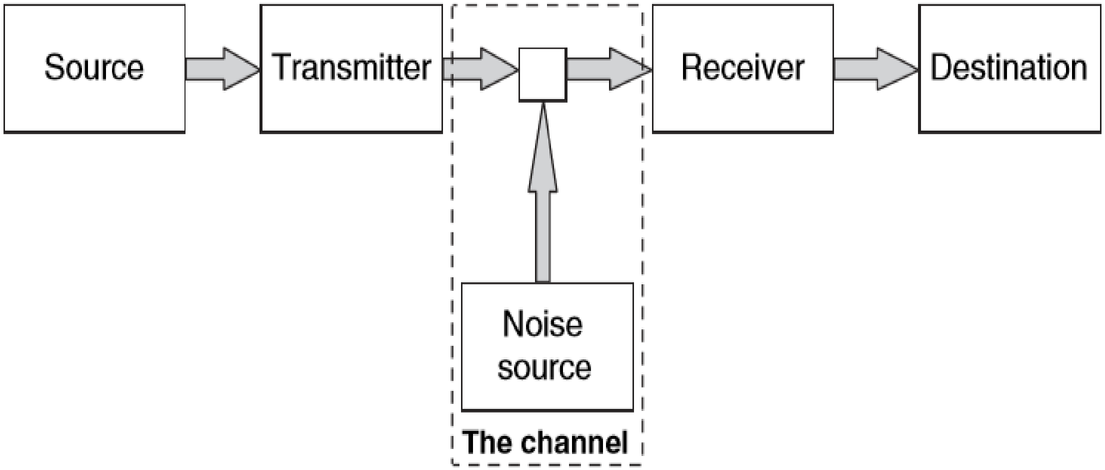


Figure 2.1: Architecture of a generic communication system [19].

The basic resource exploited in wireless communication systems is the electromagnetic spectrum. Practical radio communication takes place at frequencies between 3 kHz to 300 GHz, which corresponds to wavelengths in free space from 100 km to 1 mm. As illustrated in Figure 2.2, the power loss encountered during transmission between the base and the mobile is influenced by the path loss. This depends on the antenna height, carrier frequency and distance. The path loss is given by equation 2:1:

$$\frac{P_R}{P_T} = \frac{1}{L} = k\,\frac{h_m h_b^2}{r^4 \quad f^2} \tag{2:1}$$

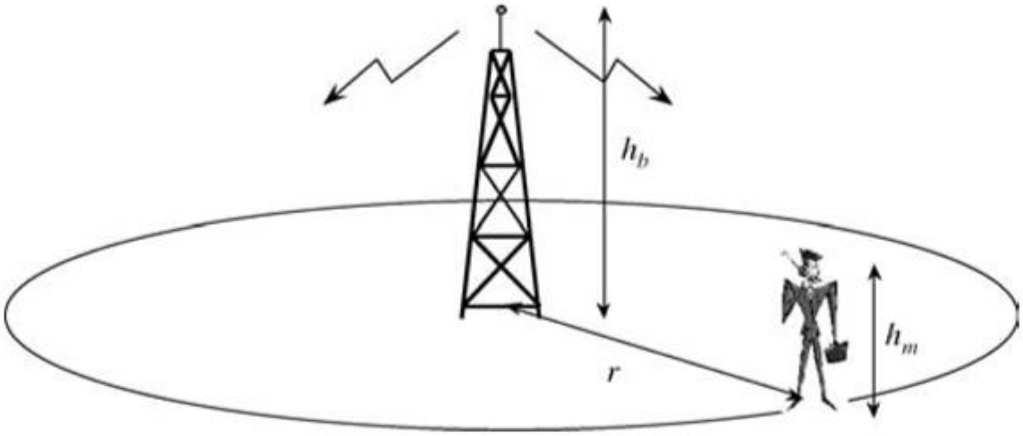


Figure 2.2: Basic geometry of cell coverage [19].

Where:

P_R the power received at the mobile input terminals (W).

P_T the base station transmitting power (W).

h_m and h_b are the mobile and base station antenna heights, respectively.

r the horizontal distance between the base station and the mobile (m).

f the carrier frequency (Hz).

The quantity L is the path loss and depends mainly on the characteristics of the path between the base station and the mobile device.

Dependencies are functions of the environment type (urban or rural). At higher frequencies, the range for a given path loss is reduced and more cells are required to cover a given area. To increase the cell radius for a given transmit power the antenna height must be adequate to clear surrounding clutter i.e., trees and buildings, but not so high as to cause excessive interference to distant co-channel cells. The antenna must be chosen with regard for the environment and local planning regulations.

2.4 Traffic

It was observed in Costa et al. [10] the number of channels required to guarantee services to every user in a particular system is almost impractically. It can be noted that the number of users requiring channels simultaneously is less than the total number of users. Traffic is measured in erlang. Indeed, one should note that one erlang (E) is equivalent to one user making a call for 100% of the time. A voice user generates around 2–30 mE of traffic during the peak hours of the wireless system.

2.5 Network Functions Virtualization (NFV)

According to Mebarkia et al. [21], technologies such as software defined networks (SDN) and network function virtualization (NFV) enable flexibility in operation, a reduction of the costly hardware to be deployed in networks and improve service agility. These Virtual network functions (VNF) operate as virtual machines on servers that are arranged in chains using service function chaining.

Researchers such as Ordenez et al. [22] observe that 5G networks require enhancements to support and meet the increasing bandwidth demands. Network softwarization realized through SDN and NFV enables the realization of this goal. As shown in [22] the resulting network incorporates network slices.

Basta et al. [23] explain that mobile operators are expected to support the increasing demand for data, embrace service innovation and achieve cost reduction. These capabilities can be realized by adopting SDN and NFV technologies. The need to support the increased demand for data services implies that there is an increase in the volume of data being transported through the mobile networks. The use of NFV enables the use of gateway in the manner as shown in Figure 2.3. The network element could either be a transport switching element or SDN network element. This increases the flexibility of data-flow handling between transport network and data centres as required by operators.

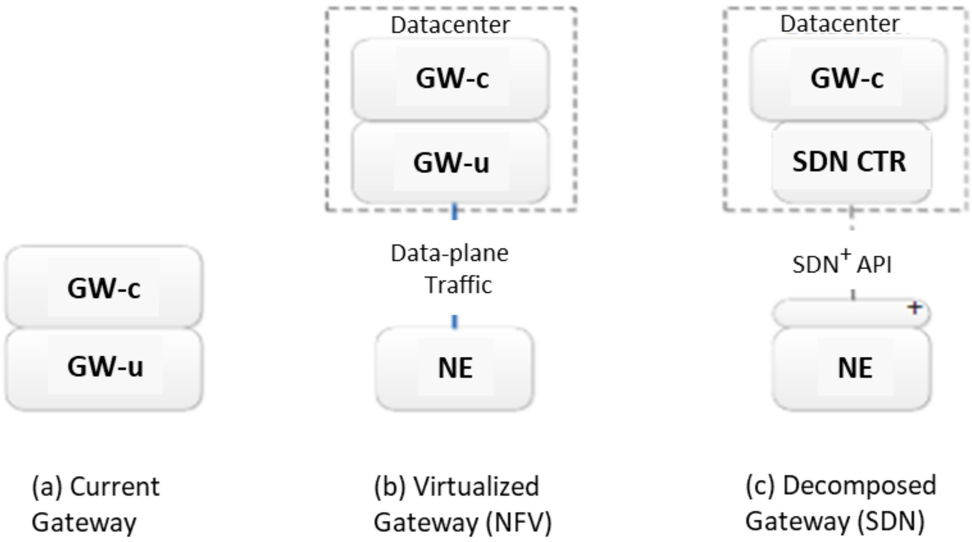


Figure 2.3: Mobile core gateways re-design [23].

Network function virtualization (NFV) enables the software implementation of previously hardware-based network roles such as routers firewalls and load balancers as shown in Figure 2.4. Hence the use of NFV can enable cost reduction in mobile networks [24]. Basta et al. [23] in their research propose the implementation, of mobile network functions as software executables that use commodity hardware or data centre resources. The deployment of NFV enhances network scalability, load optimization, energy savings and low costs. The incorporation of NFV enables the realization of dynamic virtualized platforms with features of rapid instantiation (enabling improved resource usage).

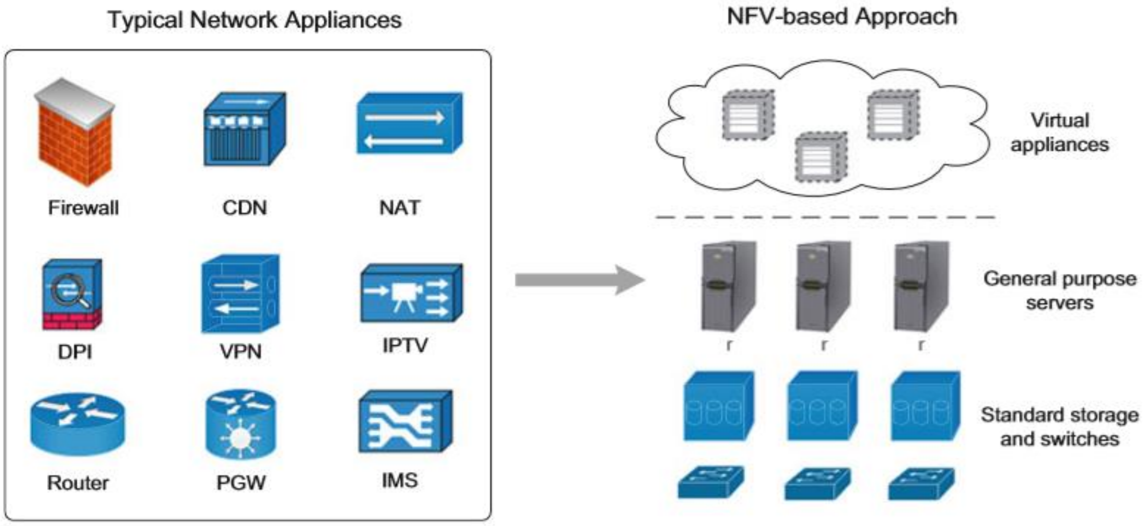


Figure 2.4: NFV-based approach of instantiating network function [24].

The discussion in Kazmi et al. [25] also recognizes that the adoption of NFV enables the reduction of capital expenditure, operational expenditure while making the system more flexible. NFV enables the realization of transforming network nodes into virtual functions. The execution of instantiation via NFV is shown in Figure 2.5.

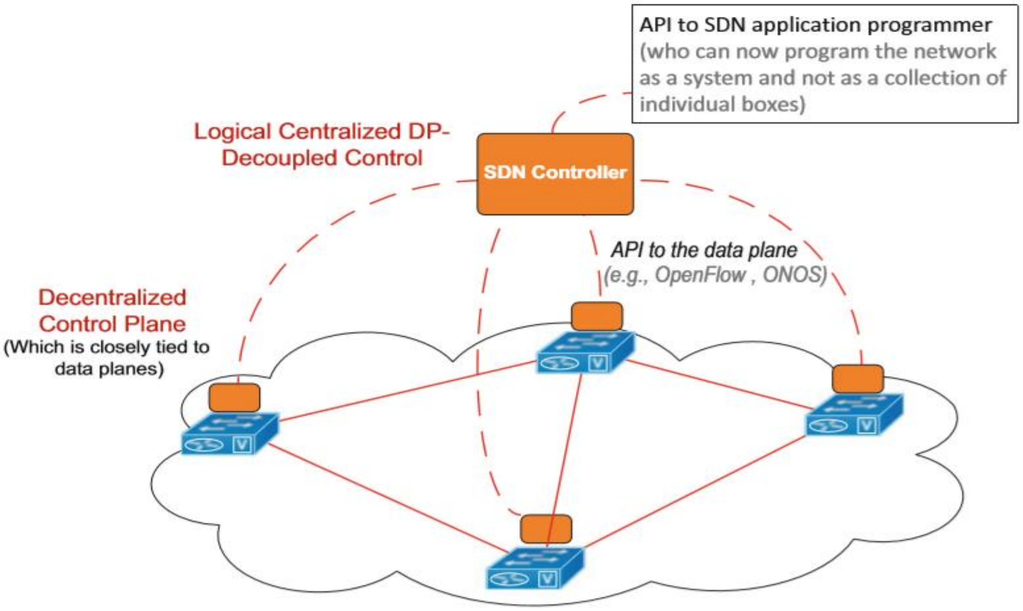

API to SDN application programmer
(who can now program the network as a system and not as a collection of individual boxes)
Logical Centralized DP-Decoupled Control
SDN Controller
API to the data plane
(e.g., OpenFlow , ONOS)
Decentralized Control Plane
(Which is closely tied to data planes)

Figure 2.5: Software defined networking [25].

2.6 Software Defined Networking SDN

The incorporation of SDN enables the acquisition of increased agility and flexibility of future networks. This will enable networks to be capable of meeting the requirements of different subscriber categories. Generally, SDN is deemed a complementary deployment concept for mobile networks. Its use enables, the introduction of capabilities such as network functions decomposition (NFD) where control–plane functions are appended to a data centre based on logically centralized controller. Data plane functions in the SDN are realized through a transport network. The inclusion of SDN enables network programmability and flexibility. This enables the dynamic steering of data thereby improving the flows within the network. This enables the realization of the provisioning of an improved quality of service and subscriber experience alongside significant cost savings. The introduction of NFV has the benefit of transporting network data to the operator’s datacentre.

This imposes additional load on the network and may have significant impact on the deliverables of the network such as throughput [25].

The use of NFV enables the decoupling of the control plane from the data plane. In this case, the logical control plane enables the easier configuration of effective management functions. This was previously executed through the design of conventional and traditional hardware that executes the required network functionalities. However, this has high costs and is also non scalable in the case of large-scale networks. The SDN framework comprises of different types of controllers such as Floodlight, Onix and Nox, while OpenFlow is the most used interface for packet flow. There are two categories of controllers in an NFV implementation. These are local and root controllers. Local controllers are connected to one or more switches which is further controlled by the root controller. The root controller other than controlling the local controller performs functions that need the network wide view of the network [15].

The capabilities of SDN and NFV enable the realization of a cloud centred network with improved robustness. This is important, considering the increasing role of software solutions and cloud computing platforms in future networks. The transition to the SDN and NFV paradigms also enhances the granularity of utilizing network resources. This results in improved network resources optimization. An important network resource in this regard is the network bandwidth resource. The SDN and NFV paradigm plays an important role in future cloud integrated radio access networks [25].

The paradigm and architecture of the SDN framework is the control entities and the multi-tier entities in future networks. The concerned networks are those that integrate the SDN and NFV paradigms. These control entities enable the determination of the way that resources i.e., the computing resources, bandwidth and power are utilized. The utilization of resources in this case is done with the goal of ensuring that quality of service derivable to the subscriber meets the highest standards possible. In this case, the desired quality of service performance is high throughput, and ultra-low latency. From a network perspective, some of the important metrics are spectral efficiency (bandwidth efficiency), and power efficiency [14].

The discussed aspects in [22] relate to the network architecture and describe the relations between the control entities and data communication entities. This does not consider the manner of how calls i.e., voice and data calls emanating from users and subscribers to the concerned wireless networks. This is important as the concept of providing the desirable level of quality of service (QoS) becomes relevant only after concerned calls have been admitted into the network.

However, the concern of addressing the utilization of the network slice in future networks incorporating SDN and NFV is important and influences the call admission performance in this regard [22].

In summary, NFV and SDN both utilize software components however they are different, NFV converts network processes by themselves into software applications, while SDN virtualizes the management of networks to benefit application-based traffic prioritization. SDN also reduces the cost of the network, as expensive switches and routers are not required while NFV increases the networks scalability and agility.

SDN and NFV are distinct technologies that address different aspects of network management. SDN focuses on network control and traffic management, while NFV is centred around virtualizing network functions. Combined they provide an essential requirement for optimizing network infrastructure, especially for evolving the telecommunication and cloud computing environments [14].

The focus of the next section is on the discussion on the role of the utilization of network slices in networks incorporating NFV and SDN, as network slices are considered the basic bandwidth unit corresponding components in SDN's.

2.7 Network performance and Quality of service

For all intents and purposes, Claudio in [26] explains that mobile communication networks have increased in application areas and vertical sectors. The advent of fifth generation communication networks (5G) has made it imperative to integrate more services in the communication network ecosystem to improve the quality of service. This can be realized by designing dedicated networks that can meet the subscriber's different preferences. Dedicated networks designed in this manner should enable the implementation of custom functionality for each subscriber. These networks operate on a common platform. The use of innovative mechanisms in 5G should be incorporated to meet the preferences of subscribers for improved realization of network slicing.

According to Chochilouros et al. [27] a network slice is defined by using technologies that are transparent to subscribers. They consist of multiple aspects of the network terminals, access network, core network and transport networks. These entities could be deployed across multiple operators and service providers. Slices host resources that can be utilized in a dedicated or shared manner, network slicing implementation in 5G finds useful applications in future evolved networks. These networks are designed to be dynamically operated and incorporate the capabilities of instantiation.

Alotaibi in [28] explains that composing logical networks are expected to meet the demands of subscribers via provisioning of the resources to meet the service level requirements for different services and applications. This can be achieved by incorporating the features of dynamic resource self–management and a slicing mechanism that allows virtual networks to provide personalized services on request to ensure high network speed, low latency, improved reliability, security, service delivery and availability of this quality of service.

The GSMA Alliance in [1] observes that network slices are expected to have different capabilities. This is necessary to meet the quality-of-service preferences for different subscriber applications, demand and use contexts [1]. Network slicing should consider the diffusive nature of 5G networks in hosting applications. 5G networks should be capable of providing improved quality of service in an end-to-end cloud computing environment. They consider the emergence of ultra–dense heterogeneous networks. The enhanced quality of service should also translate low latencies for movie uploads and downloads via mobile devices. In this case, mobile devices comprise devices such as health sensors wearables, and internet of thing-based access devices and technologies. These devices can connect to the internet.

Mebarkia et al. [29] explains that it is important to design solutions and mechanisms that enable different services. These services require access to network slices. The access to network slices should be done while considering different subscriber preferences and service categories. It is also imperative to consider that 5G networks incorporating network slices can provide services to subscribers with varying quality of service preferences. The future of 5G network is expected to support multiple users to execute voice and data calls. These users require access to network slices in varying amounts. This is because network slices serve as the main abstraction of the resources that enable call execution. The concerned network should be able to support different call preferences and resource demand patterns for the network to accommodate different subscribers and user quality of service preferences.

Pervej et al. [30] explains that Wireless communications have experienced dramatic changes over the past few decades, as user penetration keeps increasing and new applications have been consistently emerging, wireless technologies have evolved from one generation to the next to provide more and more capacity and improved user quality of service (QoS). The thrust for these changes is the crave for enhanced data rate, latency, energy efficiency, and QoS. The existing network has ensured auspicious performance delivery, the new and challenging demands on capacity and other performance have never ceased to emerge. The inadequacy of the existing technology has become apparent with the inclusion of the Internet of things (IoT). To guarantee the required performances, researchers are in search for new technologies to incorporate in the fifth generation (5G) and beyond. The need to support multiple users makes it important to ensure that the admission of calls does not result in a degradation of calls being executed. This is a basic requirement of communication networks. This performance perspective is often described by the metrics of the call blocking probability and call dropping probability.

2.8 Resource management in cellular networks

Communication network resources are scarce and should be efficiently managed to obtain an improved quality of service for subscribers. A resource is a manageable unit with own attributes and capabilities enabling the delivery of a given requirement. In addition, a resource is an entity that can be utilized to provide services in response to a subscriber’s request. Sarmah et al. [31] observes that evolution in wireless networking technology has resulted in increased traffic across the existing and future networks. The need to address the challenge of congestion is also recognized. The use of bandwidth management and allocation mechanism is recognized to be necessary. The bandwidth is the data carrying capacity of the network. Insufficient bandwidth degrades subscriber quality of service. This makes it challenging for users to execute their online tasks. The acquisition of extra and additional bandwidth is a feasible solution to this challenge. However, purchasing extra bandwidth is expensive and beyond reach for most organizations therefore the limited available bandwidth must be optimally managed to solve this problem. Bandwidth management is a technique through which resources are allocated to applications or users in an efficient and productive manner. If there is no management scheme in a network an application or users can acquire all the available bandwidth resources hence restricting other application and users from the network. Bandwidth management can be implemented by differentiating the network traffic into classes by application and service types, for each class different priority can be set, based on priority value and class that can access the network.

Mamman et al. [32] state that quality of service requirements is not very strict for all traffic types, the multi-level bandwidth adaptation technique improves the forced call termination probability as well as provides priority of the traffic classes in terms of call blocking probability without reducing the bandwidth utilization. A bandwidth model is proposed that releases multi-level of bandwidth from the existing multimedia traffic calls. The amount of released bandwidth is decided based on the priority of the requesting traffic calls and the number of existing bandwidth adaptive calls. This prioritization of traffic classes does not reduce the bandwidth utilization. The scheme reduced the forced call termination probability significantly. The proposed scheme is modelled using the Markov Chain and the results show that the proposed scheme can provide negligible handover call dropping probability as well as significantly reduced new call blocking probability of higher priority calls without increasing the overall forced call termination probability.

The discussion here recognizes the importance of designing a framework that considers the manner of using network slices in a network architecture incorporating SDNs, and NFV capabilities. In addition, the network recognizes the concept and identity of the network slice. The manner of utilization of network slices is important for SDN and NFV capable networks and should be considered for different network systems. This is because the network architecture definition can differ for different contexts where SDN and NFV play an important role. This is necessary for a generic consideration for the preferences of different network service providers.

The preferences of different network service providers are recognized by Sharmah et al,[31] where insufficient bandwidth leads to degradation of subscriber quality of service. In this case, the acquisition, of additional bandwidth is deemed necessary. However, the bandwidth acquisition costs has not been considered. This is an important concern for licensed wireless networks that incorporate SDN and NFV. The discussion in [31] recognizes the need to design bandwidth management and resource allocation. The role of resource of allocation is recognized to be important in [31]. However, this is done in the context of traffic classes. These traffic classes arise due to the variability in subscriber demand for access to networks and networked applications.

Mamman et al. [32] proposed the design of a bandwidth management and resource allocation mechanism using the Markov chain approach. The proposed mechanism and associated analysis demonstrate the dropping of calls due to handover execution for mobile users is reduced. The presented research and its performance results demonstrate the usefulness and benefits of using Markov Analysis methods in resource allocation mechanisms aboard future wireless networks.

2.9 Markov chain

Continuous Markov Chain models are used to analyse and investigate the dynamics of bandwidth management in wireless networks. Holger Hermanns et al. [33] indicates that Markov chains are used in the context of performance and reliability evaluation of systems in various nature. They are widely used in simple yet adequate models for diverse areas, including mathematics, computer science, operations research, industrial engineering, biology, and demographics. Markov chains estimate performance characteristics for example to quantify throughput of manufacturing systems, locate bottlenecks in communication systems, or to estimate reliability in aerospace systems.

According to Lundteigen et al. [34] CTMC in easy terms are:

i. Probability of being in a specific state at a future time *t* only depends ONLY on the state of the system right now (s) state and NOT at all about the states the system has had before.
ii. Memory lessness process where the next state of the process depends only on the previous state and not the sequence of the states.
iii. The next move is purely unpredictable.

Holger Hermanns et al. [33] state that, Markov chains are memoryless with discrete time setting that are reflected by probabilistic decisions independent on the outcome of decisions taken earlier but only the current occupied state decisive to determine the probability of next transitions.
In continuous-time Markov chains models time ranges over positive reals instead of discrete subsets hence the memoryless property implies that probabilities of taking next transitions do not depend on the amount of time spent in the current state. The three ways to solve Markov equations are evaluating the time dependant probabilities, calculating steady state probabilities and the mean time to first failure.

2.10 Theoretical study: Markov Chain and Network Slicing

Continuous-time Markov chain (CTMC) is a mathematical model that describes the behaviour of a system as a sequence of random transitions between different states over time. It can be used to model the performance, and throughput of a network slice, as the probability of different traffic patterns or the expected users in a particular slice. This model can be used to optimize the allocation of resources to ensure that the network is functioning effectively [35].

Sohaib et al. [36] explains that the performance of the network can be evaluated in terms of latency, data rate and packet loss rate, it can affect the quality-of-service requirements for the different slices, the CTMC model can be used to analyse the impact of data prioritization on the delay and throughput of different slices, it can be used to study the number of slices and resources to be allocated to each slice, considering the different traffic priorities and patterns As discussed in [36] the use of the Markov chain is a scenario specific and targets a smart port environment. Furthermore, Markov chain is used to proactively estimate database station handover occurrences. This is done to reduce the signalling overhead and latency associated with handover execution. The focus is not the case of network slice management for a cloud-integrated network scenario. The research described in [36] focuses on the development of a new network architecture and the subsequent analysis. This target is the reduction of the number of mobile nodes that are highly active but with low utilization. A user's primary allocation and use of network slices have not been considered. Instead of subscribers, the network slices are allocated to service function chains. In addition, call quality is not analysed.

Furthermore, Geetha et al. [37] observe that different traffic classes have a varying bandwidth demand depending on the traffic load, based on the quality-of-service requirements the traffic classes are characterized into higher and lower classes. Bandwidth prioritization is the technique of allocating more bandwidth to a certain network use. It allows the network operator to assign different levels of priority to different types of network traffic. This can be done to ensure that important traffic such as video conferencing is given priority over less important traffic such as email or file transfer. The model can be used to ensure that the available resources are used effectively for example high-class users and applications that require sufficient bandwidth to function without being impacted by congestion, throttled, or delayed.

The discussion, in Geetha et al. [37] presents a queue for the case of a single base station and multiple subscribers. Relationship between multi-class calls and resource allocation has not received significant attention as the role of the cloud in slice-based networks is yet to be considered.

Vincenzi et al. [38] note that the service demands for 5G are stringent in terms of, sub-millisecond latencies for delay-critical services, a-100-fold capacity increase to service the needs of new applications and quality of service policy control for reliable communications.5G is considered as the technology for the three main service types enhanced mobile broadband that require high through-put and mobility demands, ultra-low latency and high reliability in terms of delays and reliability and machine-to-machine communication that require low data rates for IoT deployments. The model in [38] provided a slicing mechanism with an adaptive timescale with respect to network congestion. This approach provided, the limitation of the overall requirements without a significant loss in performance, and reduction in congestion adapting the admission strategy to comply with traffic fluctuations and performance compared to an on-demand method in a cost-effective manner. This model provided the slice provision to services with strict time constraints while maintaining excellent revenue for the network operator.

Yarkina et al. [39] proposed a scheme that focuses on a flexible performance where the isolation of slices customised to reflect both QoS and SLA terms. To evaluate the performance, a CTMC model was developed with a multi class service system with state dependent priorities. In this model three slices are active at the base station and re-slicing is performed as and when required (reallocation is when a user connection is established or terminated. The proposed simulated result is show in Figure 2.6. The network performance is improved compared to static slicing in terms of blocking probability. A higher blocking probability in slice 1 compared to slice 2 under static slicing or complete sharing due to a minimum data rate in slice 1.

The presented approach in [39], is only applicable to a 5G network only. Operator preferences leading to a deviation from the presented network model and system architecture have not been considered. This is especially important as the network architecture of 5G is yet to be standardized. The redesign of a new approach to ensure network operator-friendly resource allocation requires further research attention and consideration.

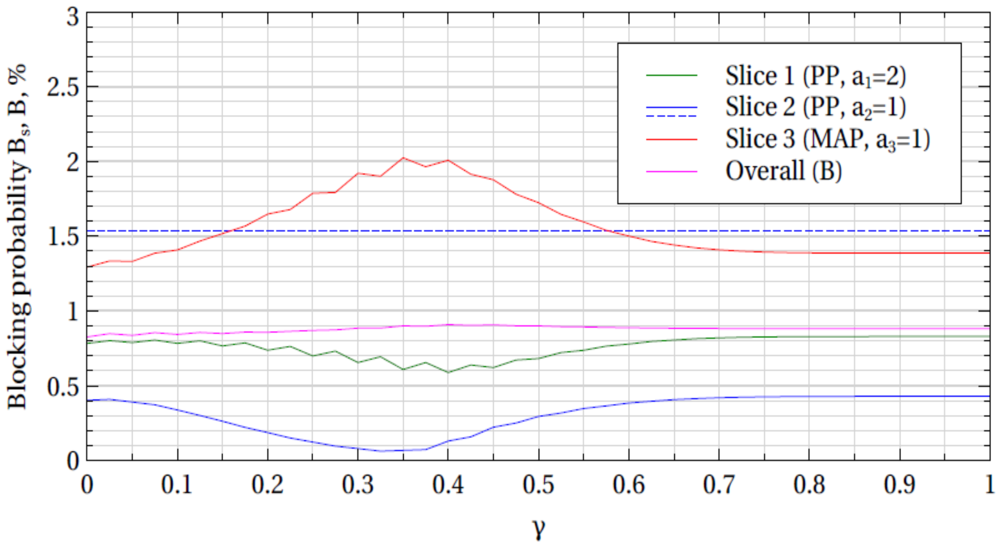


Figure 2.6: The blocking probabilities vs the booked capacity [39].

In Huynh et al. [40] the advantage of using network slicing is to enable service providers to offer network services on an as-a service basis that improves operational efficiency and hence reducing time-to-market for new services. In [40] it states that a dynamic network resource management model based on semi-Markov decision process will allow the network provider to allocate computing storage and radio resources to different slice request in a real-time manner and achieve a long-term reward with the available resources. In Figure 2.7 a network slicing model is considered with the three key parties. A network provider is the owner of the network infrastructure that provides resource slices including radio computing and storage to the tenants. Tenants request and lease resource slices to meet the service demands of their subscribers. The end users operate their functions and applications on the slices of the tenants [40].

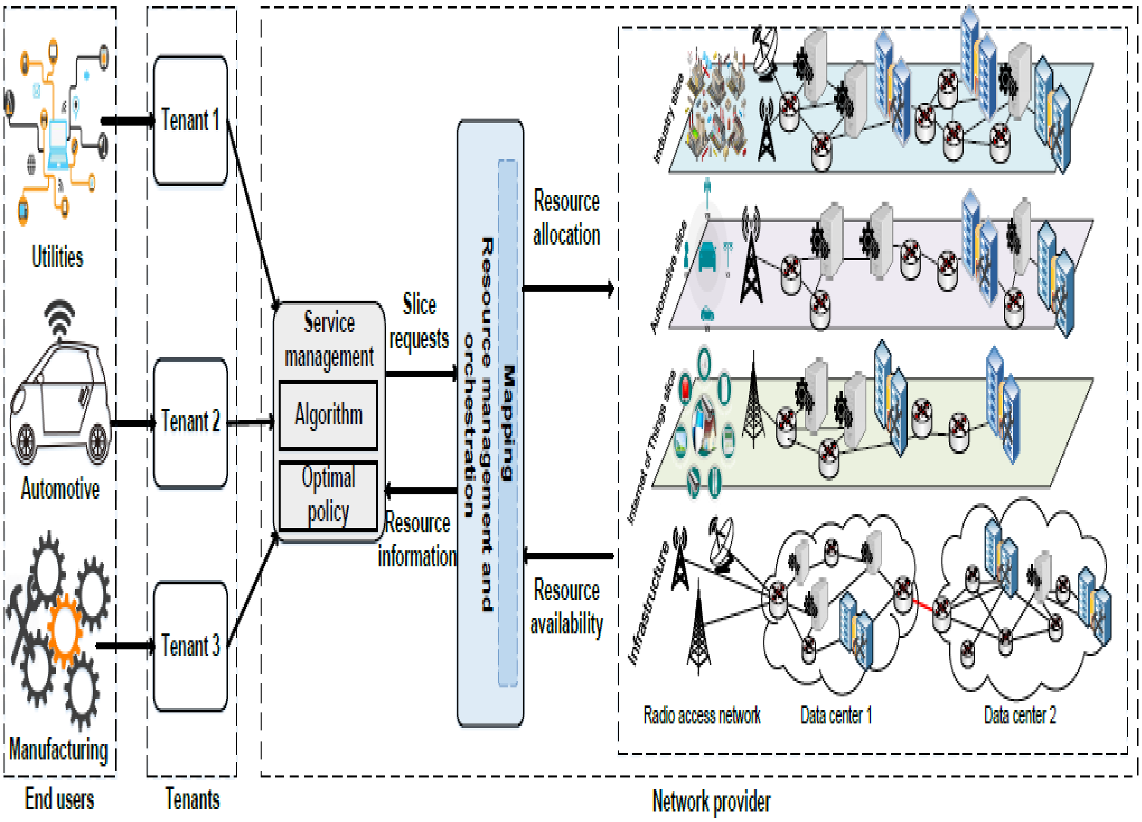


Figure 2.7: Network resource slicing system model [40].

In this model [40], three different classes of services are considered i.e., utilities, automotive and manufacturing. Each class has a specific demand and requirement, for example a vehicle will need ultra-reliable slice for telemetry driving. For the slices requested from industry, security, resilience, and reliability these are higher priority services. The tenant sends a slice request to the network provider that will specify the resources and additional service requirements hence these tenants will pay different prices for their requests. The network management component in Figure 2:7 will evaluate the requirements and decide to accept or reject the request based on its policy and prioritization class [40].

According to Han in [41] network slicing is an essential feature that enables 5G communications networks. Mobile operators can utilize physical and virtual network resources, network infrastructure, and the capacity of virtualized network functions, that provide scalability, flexibility, accountability, shareability and profitability to mobile networks.

Bakri in [42] it states that the emerging challenge of network slicing is to efficiently allocate network resources over different slices. The discussion in [41] focuses on state models of slice admission systems that operate in a synchronous method. The mobile network operator's decision to tenant requests periodically instead of immediately upon request arrivals.

Bakri in [42] observes that mobile networks have seen strong growth with the emergence of a new generation. As demand significantly increases quality of service becomes more critical. In the current "one size fits all" of architectures cannot brace these next-generation service demands. Research around 5G aims to provide adequate architectures and mechanisms.5G architecture is envisioned to provide the diverse and conflicting demands of services in terms of latency, bandwidth, and reliability which cannot be sustained by the same network infrastructure. Network slicing is provided by network virtualization that allows the infrastructure to be divided into different slices each slice is tailored to meet specific service criteria.

According to Bakri in [42], the ultimate key is to improve the network performance by introducing flexibility and a greater utilization of network resources to meet the diverse customer requirements.

The proposed Markov decision process model in [42] for network slicing with the following connotations tuples *M* = ***(S; A; T; R****)* are:

S set of states

A set of actions

T is the transition probability from state '*s'* at time '*t* 'to state *s'* at time *t + 1* when taking an action, *a* and ***R*** is the reward obtained by performing the action *a*, which leads to move from the state *s* to *s'* for the proposed system if a state *s = (n; m; l; b)* is composed of the following information where:

i. *n* is the number of accepted enhanced mobile broadband slices.

ii. *m* is the number of accepted ultra-reliable low latency communication slices.

iii. *l* is the number of accepted massive machine type communication slices.

iv. *b* is a value that can be equal to 1,2 or 3 to indicate the slice type.

a new slice request, via an agent observes the state of the system and takes an action.
To accept or reject a request, the action set is as follows:
a=1 if new arrival slice request is accepted
a=0 if new arrival slice request is rejected

Different transitions of the system occur when a new network slice arrives, and a decision is needed, if the system is in a state s = (n; m; l; b) and a new slice arrives, a decision needs to be taken (accept or reject) leading that the system transits to one of the following states:

i. *(n + 1; m; l; 1)* if a slice of eMBB is accepted.
ii. *(n; m + 1; l; 2)* if a slice of uRLLC is accepted.
iii. *(n; m; l + 1; 3)* if a slice of mMTC is accepted.
iv. *(n; m; l; 1)* if a slice of eMBB is rejected.
v. *(n; m; l; 2)* if a slice of uRLLC is rejected.
vi. *(n; m; l; 3)* if a slice of mMTC is rejected.

If a network slice leaves without taking any actions, the system moves into one of the following states:

i. *(n -1; m, l; 1)* if a slice of eMBB has left.
ii. *(n; m -1; l; 2)* if a slice of uRLLC has left.
iii. *(n; m; l - 1; 3)* if a slice of mMTC has left.

2.11 Bandwidth Management and prioritization

According to Mallapur et al. [43] the mobile industry is one of the rapidly growing areas in today’s communication technology. Multimedia applications, such as audio phone, video on demand, video conferences, video games and social media applications has resulted in spectacular strides in the progress of wireless communication systems. Data needs to be transmitted continuously demanding larger bandwidth. Since bandwidth is the critical resource in wireless networks it is necessary for efficient utilization. The goal of wireless communications is to provide reliable connectivity at any place and at any time.

Sarmah et al. [44] notes the advantage and enhancement of IEEE 802.11 wireless LAN that has become a popular and demanding technology for all types of end users. A smooth congestion free traffic in any organization bandwidth is the main network resource. Correct management and allocation mechanism needed in every organization. Users get frustrated when they cannot perform their tasks properly due to lack of bandwidth. Purchasing extra bandwidth is expensive and beyond reach for most organizations therefore the limited available bandwidth must be managed. Bandwidth management is a technique through which resources are allocated to applications or users in an efficient and productive manner. If there is no management scheme in a network an application or users can acquire all the available bandwidth resources hence restricting other application and users from the network. Bandwidth management can be implemented by differentiating the network traffic into classes by application and service types, for each class different priority can be set, based on priority value and class that can access the network. A scheme was proposed where users are categorized into three main types, high priority, medium priority, and low priority. Based on priority of the users the admission controller will allow the users to access the resource. The admission controller will maintain a priority table of all the stations that will get connected to the access point. Bandwidth management is a technique whereby resources are allocated to applications or users in a prioritized manner.

2.12 Related work to this study

According to Sharma et al. [45] data is the transfer of e-mail messages and files, video, and voice communications. Networks that function on copper, fibre optic or wireless networks convey information between computer devices. Hence, wireless is the new technology of networking. People can enjoy the benefits where they live and work It is useful for people to communicate and access applications and information without wires, people can interact from a location that they prefer. Wireless networks are not bound to a channel to follow like wired networks. In addition, it is less expensive and much easier to install than wired networks.

According to David in [16] wireless communication is the most vibrant areas in the communication field today. The past decade has seen a surge of research activities in this area, due to a confluence of several factors which include explosive increase in demand for connectivity, driven so far mainly by cellular telephony but is expected to be soon eclipsed by wireless data applications, the dramatic progress in VLSI technology has enabled small area and low-power implementation of sophisticated signal processing algorithms and coding techniques and the success of second-generation digital wireless standards the IS-95 code division multiple access standard, provides a concrete demonstration that good ideas from communication theory can have a significant impact in practice.

Research in the above-mentioned sections presented a comprehensive knowledge of existing literature related to the various components and concepts, critical to our study that include theoretical alternatives in telecommunication systems, wireless propagation, traffic management, network performance, quality of service and resource management in mobile networks. This review sets the foundation to the study of Markov chain-based network slicing, bandwidth management and data prioritization strategies.

2.13 Summary

The research being presented in this study recognizes the need to design a resource allocation mechanism for network slices in future networks incorporating SDN and NFV. This is deemed important for networks that support multi–priority call classes. In this regard, the design of allocation mechanisms ensures that subscriber calls (in an individual or group consideration) achieve the desired level of service. In addition, it is recognized that existing research has analysed the design and performance of resource allocation mechanisms as seen in [37]. However, the case of different call priorities alongside the incorporation of an enhancement capability i.e., a buffer is required. The metrics of call blocking probability and call dropping probability are being considered. These metrics are evaluated in the case of existing mechanism described by the model in [37] and the proposed mechanism where a buffer is incorporated.

This research proposes the use of buffering for enhancing call quality in future networks incorporating network slicing where slices host the resources enabling data and voice execution. The contribution of this research is the design of a mechanism considering different service priorities in 5G networks comprising network slices. The proposal of a network buffering mechanism for resource access in the context of multi-subscriber priorities in 5G networks. The presented research develops an analytic model for evaluating the role of a buffer in a 5G network comprising network slices which also executes voice calls and data calls having varying resource usage preferences (i.e., priorities).

The role of a buffer is examined in this case and important metrics are formulated. These metrics are the call blocking probability and call dropping probability. The identified metrics are investigated in the absence (existing mechanism) and in the presence of a buffer (proposed mechanism). The call blocking probability and call dropping probability are evaluated considering different call arrival rates, service rates and buffering rates. Furthermore, numerical evaluation of the call blocking probability and call dropping probability is executed in the case of existing and proposed mechanisms.

Chapter three presents conceptual framework and simulation models related to the study.

Chapter 3 - Research Methodology and Simulations

3.1 Methodology

The discussion in this section presents the performance formulation and associated results. The performance formulation and evaluation aim to investigate the call blocking probability and call dropping probability. The call blocking probability (CBP) and call dropping probability (CDP) are important parameters for a communication system. In this case, the CBP and CDP are formulated for different categories of calls. The considered calls have varying priority levels. These priority levels are defined by the bandwidth resources allocated to each call. The calls have three priority levels. These levels are Class A calls (high priority and allocated the highest bandwidth), Class B calls (medium priority and allocated reasonably high bandwidth) and Class C calls (lowest priority and allocated lowest bandwidth resources). The investigation of the CBP and CDP is done via the use of the Markov chain that describes the transitions between call contexts (comprising the three calls in different states).

The Markov chain approach is used for the existing case (without buffer) and proposed case (with buffer). The rest of the discussion here is divided into three subsections. The first subsection presents the Markov chain developed for the case of the existing and proposed cases. The second subsection discusses the CBP and CDP in the existing and proposed cases. The third subsection describes the contexts and associated scenarios being considered in the evaluation of system performance and analysis.

3.2 Development and Formulation of the Markov Chain

The Markov chain presenting the transition is developed in this section. This is done for the cases of the existing solution and the proposed solution. The proposed solution differs from the existing solution in the use of buffers to enable call waiting when bandwidth resources are unavailable or insufficient. The developed and presented Markov chain makes use of the following transition variables:

i. λ_A and μ_A –These are the arrival rates and departure rates (service rate) of Class A (having the highest priority) calls into the considered network. From the perspective of network slicing, Class A calls have the highest number of reserved resources (bandwidth).

ii. λ_B and μ_B –These are the arrival rates and departure rates (service rate) of Class B (with medium priority) calls into the considered network. In the context of network slicing, Class B calls have the highest number of reserved network resources i.e., bandwidth after Class A calls.

iii. λ_C and μ_C –These are the arrival rates and departure rates (service rate) of Class C (having lowest priority) calls into the considered network. Class C calls have the least number of reserved bandwidth network resources and require resources lower than Class A and Class B call categories.

iv. λ_B' –This is the arrival rate of Class B calls into the buffer. The buffer is introduced in the proposed solution. The buffering of Class B calls becomes necessary when there are a significant number of Class A calls requiring access to bandwidth resources.

v. λ_C' – This is the arrival rate of Class C calls into the buffer. The buffer is introduced in the proposed solution. The buffering of Class C calls becomes necessary when there are a significant number of Class A and Class B calls requiring access to bandwidth resources.

In the development of the Markov chain, the transitions consider that only a class of calls i.e., either Class A call, Class B call or Class C call can either arrive or depart from the network system. Furthermore, cases where the increase in the number of calls is clearly indicated are states where the system capacity i.e., bandwidth resources (indicating the amount of network slices) have been fully utilized. The Markov transition for the case of the existing solution (without the use of buffering) is presented in Figure 3.1 [40],[44].

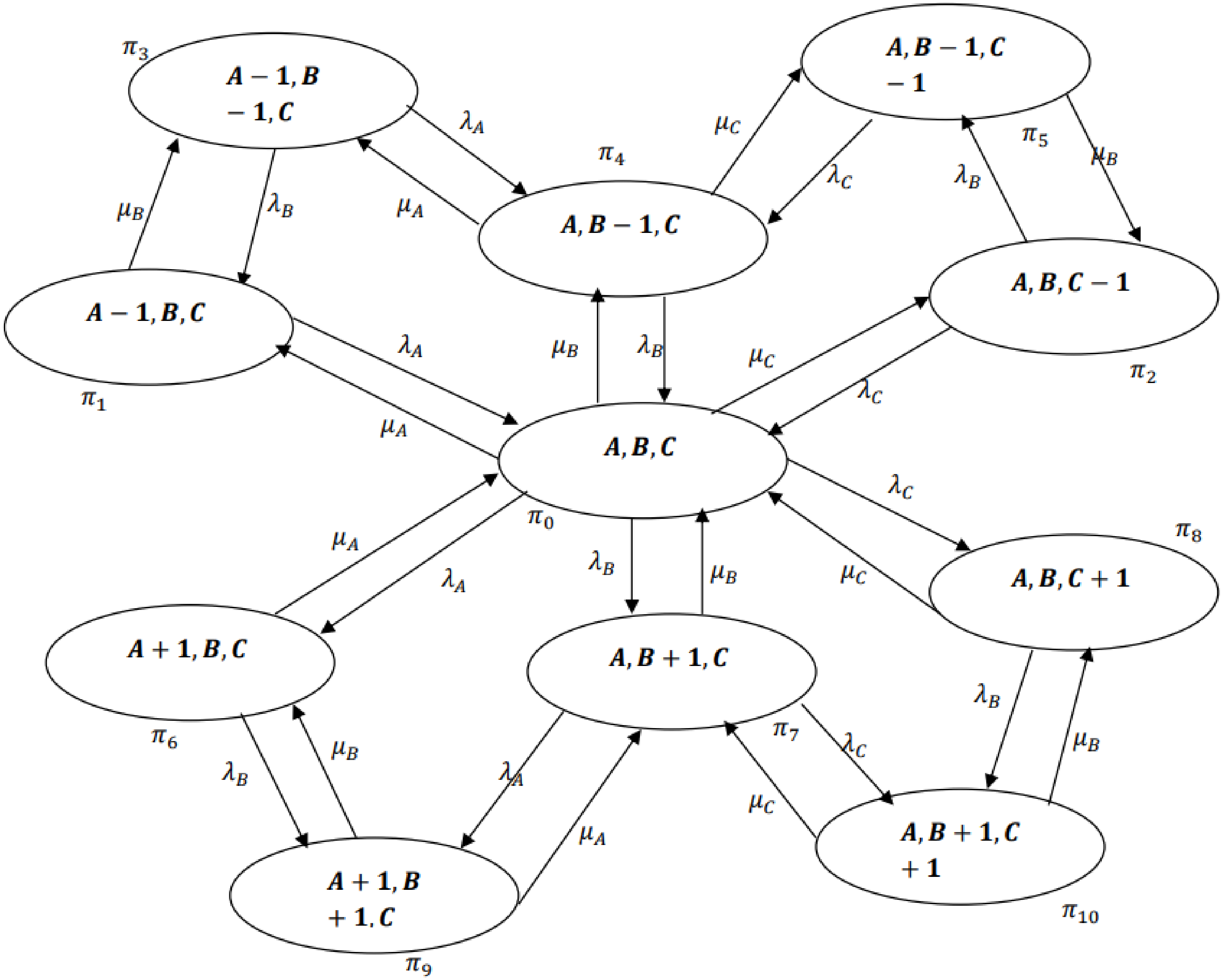


Figure 3.1: Markov Chain Describing the Case of the existing mechanism.

In the case of the existing solution presented in Figure 3.1 the arrival of more Class A calls triggers different transitions in the network. A case where Class A calls require more network slices (bandwidth resources) exceeding the initially provided number of resources is feasible. In such a case, a blocking or dropping of Class B calls and Class C calls results in the network. A similar scenario can arise in the case of Class B calls. This results in the dropping or blocking of Class C calls (having the lowest priority). It is important to design solutions that reduce the CBP and CDP arising from these contexts. This is the motivation for developing and presenting the Markov chain for the proposed scheme shown in Figure 3.2. The variables λ'_B and λ'_C are introduced to indicate the transfer of calls into the buffer. Buffered calls can continue the use of network slice i.e., resources after the execution of call classes with higher slice access priority.

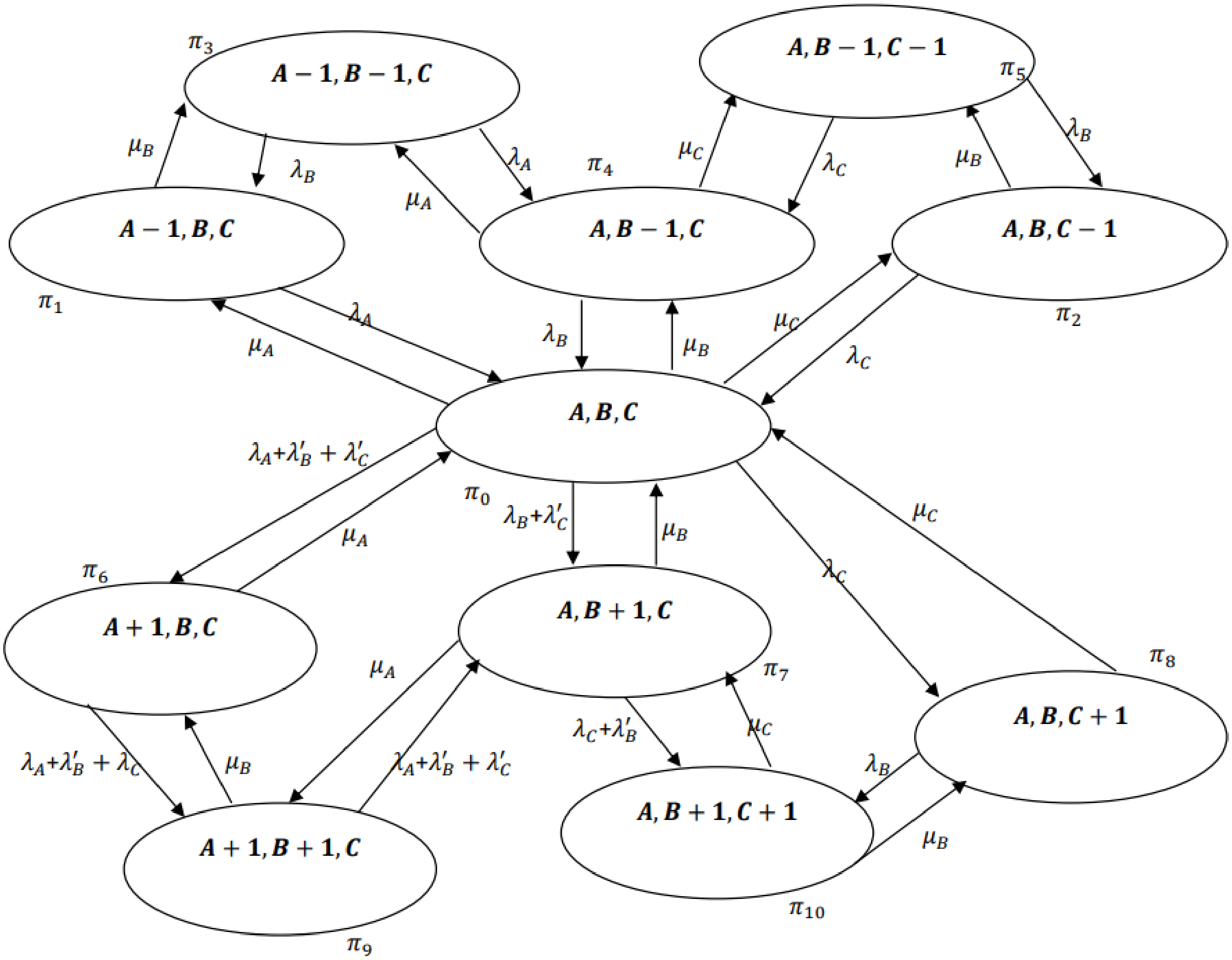


Figure 3.2: Markov State Diagram for the proposed case considering the role of the Buffer.

The state equations for the concerned Markov chain shown in Figures 3.1 and 3.2 are presented and solved. In the case of the existing solution Figure 3.1 the state equations are given as:

State Equations for the consideration of the existing case – In the case without Buffer

$$A_1\pi_0 - \lambda_C\pi_2 - \mu_C\pi_8 - \mu_B\pi_7 - \mu_A\pi_6 - \lambda_A\pi_1 - \lambda_B\pi_4 = 0\ ;\ A_1 = (\mu_A + \mu_B + \mu_C + \lambda_A + \lambda_B + \lambda_C) \quad (3.1)$$

$$A_2\pi_1 - \lambda_B\pi_3 - \mu_A\pi_0 = 0\ ; A_2 = (\mu_B + \lambda_A) \quad (3.2)$$

$$\pi_2 = A_{19}\pi_0 + A_{20}\pi_5\ ; A_{19} = \mu_C(\lambda_B + \lambda_C)^{-1}\ \ ; A_{20} = \mu_B(\lambda_B + \lambda_C)^{-1} \quad (3.3)$$

$$A_3\pi_3 - \mu_A\pi_4 - \mu_B\pi_1 = 0\ ; A_3 = (\lambda_B + \lambda_A) \quad (3.4)$$

$$A_4\pi_4 - \mu_B\pi_0 - \lambda_A\pi_3 = 0\ ; A_4 = (\lambda_B + \mu_A + \mu_C) \quad (3.5)$$

$$A_5\pi_5 - \lambda_B\pi_2 - \mu_C\pi_4 = 0\ ; A_5 = (\lambda_C + \mu_B) \quad (3.6)$$

$$A_6\pi_6 - \lambda_A\pi_0 - \mu_B\pi_9 = 0\ ; A_6 = (\mu_A + \mu_B) \quad (3.7)$$

$$A_7\pi_7 - \lambda_B\pi_0 - \mu_A\pi_9 = 0\ ; A_7 = (\mu_B + \lambda_A + \lambda_C) \quad (3.8)$$

$$A_8\pi_8 - \lambda_C\pi_0 - \mu_B\pi_{10} = 0\ ; A_8 = (\mu_C + \lambda_B) \quad (3.9)$$

$$A_9\pi_9 - \lambda_B\pi_6 - \lambda_A\pi_7 = 0\ ; A_9 = (\mu_A + \mu_B) \quad (3.10)$$

$$A_{10}\pi_{10} - \lambda_C\pi_7 - \lambda_B\pi_8 = 0\ ; A_{10} = (\mu_B + \mu_C) \quad (3.11)$$

$$\pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 + \pi_6 + \pi_7 + \pi_8 + \pi_9 + \pi_{10} = 1 \quad (3.12)$$

The system of equations in (3.1) – (3.12) is solved via multiple steps of substitution and simplification to yield the solutions of the concerned state probability variables.

The presented system of equations in (3.1) – (3.12) is solved analytically to obtain an algebraic closed form expression for each state probability. The algebraic expression for each state probability can be used to further evaluate the probability associated with a certain context in the network.

The presented system of equations is solved analytically in a manner to obtain a closed form expression for each state probability the derived metrics. A graphical or tabular based solution arising from a numerical procedure though recognized to be feasible has not been considered. This is because such an approach requires making simplifying assumptions and negates the challenge of ensuring the design of a multi-scenario in the concerned network. Therefore, the use of such an approach extends into imposing different conditions on the architecture and capability of the multi-subscriber and multi-priority call network. This limits the network operation contexts capable of being capable in the presented solution and the corresponding analysis. Moreover, the concerned network is deemed to have sufficient network slices i.e., bandwidth resources to support the concerned state transitions.

Hence, the system of equations in (3.1) – (3.12) is solved in an analytical manner**.**

The following variables are used in the simplification steps. The variables used in the simplification are presented in Table 3.1.

Table 3.1: Defined variables associated the solution of the system of equations (no buffer case)

<table>
<tr><td>$A_{11} = \frac{\lambda_A}{A_1}$</td><td>$A_{12} = \frac{\lambda_C}{A_1}$</td><td>$A_{13} = \frac{\lambda_B}{A_1}$</td><td>$A_{14} = \frac{\mu_A}{A_1}$</td><td>$A_{15} = \frac{\mu_B}{A_1}$</td><td>$A_{16} = \frac{\mu_C}{A_1}$</td><td>$A_{17} = \frac{\mu_A}{A_2}$</td><td>$A_{18} = \frac{\lambda_B}{A_2}$</td><td>$A_{19} = \frac{\mu_C}{\lambda_B + \lambda_C}$</td></tr>
<tr><td>$A_{21} = \frac{\mu_B}{A_3}$</td><td>$A_{22} = \frac{\mu_A}{A_3}$</td><td>$A_{23} = \frac{\mu_B}{A_4}$</td><td>$A_{24} = \frac{\lambda_A}{A_4}$</td><td>$A_{25} = \frac{\lambda_C}{A_4}$</td><td>$A_{26} = \frac{\lambda_B}{A_5}$</td><td>$A_{27} = \frac{\mu_C}{A_5}$</td><td>$A_{28} = \frac{\mu_B}{A_6}$</td><td>$A_{20} = \frac{\mu_B}{\lambda_B + \lambda_C}$</td></tr>
<tr><td>$A_{31} = \frac{\mu_A}{A_7}$</td><td>$A_{32} = \frac{\mu_C}{A_7}$</td><td>$A_{33} = \frac{\lambda_B}{A_9}$</td><td>$A_{34} = \frac{\lambda_A}{A_9}$</td><td>$A_{35} = \frac{\lambda_C}{(\mu_B + \mu_C)}$</td><td>$A_{36} = \frac{\lambda_B}{(\mu_B + \mu_C)}$</td><td></td><td></td><td>$A_{29} = \frac{\lambda_A}{A_6}$</td></tr>
<tr><td>$A'_{31} = \frac{\lambda_C}{A_8}$</td><td>$A'_{32} = \frac{\mu_B}{A_8}$</td><td></td><td></td><td></td><td></td><td></td><td></td><td>$A_{30} = \frac{\lambda_B}{A_7}$</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="3">$B_1 = \frac{A_{19}(1 - A_{18}A_{21})}{A_{18}A_{19}A_{22} - A_{17}A_{20}A_{27}}$</td><td colspan="3">$B_2 = \frac{A_{17}(1 - A_{20}A_{26})}{A_{18}A_{19}A_{22} - A_{17}A_{20}A_{27}}$</td><td colspan="3">$B_3 = \frac{A_{22}A_{23}}{1 - A_{22}A_{24}}$</td></tr>
<tr><td colspan="3">$B_4 = \frac{A_{21} + A_{22}A_{25}A_{27}B_1}{1 - A_{22}A_{24}}$</td><td colspan="3">$B_5 = \frac{A_{22}A_{25}(A_{26} - A_{27}B_2)}{1 - A_{22}A_{24}}$</td><td colspan="3">$B_6 = \frac{A_{17} + A_{18}B_3}{1 - A_{18}B_4}$</td></tr>
<tr><td colspan="3">$B_7 = \frac{A_{18}B_5}{1 - A_{18}B_4}$</td><td colspan="3">$B_8 = \frac{A_{19}}{1 - A_{20}A_{26} + A_{20}A_{27}B_2}$</td><td colspan="3">$B_9 = \frac{A_{20}A_{27}B_1}{1 - A_{20}A_{26} + A_{20}A_{27}B_2}$</td></tr>
<tr><td colspan="3">$B_{10} = A_{27}B_1$</td><td colspan="3">$B_{11} = A_{26} - A_{27}B_2$</td><td colspan="3"></td></tr>
<tr><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td></tr>
<tr><td colspan="3">$C_1 = \frac{B_6 + B_7B_8}{1 - B_7B_9}$</td><td colspan="3">$C_2 = B_8 + B_9C_1$</td><td colspan="3">$C_3 = B_3 + B_4C_1 + B_5C_2$</td></tr>
<tr><td colspan="3">$C_4 = B_1C_1 - B_2C_2$</td><td colspan="3">$C_5 = B_{10}C_1 + B_{11}C_2$</td><td colspan="3"></td></tr>
<tr><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td></tr>
<tr><td colspan="3">$D_0 = \frac{A'_{31}A_{36}}{1 - A'_{32}}$</td><td colspan="3">$D_1 = \frac{A_{35}}{1 - A'_{32}}$</td><td colspan="3">$D_2 = \frac{A_{29}D_0}{1 - A_{28}A_{33}}$</td></tr>
<tr><td colspan="3">$D_3 = \frac{A_{28}A_{34} + A_{29}D_1}{1 - A_{28}A_{33}}$</td><td colspan="3">$D_4 = \frac{A_{33}D_2 + A_{34}D_2}{1}$</td><td colspan="3">$D_5 = \frac{A_{33}D_3 + A_{34}D_3}{1}$</td></tr>
<tr><td colspan="3">$D_6 = A_{31} + A_{32}D_0$</td><td colspan="3">$D_7 = A_{32}D_1$</td><td colspan="3">$D_8 = \frac{A_{30} + A_{31}D_4 + A_{32}D_0}{1 - A_{31}D_5 - A_{32}D_1}$</td></tr>
</table>

The solutions to the state variables (state probability values) are given as:

$$\pi_0 = \left(\left(\sum_{n=1}^{5} C_n \right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7)) \right)^{-1} \qquad (3.13)$$

$$\pi_1 = C_1 \left(\left(\sum_{n=1}^{5} C_n \right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7)) \right)^{-1} \qquad (3.14)$$

$$\pi_2 = C_2\left(\left(\sum_{n=1}^{5} C_n\right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7))\right)^{-1} \quad (3.15)$$

$$\pi_3 = C_3\left(\left(\sum_{n=1}^{5} C_n\right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7))\right)^{-1} \quad (3.16)$$

$$\pi_4 = C_4\left(\left(\sum_{n=1}^{5} C_n\right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7))\right)^{-1} \quad (3.17)$$

$$\pi_5 = C_5\left(\left(\sum_{n=1}^{5} C_n\right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7))\right)^{-1} \quad (3.18)$$

$$\pi_6 = (D_2 + D_3D_8)\left(\left(\sum_{n=1}^{5} C_n\right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7))\right)^{-1} \quad (3.19)$$

$$\pi_7 = (D_8)\left(\left(\sum_{n=1}^{5} C_n\right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7))\right)^{-1} \quad (3.20)$$

$$\pi_8 = (D_6 + D_7D_8)\left(\left(\sum_{n=1}^{5} C_n\right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7))\right)^{-1} \quad (3.21)$$

$$\pi_9 = (D_4 + D_5D_8)\left(\left(\sum_{n=1}^{5} C_n\right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7))\right)^{-1} \quad (3.22)$$

$$\pi_{10} = (D_0 + D_1D_8)\left(\left(\sum_{n=1}^{5} C_n\right) + (D_0 + D_2 + D_4 + D_6 + D_8 + D_8(D_1 + D_3 + D_5 + D_7))\right)^{-1} \quad (3.23)$$

In the case of the proposed solution (considering the use of a buffer) i.e., Figure 3.2, the state equations are given as:

$$\pi_0 = A_1\pi_1 + A_2\pi_2 + A_3\pi_4 + A_4\pi_6 + A_5\pi_7 + A_6\pi_8 \quad (3.24)$$

$$\pi_1 = A_7\pi_1 + A_8\pi_3 \quad (3.25)$$

$$\pi_2 = A_9\pi_0 + A_{10}\pi_5 \quad (3.26)$$

$$\pi_3 = A_{11}\pi_1 + A_{12}\pi_4 \quad (3.27)$$

$$\pi_4 = A_{13}\pi_0 + A_{14}\pi_3 + A_{15}\pi_5 \quad (3.28)$$

$$\pi_5 = A_{16}\pi_2 + A_{17}\pi_4 \quad (3.29)$$

$$\pi_6 = A_{18}\pi_0 + A_{19}\pi_9 \quad (3.30)$$

$$\pi_7 = A_{20}\pi_0 + A_{21}\pi_9 + A_{22}\pi_{10} \quad (3.31)$$

$$\pi_8 = A_{23}\pi_0 + A_{24}\pi_{10} \quad (3.32)$$

$$\pi_9 = A_{25}\pi_6 + A_{26}\pi_7 \quad (3.33)$$

$$\pi_{10} = A_{27}\pi_7 + A_{28}\pi_8 \quad (3.34)$$

$$\pi_0 + \pi_1 + \pi_2 + \pi_3 + \pi_4 + \pi_5 + \pi_6 + \pi_7 + \pi_8 + \pi_9 + \pi_{10} = 1 \quad (3.35)$$

In this case, the concerned simplifying variables are presented in Table 3.2 (proposed case using the buffer)

Table 3.2: Defined variables associated the solution of the system of equations (buffer case)

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="4">$A_0 = \mu_B + \mu_C + \lambda_A + \lambda_B + \lambda_C + \lambda'_B + 2\lambda'_C$</td><td>$A_1 = \frac{\lambda_A}{A_0}$</td><td>$A_2 = \frac{\lambda_C}{A_0}$</td><td>$A_3 = \frac{\lambda_B}{A_0}$</td><td>$A_4 = \frac{\mu_A}{A_0}$</td><td>$A_5 = \frac{\mu_B}{A_0}$</td></tr>
<tr><td>$A_6 = \frac{\mu_C}{A_0}$</td><td colspan="2">$A_7 = \frac{\mu_A}{\lambda_A + \mu_B}$</td><td colspan="2">$A_8 = \frac{\lambda_B}{\lambda_A + \mu_B}$</td><td colspan="2">$A_9 = \frac{\mu_C}{\lambda_C + \mu_B}$</td><td colspan="2">$A_{10} = \frac{\lambda_B}{\lambda_C + \mu_B}$</td></tr>
<tr><td colspan="2">$A_{11} = \frac{\mu_B}{\lambda_A + \lambda_B}$</td><td colspan="2">$A_{12} = \frac{\mu_A}{\lambda_A + \lambda_B}$</td><td colspan="2">$A_{13} = \frac{\mu_B}{\lambda_B + \mu_A + \mu_C}$</td><td colspan="2">$A_{14} = \frac{\lambda_A}{\lambda_B + \mu_A + \mu_C}$</td><td>$A_{16} = \frac{\mu_B}{\lambda_B + \lambda_C}$</td></tr>
<tr><td colspan="2">$A_{15} = \frac{\lambda_C}{\lambda_B + \mu_A + \mu_C}$</td><td colspan="2">$A_{17} = \frac{\mu_C}{\lambda_B + \lambda_C}$</td><td colspan="2">$A_{18} = \frac{\lambda_A + \lambda'_B + \lambda'_C}{\mu_A + \lambda_A + \lambda'_B + \lambda_C}$</td><td colspan="3">$A_{19} = \frac{\mu_B}{\mu_A + \lambda_A + \lambda'_B + \lambda_C}$</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="3">$A_{20} = \frac{\lambda_B + \lambda'_C}{\mu_A + \mu_B + \lambda_C + \lambda'_B}$</td><td colspan="3">$A_{21} = \frac{\lambda_A + \lambda'_B + \lambda'_C}{\mu_A + \lambda_A + \lambda'_B + \lambda_C}$</td><td colspan="3">$A_{22} = \frac{\mu_C}{\mu_A + \lambda_A + \lambda'_B + \lambda_C}$</td></tr>
<tr><td colspan="3">$A_{23} = \frac{\lambda_C}{\lambda_B + \mu_C}$</td><td colspan="3">$A_{24} = \frac{\mu_B}{\lambda_B + \mu_C}$</td><td colspan="3">$A_{25} = \frac{\lambda_A + \lambda'_B + \lambda_C}{\mu_B + \lambda_A + \lambda'_B + \lambda'_C}$</td></tr>
<tr><td colspan="3">$A_{26} = \frac{\mu_A}{\mu_B + \lambda_A + \lambda'_B + \lambda'_C}$</td><td colspan="3">$A_{27} = \frac{\lambda'_B + \lambda_C}{\mu_B + \mu_C}$</td><td colspan="3">$A_{28} = \frac{\lambda_B}{\mu_B + \mu_C}$</td></tr>
<tr><td colspan="3"></td><td colspan="3"></td><td colspan="3"></td></tr>
<tr><td colspan="3">$B_1 = A_8A_{12}(1 - (A_3 + A_8A_{11}))^{-1}$</td><td colspan="3">$B_2 = A_9(1 - A_{10}A_{16})^{-1}$</td><td colspan="3">$B_3 = A_{10}A_{17}(1 - A_{10}A_{16})^{-1}$</td></tr>
<tr><td colspan="3">$B_4 = A_{11}B_1 + A_{12}$</td><td colspan="3">$B_5 = \frac{A_{13} + A_{15}A_{16}B_2}{(1 - A_{15}A_{16}B_3 - A_{15}A_{17} - B_4)}$</td><td colspan="3">$B_6 = A_{17}B_5 + A_{16}B_2 + A_{16}B_3B_5$</td></tr>
<tr><td colspan="3">$C_1 = \frac{A_{18}}{1 - A_{19}A_{25}}$</td><td colspan="3">$C_2 = \frac{A_{19}A_{26}}{1 - A_{19}A_{25}}$</td><td colspan="3">$C_3 = \frac{A_{23}A_{28}}{1 - A_{24}A_{28}}$</td></tr>
<tr><td colspan="3">$C_4 = \frac{A_{27}}{1 - A_{24}A_{28}}$</td><td colspan="3">$C_5 = A_{25}C_1$</td><td colspan="3">$C_6 = A_{25}C_2 + A_{26}$</td></tr>
<tr><td colspan="3">$C_7 = A_{24}C_3 + A_{23}$</td><td colspan="3">$C_8 = A_{24}C_4$</td><td colspan="3">$E_0 = \frac{(A_{20} + A_{21}C_5 + A_{22}C_3)}{1 - (A_{21}C_6 + A_{22}C_4)}$</td></tr>
</table>

The solutions to the state variables (state probability values) are given as:

$$\pi_0 = (B_2 + B_5 + B_6 + B_5(B_1 + B_3 + B_4) + C_1 + C_3 + C_5 + C_7 + E_0 + E_0(C_2 + C_4 + C_6 + C_8))^{-1} \quad (3.36)$$

$$\pi_1 = B_1B_5(B_2 + B_5 + B_6 + B_5(B_1 + B_3 + B_4) + C_1 + C_3 + C_5 + C_7 + E_0 + E_0(C_2 + C_4 + C_6 + C_8))^{-1} \quad (3.37)$$

$$\pi_2 = (B_2 + B_3B_5)\pi_0 \quad (3.38)$$

$$\pi_3 = B_4B_5(B_2 + B_5 + B_6 + B_5(B_1 + B_3 + B_4) + C_1 + C_3 + C_5 + C_7 + E_0 + E_0(C_2 + C_4 + C_6 + C_8))^{-1} \quad (3.39)$$

$$\pi_4 = B_5(B_2 + B_5 + B_6 + B_5(B_1 + B_3 + B_4) + C_1 + C_3 + C_5 + C_7 + E_0 + E_0(C_2 + C_4 + C_6 + C_8))^{-1} \quad (3.40)$$

$$\pi_5 = B_6(B_2 + B_5 + B_6 + B_5(B_1 + B_3 + B_4) + C_1 + C_3 + C_5 + C_7 + E_0 + E_0(C_2 + C_4 + C_6 + C_8))^{-1} \quad (3.41)$$

$$\pi_6 = (C_1 + C_2E_0)\pi_0 \quad (3.42)$$

$$\pi_7 = (E_0)\pi_0 \quad (3.43)$$

$$\pi_8 = (C_7 + C_8E_0)\pi_0 \quad (3.44)$$

$$\pi_9 = (C_5 + C_6E_0)\pi_0 \quad (3.45)$$

$$\pi_{10} = (C_3 + C_4E_0)\pi_0 \quad (3.46)$$

3.3 Derivation of the CBP and CDP Performance Metrics Existing Case

In the existing case, a call blocking event affects Class B and Class C calls but not Class A calls. This is because Class A calls have the highest priority. The occurrence of call dropping events also affects Class B and Class C calls but not Class A calls. The occurrence of calling block events is given by the events described by the transition from state π_3 to π_4 OR π_1 to π_0 OR π_6 to π_0 OR π_6 to π_9 OR π_7 to π_9. In addition, the occurrence of call blocking is also described by being in either of the states π_6 ,π_9 or π_{10}. The call blocking probability in the case of the existing solution i.e., without the use of buffer for Class B or Class C calls is denoted CBP_1 and given as:

$$CBP_1 = \pi_3\pi_4 + \pi_1\pi_0 + \pi_6\,\pi_9 + \pi_7\,\pi_9 + \pi_6 + \pi_9 + \pi_{10} \quad (3.47)$$

In the case of the existing solution i.e., without buffer, the occurrence of call dropping event affects Class B and Class C calls only. Call dropping event occurs for the transition from state π_0 to π_7 OR π_4 to π_0 OR π_2 to π_5. In addition, the occurrence of call dropping is also described by being in either of the states π_7 ,π_8 or π_{10}. The call dropping probability in the case of the existing solution i.e., without the use of buffer for Class B or Class C calls is denoted CDP_1 and given as:

$$CDP_1 = \pi_0\pi_7 + \pi_4\pi_0 + \pi_2\pi_5 + \pi_7 + \pi_8 + \pi_{10} \tag{3.48}$$

For the case of the proposed solution i.e., including the use of the buffer for Class B and Class C calls, the call blocking probability and call dropping probability are denoted CBP_2 and CDP_2, respectively. In deriving the call blocking probability, the role of the buffer prevents the call blocking transitions described in π_6 to π_9 OR π_7 to π_9. Similarly, the role of the buffer prevents the calls dropping transitions. This is achieved by considering the transition associated with the states where call dropping occurs. The concerned states where the role of the buffer is executed are π_7 ,π_8 or π_{10}. However, the role of the buffer is not included in the states π_1 ,π_2 ,π_3,π_4 or π_5. The expression for the parameters CBP_2 and CDP_2 is given as:

$$CBP_2 = \pi_3\pi_4 + \pi_1\pi_0 \tag{3.49}$$

$$CDP_2 = \pi_1\pi_3 + \pi_4\pi_0 + \pi_2\pi_5 \tag{3.50}$$

3.4 Performance Evaluation for Considered Scenarios

The performance evaluation considers the evaluation of the call blocking probability and call dropping probability are presented in different contexts. Three contexts comprising different scenarios having varying system parameters have been considered. These contexts are namely Context 1, Context 2, and Context 3.

Context 1 considers one scenario that examines the values of the call blocking probability and call dropping probability in a case where there is a varying value of the class A call arrival rate. In this consideration, network slices assigned to class A calls are rapidly utilized with values varying from 0.1 at minimum to 1.0 at maximum. In this case either class B calls, or class C calls can make use of the resources of the buffer introduced in the proposed solution.

Context 2 models the case of the network system while examining how a varying class B call buffering rate influences the call blocking probability and call dropping probability. The condition being considered here is such that class B calls require having access to the buffer resources. This can arise when network slices assigned to class B calls have all being utilized or when there is a significant number of class A calls being admitted and requiring access to network slices. In this case, varying values of class B call arrival rates are utilized in different scenarios within this context.

Context 3 considers the case where class C calls have a varying arrival rate for a given class A call arrival rate. In this case, the performance of the system metrics is examined for varying class B call buffering rates.

Each context and its associated scenarios are described by a different set of system model related parameters.

Chapter 4 - Results, data analysis and interpretation

The discussion in this section focuses on performance analysis and evaluates the benefits of using the proposed solution that incorporates buffer. It is recognized that the determination of the buffer optimum size for the proposed mechanism is essential. However, the focus of the discussion is examining the performance benefits of the proposed solution. The performance benefit is analysed for multiple contexts. Each of the considered context has its own optimum buffer size. The determination of the optimum buffer size requires evaluating the derivative of the concerned state equation for each of the concerned context. This requires the development and use of novel derivative techniques and is beyond the scope of the presented research.

The discussion is divided into two aspects. The first aspect presents the considered context and discusses the scenarios under each context. The second aspect presents and analyses the performance benefits via the simulation results. The performance evaluation and benefits analysis are done for the following scenarios:

4.1 Considered Contexts and Associated Scenarios

Context 1:

Scenario 1: In this case, class A calls are admitted into the system in a manner that they require access to an increasing number of network slices. The network slices in this context infer network resources i.e., bandwidth required for successful call execution. The call arrival rates, buffering rates and service rates associated with class B calls and class C calls are non-varying as the network delivers the expected functionality. This implies that the number of network slices being used by class B calls and class C calls remain constant. The scenario being considered is one in which the class A calls have the highest service rate (indicating allocation of high bandwidth resources). The arrival rate of Class A calls is varying spanning the values 0.1 – 1.0. In this case, class B calls have a higher arrival rate than class C calls. A similar context applies to the rate at which class B calls transit into the introduced buffer. In a similar manner, class B calls have a higher service rate than class C calls. The scenario in this case is described by the set of parameters given as $\lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.90, \mu_B = 0.56, \mu_C = 0.45,\ \lambda'_A = 0,\ \lambda'_B = 0.76,\ \lambda'_C = 0.45.$ In this case, the performance of the network system is examined.

Context 2: The performance evaluation in this case considers the effect of varying class B calls buffering rates on the call blocking probability and call dropping probability. In addition, a regime of different class A call arrival rate is considered. Class A calls are considered to have different arrival rates that signify the different rates of utilizing network slices. The scenarios being considered are Scenarios 2, 3 and 4.

Scenario 2: The scenario being considered in this case is one in which the buffering rate of class B calls is varied while other system related parameters remain unchanging. However, the class A call (highest priority class of calls) have a low demand for network slices. Essentially, a low utilization of the network slices allocated to class A calls is observed. In this case, the variable system related parameters refer to the arrival rate of class A calls, arrival rate of class B calls and arrival rate of class C calls. This also includes the service rates associated with these call classes and the buffering rate of class C calls. The performance evaluation procedure in this context aims to evaluate the system performance in a case where the class A call arrival rate is low. This implies that there is low demand for network slices by the calls of class pre – defined to have the highest priority. The scenario in this case is described by the set of parameters given as $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45$, $\lambda'_A = 0, \lambda'_C = 0.45$. In this regard, the low value of class A call arrival rate i.e. $\lambda_A = 0.1$ is used in the simulation.

Scenario 3: The case in scenario 3 is like the network context that is considered in scenario 2. The significant difference in this case being an increase in class A call arrival rate i.e., increasing the utilization of the network slices allocated to class A calls. In this case, the value of class A call arrival rate increases from $\lambda_A = 0.1$ to $\lambda_A = 0.2$. However, the value of the rest of other system related parameters remains the same and is described as$\lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45$, $\lambda'_A = 0, \lambda'_C = 0.45$. The consideration in this scenario differs significantly from that in Scenario 2 because of the increase in the value of class A call arrival rate.

Scenario 4: The context being considered here is a regime where there is an increase in the value of class A call arrival rate. In the regime, the value of the class A call arrival rate i.e., $\lambda_A = 0.2$ is now 0.3 instead of the previous value of 0.2 that is used in Scenario 3. The variation and its influence on the system performance is examined for varying class B buffering rates. The value of other system related parameters remains the same as utilized in Scenario 3 being given as $\lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45$, $\lambda'_A = 0, \lambda'_C = 0.45$.

Context 3: The performance evaluation in this case considers the effect of varying class B calls buffering rates on the call blocking probability and call dropping probability. In addition, a regime of different class C call arrival rate is considered. Class C call arrival rate values have been considered to investigate how transitions in the call class with the lowest network slices access priority influence the call blocking probability and call dropping probability. In this case, the considered scenarios are Scenario 5, Scenario 6, and Scenario 7.

Scenario 5: The case being considered is one where class C calls having the lowest priority have lowest utilization of assigned network slices i.e., network bandwidth resources. In addition, only class B calls are being admitted into the buffer because they have highest arrival rate in the scenario consideration. Class C calls have a low rate of network slice utilization and do not need to make use of the buffer's resources. In this case, the parameters are given as: $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.10, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda'_A = 0, \lambda'_C = 0.45.$

Scenario 6: In this case, there is an increase in the value of network slice resources assigned to class C calls. Nevertheless, only class B calls make use of the buffer resources i.e., are admitted into the buffer. This scenario is described by the parameter values given as: $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.20, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda'_A = 0, \lambda'_C = 0.45.$

Scenario 7: The consideration in this case is associated with an increase in the arrival rate of class C calls. This increase does not result in the overwhelming use of network slices allocated to class C calls. Hence, only class B calls are buffered. The concerned scenario is described by the parameter values given as: $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.35, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda'_A = 0, \lambda'_C = 0.45.$

4.2 Presentation and Analysis of Simulation Results

The performance evaluation results are presented and analysed in this subsection. The evaluation is done via the MATLAB simulation tool because of its flexibility and versatility. The evaluation is done with the aim of determining the CBP and CDP in the case of the existing solution and proposed solution. The CBP and CDP are examined for varying values of arrival rate of Class A calls i.e., λ_A and the buffering rate of Class B calls λ'_B . This is done to evaluate how increasing values of the arrival rate of Class A calls (λ_A) influences the CBP and CDP obtained via the simulation procedure. In addition, the evaluation examines how increasing values of the buffering rate of Class B calls (λ'_B). Furthermore, the CBP and CDP are examined for varying values of the buffering rate of Class B calls (λ'_B). This is done to determine how changing values of λ'_B i.e., the role of the buffer in the proposed solution influences the CBP and CDP.

The CBP and CDP are examined for the case of the parameters in the case of a varying λ_A given the parameters$\lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.90, \mu_B = 0.56, \mu_C = 0.45,\ \lambda'_A = 0,\ \lambda'_B = 0.76,\ \lambda'_C = 0.45.$

The results obtained for the CBP and CDP in this case are presented in Figure 4.1 and Figure 4.2, respectively. The results presented in Figure 4.1 and Figure 4.2 show that the incorporation of the proposed case (with buffer) reduces the call blocking probability and call dropping probability. In addition, the results presented shows that the introduction of the buffer results in a reducing CBP for an increasing value of Class A call arrival rates. In the case of the existing case, the absence of the buffer results in an increasing CBP for increasing value of Class A call arrival rates. Analysis of the presented results shows that using the buffer as introduced in the proposed solution reduces the CBP and CDP by 92.3% and 86.6% on average, respectively.

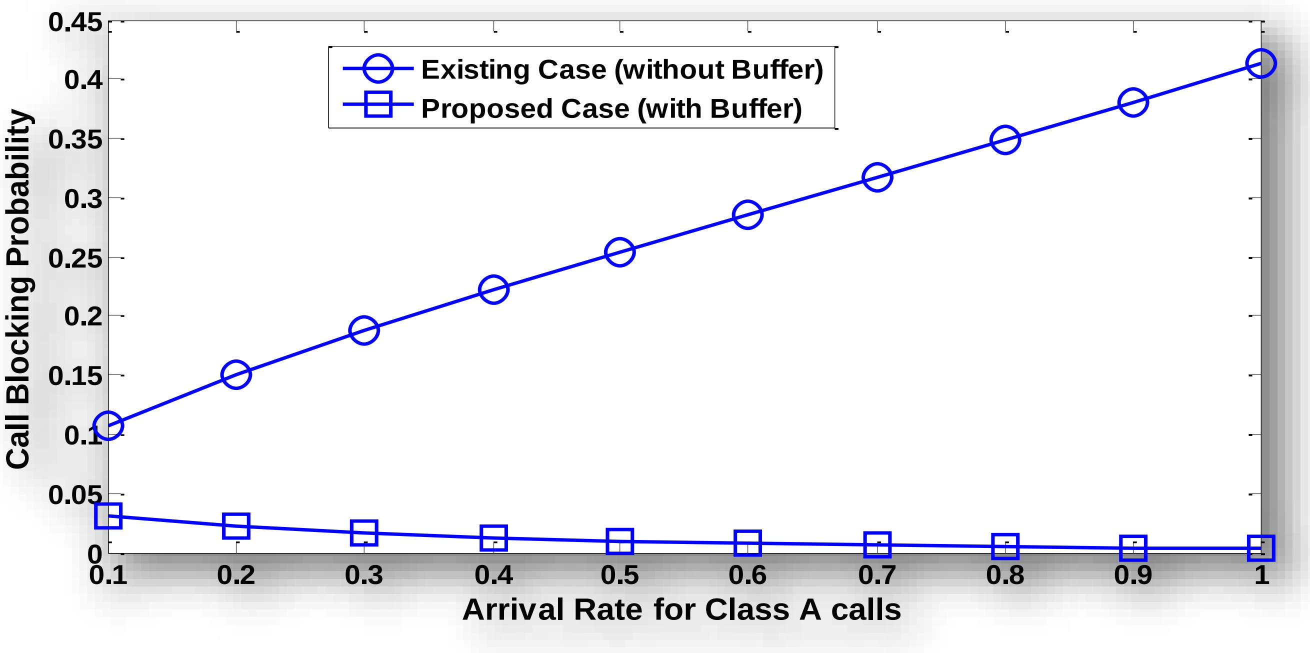


Figure 4.1: Context 1, Scenario 1 evaluation showing the call blocking probability results obtained in the simulation procedure when $\lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.90, \mu_B = 0.56, \mu_C = 0.45,\ \lambda'_A = 0,\ \lambda'_B = 0.76,\ \lambda'_C = 0.45$.

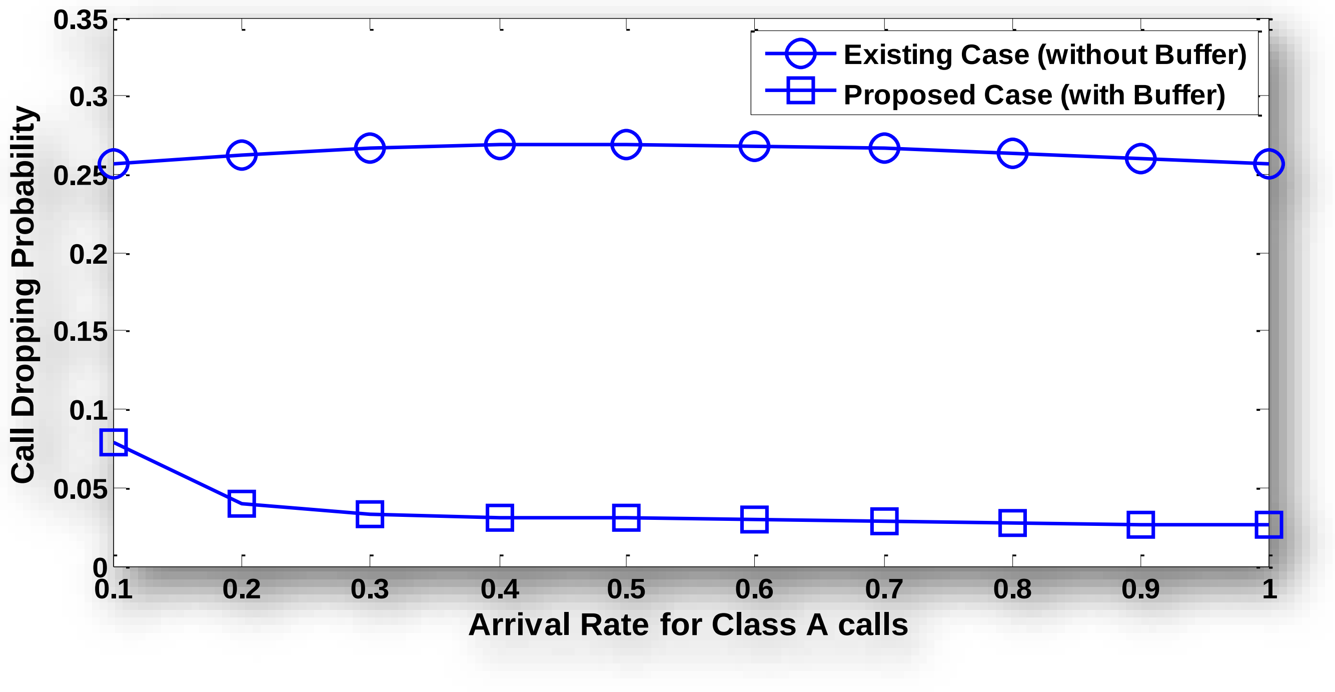


Figure 4.2: Context 1, Scenario 1 evaluation showing the call dropping probability results obtained in the simulation procedure when $\lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.90, \mu_B = 0.56, \mu_C = 0.45,\ \lambda_A' = 0,\ \lambda_B' = 0.76,\ \lambda_C' = 0.45.$

Analysis of the results in Figure 4.1, and Figure 4.2 shows that the use of the buffer reduces the CBP and CDP in the case of increasing rate of buffering for class B calls. In the evaluation, it is assumed that the buffering rate of Class A calls does not overwhelm the buffer capacity. This is feasible if the buffer capacity is over provisioned for the number of Class A calls. In the existing solution, increasing Class A call arrival results in a case where network slices are insufficient to execute subscriber demand. This results in the occurrence of call blocking or call dropping. The use of the buffer prevents some of the previously dropped calls from being un–served in the network. Analysis shows that the introduction of the buffer in this case reduces the CBP and CDP by 60.6% and 54.7% on average, respectively.

The performance evaluation procedure also examined the influence of the buffer being introduced in the proposed solution. In this case, the values of λ_B' is examined for three values of λ_A i.e. 0.1, 0.2 and 0.3 alongside the values given as $\lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda_A' = 0, \lambda_C' = 0.45$. The results of the CBP and CDP for the case of the parameter values $\lambda_A = 0.1,\ \lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda_A' = 0, \lambda_C' = 0.45$ are presented in Figure 4.3 and Figure 4.4, respectively.

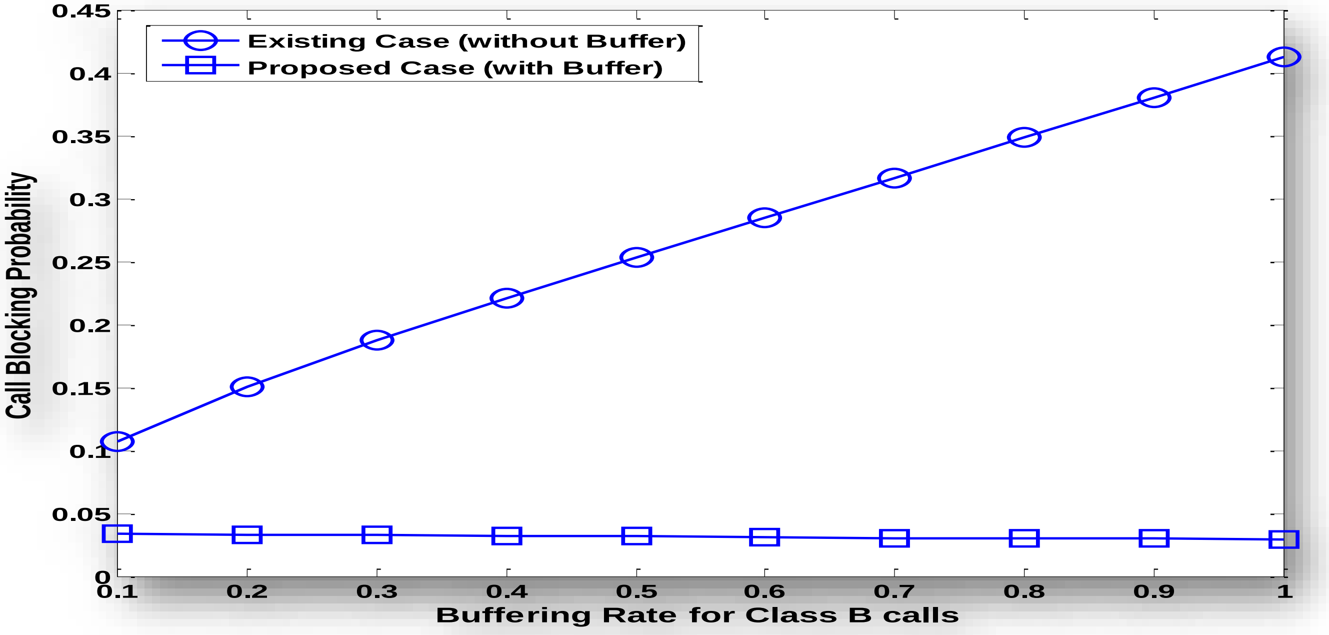


Figure 4.3: Context 2, Scenario 2::evaluation showing the call blocking probability results obtained in the simulation procedure when $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45 , \lambda_A' = 0, \lambda_C' = 0.45$.

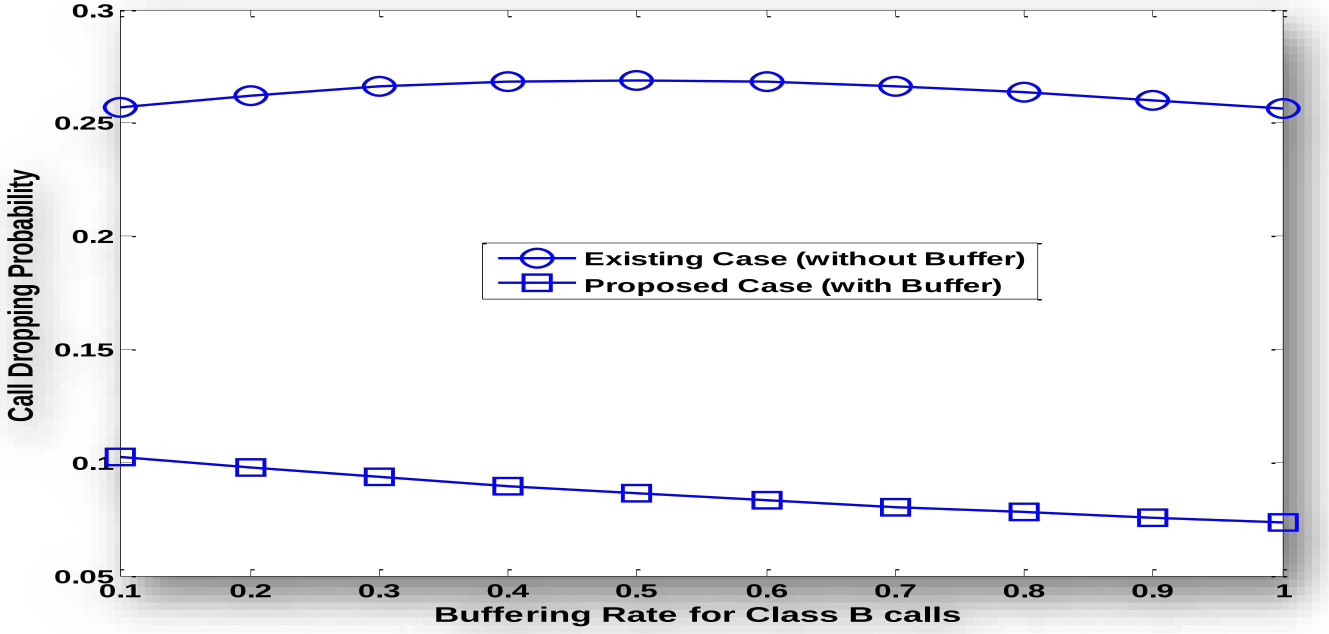


Figure 4.4: Context 2, Scenario 2: evaluation showing the call dropping probability results obtained in the simulation procedure when $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45 , \lambda_A' = 0, \lambda_C' = 0.45$.

Analysis of the results in Figure 4.3 and Figure 4.4 shows that the introduction of the buffer reduces the CBP and CDP in the case of increasing rate of buffering for class B calls. In the evaluation, it is assumed that the buffering rate of Class B calls does not overwhelm the buffer capacity. This is feasible if the buffer capacity is over provisioned for the number of Class B and Class C calls being buffered in the network. The presented results show that the CBP and CDP reduces with the introduction of the buffer in the proposed scheme for increasing values of the buffering rate of Class B calls. In the case of the existing solution, increasing arrival of Class B calls results in a case where network slices are increasingly insufficient to execute increasing subscriber demand. Hence, call blocking or call dropping results. However, the provision of the buffer in the proposed scheme prevents some of these previously dropped calls from being un–served in the network. Analysis shows that the introduction of the buffer in this case reduces the CBP and CDP by 70.6% and 66.5% on average, respectively.

In addition, the performance benefit of using the buffer in the proposed solution is evaluated for increasing value of buffering rate of Class B calls when Class A call arrival rates increases by 50% from 0.1 to 0.2. In this case, the additional parameters being considered are: $\lambda_A = 0.20, \lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda'_A = 0, \lambda'_C = 0.45$. The results for the CBP and CDP for the existing solution and proposed solution are presented in Figure 4.5 and Figure 4.6, respectively. In this case, the CBP and CDP are reduced by 83.7% and 83.9% on average, respectively.

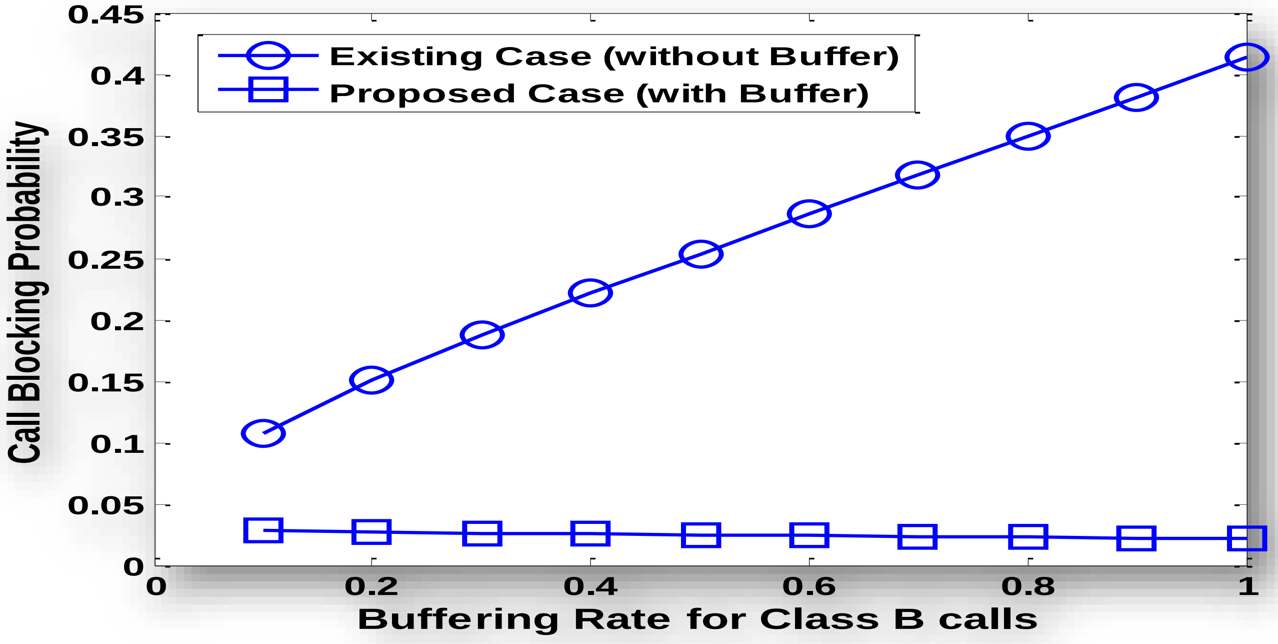


Figure 4.5: Context 2, Scenario 3:evaluation showing the call blocking probability results obtained in the simulation procedure for varying Class B buffering rates when $\lambda_A = 0.2, \lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda_A' = 0, \lambda_C' = 0.45$.

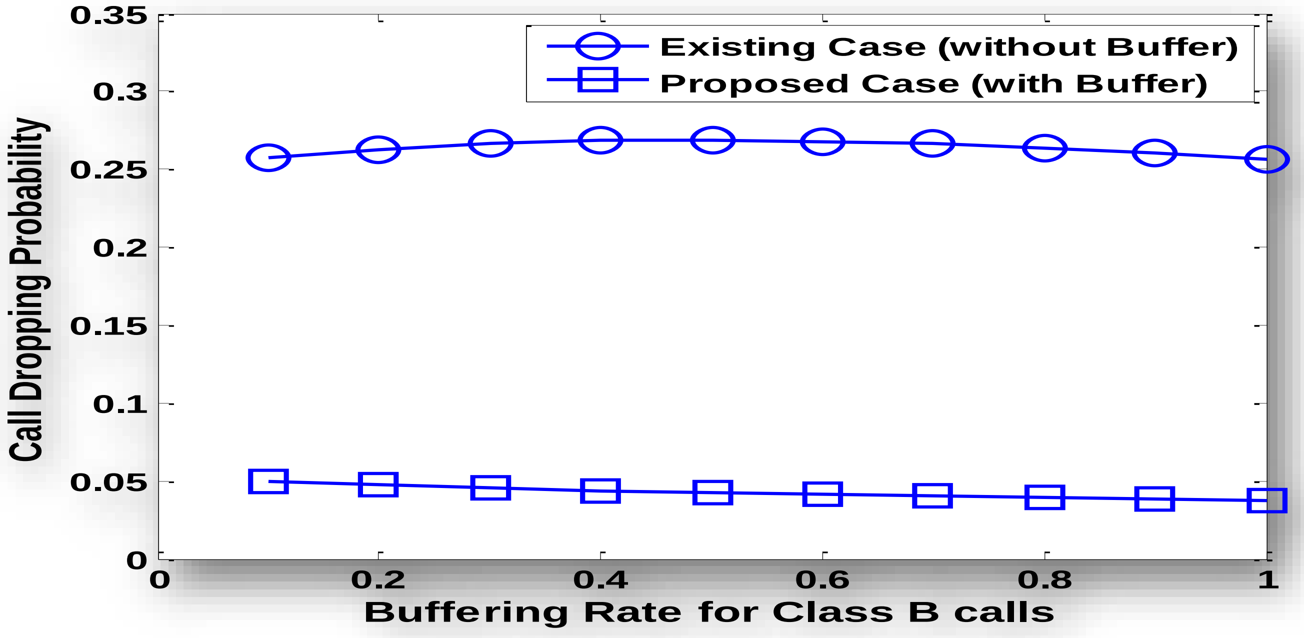


Figure 4.6: Context 2, Scenario :evaluation showing the call dropping probability results obtained in the simulation procedure for varying Class B buffering rates when $\lambda_A = 0.2, \lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda_A' = 0, \lambda_C' = 0.45$.

Furthermore, the performance benefit of introducing buffer is examined for an increase in the value of the class A call arrival rate by 33.3% from 0.2 to 0.3. The additional parameters remain the same being $\lambda_A = 0.30, \lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\,, \lambda'_A = 0, \lambda'_C = 0.45$. In this case, the results for the CBP and CDP for the existing solution and proposed solution are presented in Figure 4.7 and Figure 4.8, respectively. Analysis of the presented results also shows that the use of the buffer in the proposed solution reduces the CBP and CDP. Analysis of the results shows that the use of the buffer reduces the CBP and CDP by 70.2% and 68.4% on average, respectively.

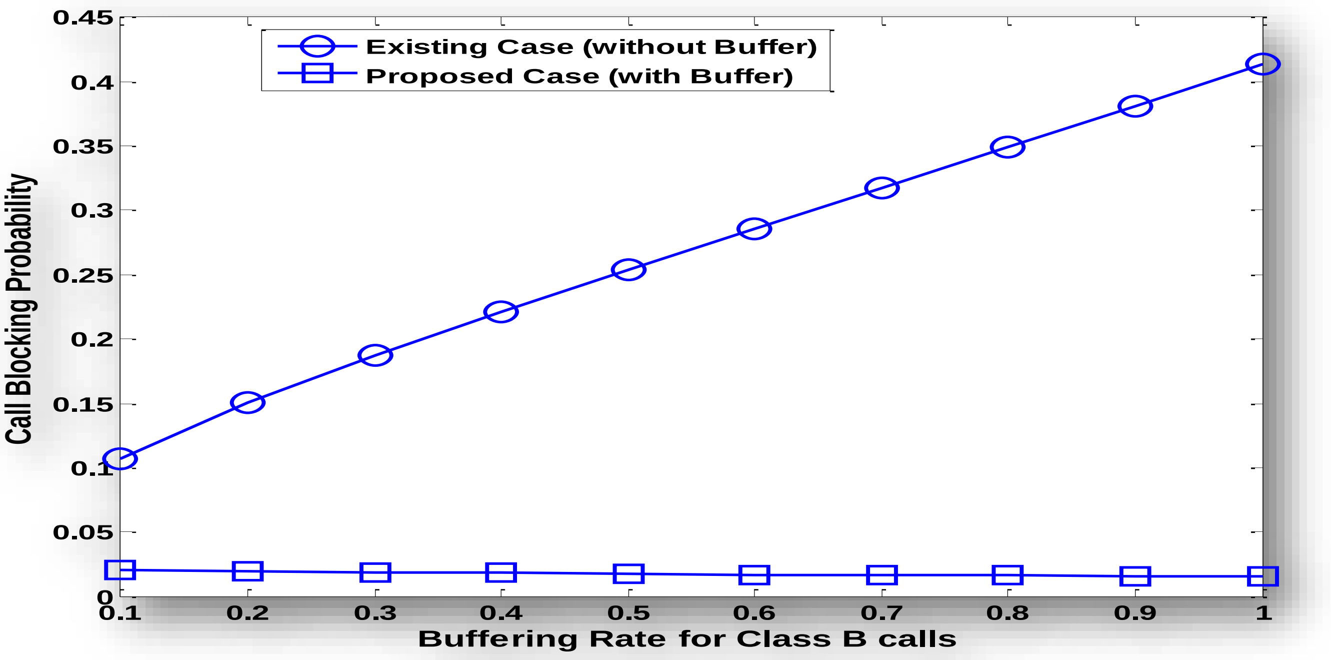


Figure 4.7: Context 2, Scenario 4 evaluation showing the call blocking probability results obtained in the simulation procedure for varying Class B call buffering rates when $\lambda_A = 0.3, \lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\,, \lambda'_A = 0, \lambda'_C = 0.45$.

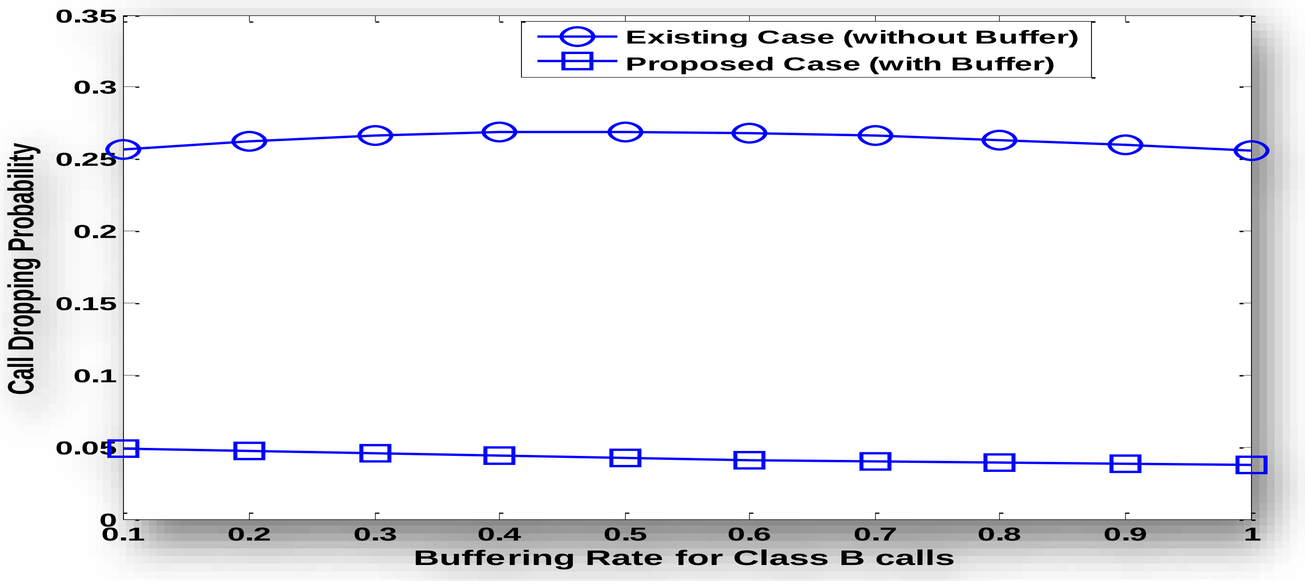


Figure 4.8: Context 2, Scenario 4 evaluation showing the call dropping probability results obtained in the simulation procedure for varying Class B buffering rates when $\lambda_A = 0.3, \lambda_B = 0.65, \lambda_C = 0.30, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45$, $\lambda'_A = 0, \lambda'_C = 0.45$.

In addition, the CBP and CDP are examined for the case of varying values of Class C call arrival rates. The value of the variable λ_C is increased from 0.10 to 0.20 (by 50%) and from 0.20 to 0.35 (42.9%). This is done to examine the effects of varying the values of λ_C given a set of values for Class B calls buffering rates. The results of the CBP and CDP for the case described as: $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.10, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45$, $\lambda'_A = 0, \lambda'_C = 0.45$ is presented in Figure 4.9 and Figure 4.10, respectively. The results of the CBP and CDP for the case described as $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.20, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45$, $\lambda'_A = 0, \lambda'_C = 0.45$ are presented in Figure 4.11 and Figure 4.12, respectively. In addition, the CBP and CDP obtained via simulation for the case described by $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.35, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45$, $\lambda'_A = 0, \lambda'_C = 0.45$ are presented in Figure 4:13 and Figure 4:14, respectively.

The results presented show that the use of the buffer in the proposed solution reduces the CBP and CDP. This observation is consistent with the earlier observed results. In addition, the discussion also analyses the performance results. The result of analysis shows that the CBP and CDP are reduced by an average of 74.7% and 51.6%, respectively. This performance benefit is obtained for the case of parameters given as $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.10, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45$, $\lambda'_A = 0, \lambda'_C = 0.45$. In addition, the performance benefit is also examined for the case given as $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.20, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45$, $\lambda'_A = 0, \lambda'_C = 0.45$. The analysis in this case shows that the use of the proposed solution reduces the CBP and CDP by 71.7% and 61.7% on average, respectively. Furthermore, the performance benefit is examined for the scenario described as $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.35, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda'_A = 0, \lambda'_C = 0.45$. In this case, the use of the proposed solution (including the buffer) instead of the existing solution (without buffer) reduces the CBP and CDP by 70.3% and 68.4% on average, respectively.

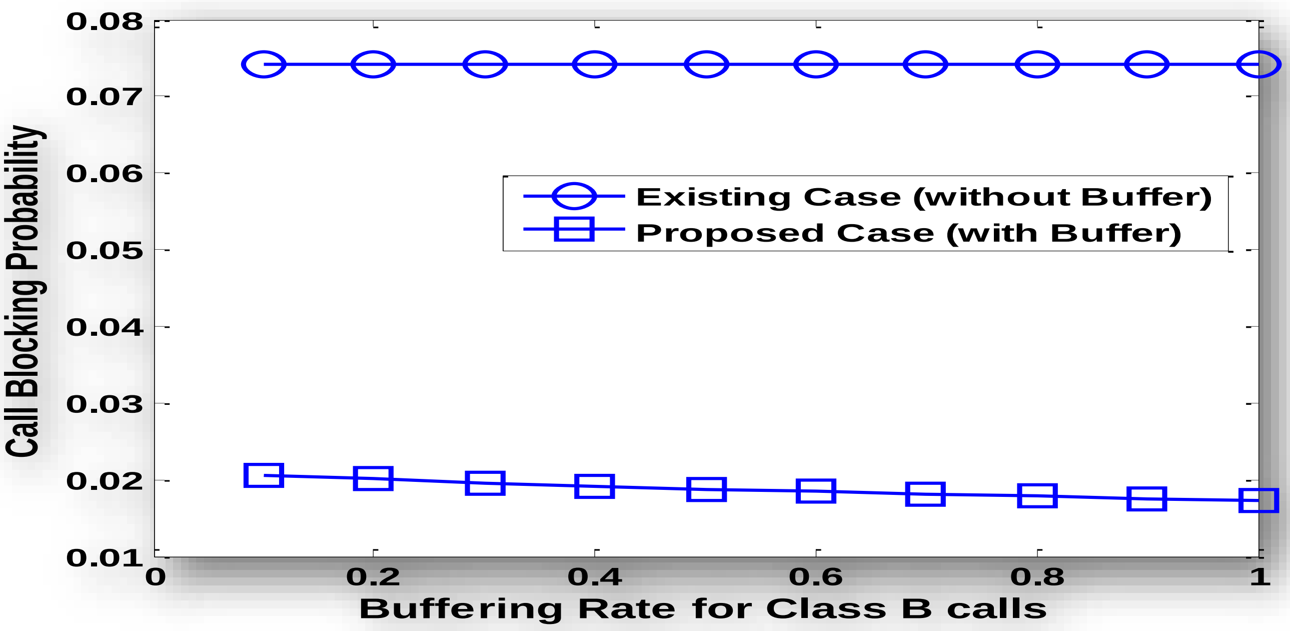


Figure 4.9: Context 3, Scenario 5 evaluation showing the call blocking probability results obtained in the simulation procedure for varying Class B buffering rates when $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.10, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda'_A = 0, \lambda'_C = 0.45$

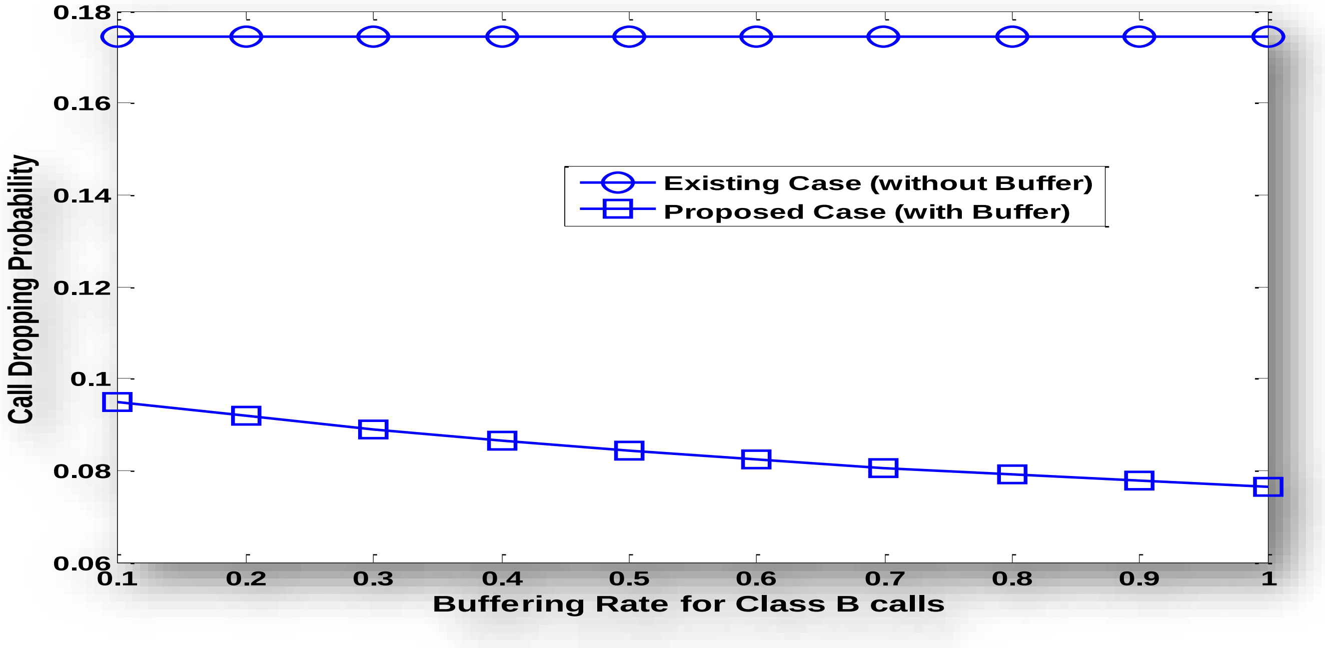


Figure 4.10: Context 3,Scenario 5 evaluation showing the call dropping probability results obtained in the simulation procedure for varying Class B buffering rates when $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.10, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda'_A = 0, \lambda'_C = 0.45$

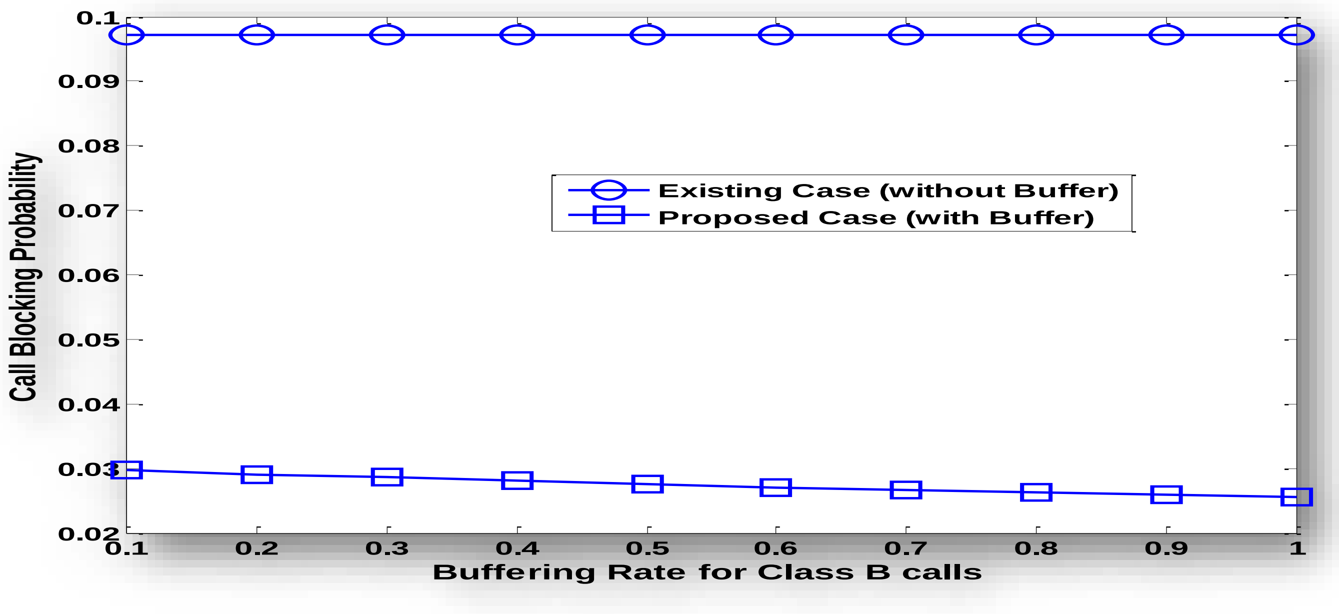


Figure 4.11: Context 3, Scenario 6evaluation showing the call blocking probability results obtained in the simulation procedure for varying Class B buffering rates when $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.20, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda'_A = 0, \lambda'_C = 0.45$

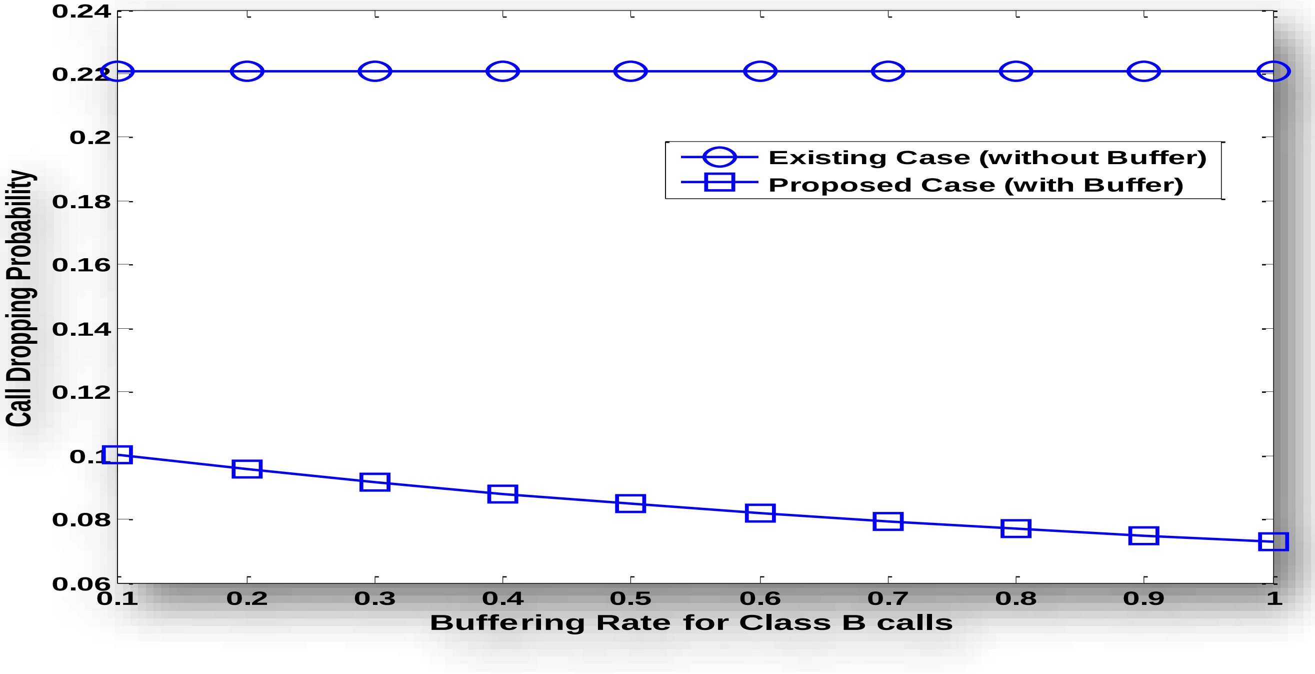


Figure 4.12: Context 3, Scenario 6 evaluation showing the call dropping probability results obtained in the simulation procedure for varying Class B buffering rates when $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.20, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda'_A = 0, \lambda'_C = 0.45$

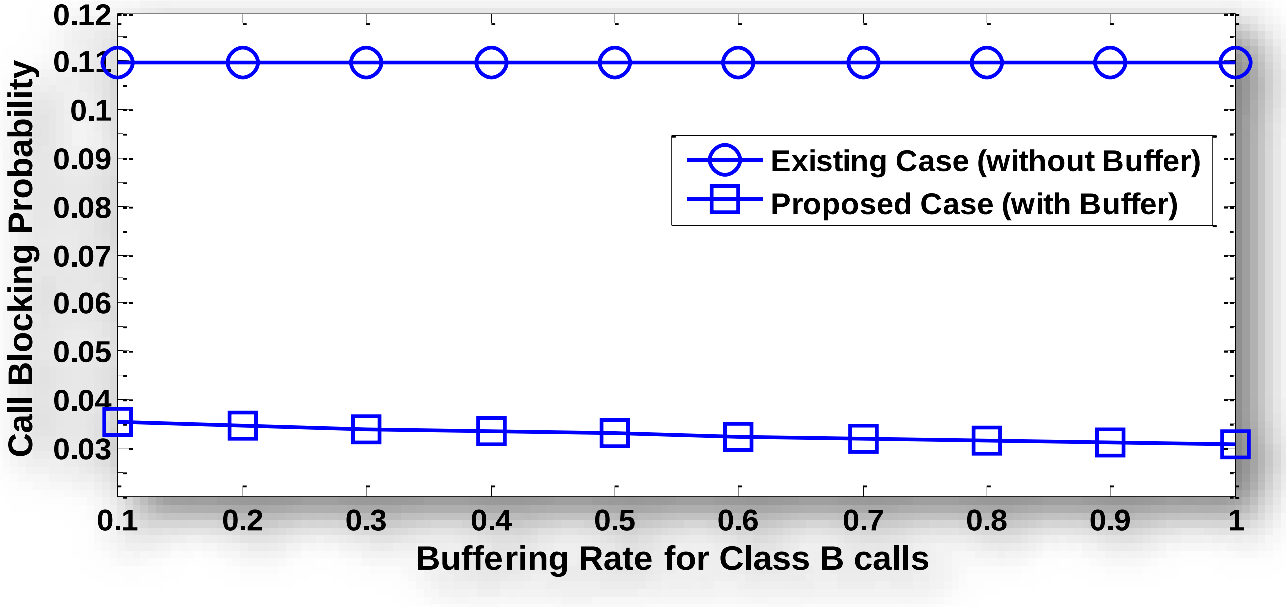


Figure 4.13: Context 3, Scenario 7 evaluation showing the call blocking probability results obtained in the simulation procedure for varying Class B buffering rates when $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.35, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45\ , \lambda'_A = 0, \lambda'_C = 0.45$

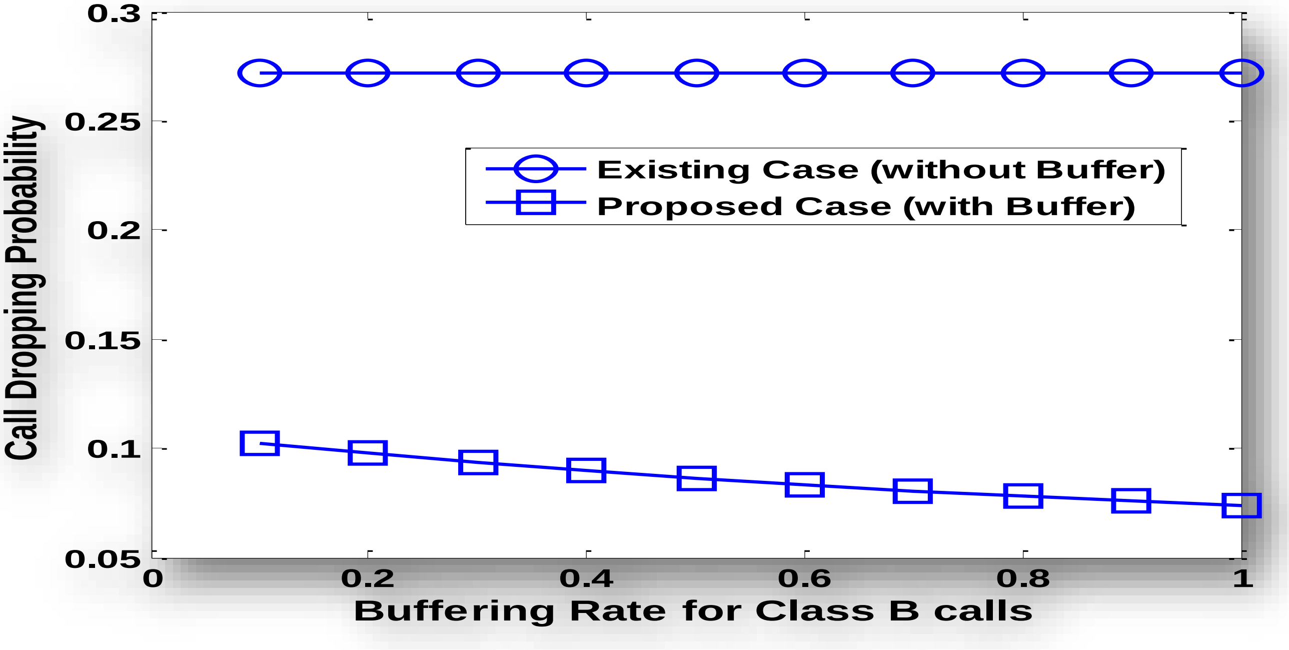


Figure 4.14: Context 3, Scenario 7 evaluation showing the call dropping probability results obtained in the simulation procedure for varying Class B buffering rates when $\lambda_A = 0.1, \lambda_B = 0.65, \lambda_C = 0.20, \mu_A = 0.9, \mu_B = 0.56, \mu_C = 0.45 , \lambda'_A = 0, \lambda'_C = 0.45$

The discussion in this aspect presents the formulated Markov chain analysis, derives the state equations, and gives the solution results. The solutions present the expressions for the state probabilities in the existing case (without buffer) and proposed case (with buffer). In addition, the discussions also present the results of performance analysis and benefits. The performance analysis is done considering different values that describe the arrival rate and departure rate of different call classes into the network under consideration. This is done for three perspectives. The first perspective examines the CBP, and CDP given an increase in the arrival rate of Class A calls. The second perspective considers the CBP, and CDP given an increase in the arrival rate of Class B calls for varying call B buffering rates. This is done for varying values of the arrival rate of class A calls. In the third perspective, the CBP and CDP are examined given an increase in the arrival of class C calls for varying call B buffering rates.

The performance analysis is done to examine how the use of the proposed solution reduces the CBP and CDP in each of the considered perspectives. The result of analysis shows that the average reduction of the CBP and CDP in the case of the first perspective is given as 92.3% and 86.6%, respectively. From the second perspective, the introduction of the buffer reduces the CBP and CDP by an average of (70.2 – 83,7) % and (68.4 – 83.9) %, respectively. In a similar manner, a reduction of CBP and CDP by an average of (70.3 – 74.7) % and (51.6 – 68.4) %, respectively. From the presented results and for the variations in the call related parameters in the presented network, the use of the buffer reduces the CBP and CDP. This is beneficial and enhances the quality of the network model that has been considered.

Chapter 5 - Conclusion and Recommendations for Future Work

The discussion in this aspect concludes the presented research, identifies, and discusses areas for future work. The rest of the discussion has two aspects. The first aspect presents insights on the conclusion of the presented research. The second aspect identifies and discusses aspects of the presented research that requires future research.

5.1 Conclusion

The presented research addresses the challenge of enhancing the quality of service for communications system in the context of future networking. The proposed communication system has the capability of supporting multiple subscribers. In the communication system, resources enabling the execution of voice calls and data calls are in form of network slices. These network slices describe the network bandwidth enabling successful communications and content access in data calls.

The considered subscribers are in different quality of service demand preferences. These preferences imply that each subscriber category can access a varying number of slices. The slices describe the number of network resources that can be used to execute either voice calls or data calls arising from subscribers in each network. The communication system accesses resources to meet subscriber demands in form of network slices. These network slices describer the number of resources accessible to subscriber call in different categories. The presented research has considered three different call categories. These are Class A calls (highest priority), Class B calls (medium priority) and Class C (lowest priority). Calls having the highest priority can access more network slices (resources) than lower priority calls. The challenge being addressed also considers a scenario where there is a significant number of calls having different priorities and requiring access to network resources. In this case, it is recognized that a significant number of calls will either be dropped or blocked. This is undesirable because it results in a high number of dropped calls and blocked calls thereby degrading network quality of service. The research proposes the use of buffer to enhance system performance when a significant number of calls require access to the communications network. The buffer provides temporary storage for calls that could have otherwise been blocked or dropped in the communication system. In its action, Class B calls (medium priority) and Class C calls (lowest priority) are buffered. Class A calls (highest priority) are not buffered as they have premium access to network slices (bandwidth resources) for executing either voice calls or data calls.

In addition, the research investigates the performance of using the buffer in the proposed communication network. This is done via the use of the Markov chain (Continuous Time Markov Chain) approach. The use of the Continuous Time Markov Chain enables the consideration of the transition between states (describing different call context) in the communication system. The transition enables the consideration of arrival rate and departure rate events of different calls. This is done for the case of all calls being considered i.e., Class A calls, Class B calls, and Class C calls. Furthermore, the research formulates the call blocking probability and call dropping probability as the main performance metrics. These metrics are formulated for the case where the network does not utilize the proposed buffer and when the network incorporates the proposed buffer. In addition, the evaluation was done in MATLAB via .m files written for this purpose. Analysis shows that the introduction of the proposed buffer reduces the call blocking probability and call dropping probability by an average of (70.2–83.7) % and (51.6–83.9) %, respectively.

The result of performance analysis shows that the introduction of the buffer significantly reduces the call blocking probability and call dropping probability. Therefore, the introduction of the buffer enhances the performance of the communication system in executing data calls and voice calls.

5.2 Aspects of Future Work

The following identified aspects of the presented research require consideration in future research work:

i. Buffer Sizing – The relations between the size of the buffer and the performance metrics require further consideration. In the presented research, the formulation and analysis are done considering that call arrival rates and buffering rates do not overwhelm the buffer capacity. This perspective does not assume an infinite buffer capacity as an infinite number of calls are not assumed to exist in the considered scenario.

ii. Increased Call Differentiation – The existing research considers that there are three categories of calls. However, more call categories can arise in a real-life scenario. The consideration of more call categories also requires a more rigorous development and analysis of the resulting Continuous Time Markov Chain. In addition, it is important to investigate the performance of the call blocking probability and call dropping probability in a case with more call categories. Increased call differentiation can arise in the case of a mobile virtual network operator with dynamic priority identities and seeking to access resources from a mobile virtual network enabler.

iii. Re–admission of Dropped or Blocked Calls – The presented research aims to ensure that call blocking or dropping is significantly minimized. However, it is still feasible that significant number of calls are dropped or blocked. In this case it is important to design an additional intelligent mechanism enabling the buffer to be capable of re-initiating a new attempt from previously dropped or blocked calls to access network slices (bandwidth resources) for voice call or data call execution.

iv. Buffer Utilization – It is also important for future research to address how the network responds to different states of buffer utilization. Three states of buffer utilization are considered. These are buffer underutilization, buffer in normal state of utilization and buffer overutilization. In the case of buffer underutilization, it is important that the buffer in the proposed network is able hold calls from another network. This can be realized via an intelligent network that incorporates call offloading.

v. Cloud Implementation of Buffer – The presented research has not recognized the role of the proposed buffer in a cloud integrated radio access network or optic – fibre-based communication network. In this case, the use of cloud computing infrastructure enables the realization of an approximated infinite buffer. However, the use and role of such a buffer within the context of a software defined network requires additional consideration.

References

1) GSMA, “An Introduction to Network Slicing Association,” 3,2017.
2) GSA White Paper, “5G Network Slicing for Vertical Industries,” 2017.
3) Massa, “Fibre Optic Telecommunication module 1,” pp 293,2000.
4) Intel White paper,5G End-to-End Network slicing [online] available https://www.intel.com >guides.
5) GSA White Paper, Contributions from Ericsson, Huawei, and Nokia, “5G Network Slicing for Vertical Industries,” September 2017.
6) G. Liu and D. Jiang, “5G: Vision and Requirements for Mobile Communication System towards Year 2020,” Volume 2016, Article ID 5974586, March 2016.
7) J.O. Lucena, P. Ameigerans, D. Lopez, J. J Ramos-Munzo, J. Lorca, and J. Folyueira, “Network Slicing for 5G with SDN/NFV: Concepts, Architectures and Challenges,” IEEE Communications Magazine,2017.
8) V.G Nguyen and K.J. Grinnemo, “SDN/NFV-Based Mobile Packet Core Network Architectures,” IEEE Communications surveys and tutorials, vol 19, (3),2011.
9) C. Collicutt, ”5G Network Slicing,” 5G network slicing and Openstack, Inter dynamic systems,2018.
10) J. Costa-Requena, J. Llorente Santos, V. F Guasch, K. Ahokas, G. Premsank, S. Luukkainen, O. Lopez Perez, M. U. Itzazelaia, L. Ahmad, M. Liyanage, M. Ylianttila, E. Montes de Costa, “SDN and NFV integration in generalized mobile network architecture,” European conference on networks and communications (EuCNC) ,154-158,2015.
11) M. Jiang, M. Condoluci, and T. Mahmoodi, “Network slicing management & prioritization in 5G mobile systems,” European Wireless conference pp 1-6,2016.
12) K. Katsalis, N. Nikaein, E. Schiller, A. Ksentini and T. Braun, “Network Slices Towards 5G Communications: Slicing the LTE Network,” I EEE Communications Magazine 55 (8) pp 146-154,2017.
13) X.Li, M. Samaka, H. Chan, D. Bhamare, G. Gupta, C. Guo, and R. Jain, “Network Slicing for 5G: Challenges and Opportunities,” IEEE Computer Society 21 (5) pp 20-27,2017.
14) P. Sharma, “Evolution of Mobile Wireless Communication Networks-1G to 5G as well as Future Prospective of Next Generation Communication Network,” Vol. 2, Issue 8, pg.47 – 53, August 2013.
15) F. Z, Yousaf, M. Bredel, S. Schaller, and F. Schneider, “NFV and SDN Key Technology Enablers for 5G Networks,” vol. 35, no. 11, pp. 2468-2478, Nov. 2017.
16) D. Tse, “Fundamentals of Wireless Communication,” University of California, Berkeley September 10, 2004.

17) L. Clara, W. Geng, A. Papathanassiou, U. Mukherjee, “An End -to-End network slicing Framework for 5G wireless systems communications systems,”1 August 2016.

18) C.Mannweiler, P.D.S. Rost Michalopoulas, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, “Network Slicing to enable Scalability and Flexibility in 5G Mobile Networks,” IEEE Communications magazine 55 (5) pp 72-79,2017.

19) R. Simon, “Antennas Propagation for wireless communication systems,” 2nd edition, University of Surrey Guildford UK, John Wiley, and sons ltd 2007.

20) R. Chundury, “Mobile broadband backhauls,” 3,2008.

21) K. Mebarkia, and Z. Zsoka, “Analysis of Network’s QoS in Service Chains,” International Symposium on Performance Evaluation of Computer and Telecommunication Systems, 20 – 22 July 2020, Madrid Spain, pp 1-9.

22) J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J.J. Ramos-Munoz, J. Lorca, J. Folgueira, “Network Slicing for 5G with SDN/NFV: Concepts, Architectures and Challenges,” IEEE 2017.

23) A. Basta, W. Kellerer, “Applying NFV and SDN to LTE Mobile Core Gateway,” proceeding of the 4th workshop on all things cellular, applications and challenges,33-38,2014.

24) M.R Bhalla and A.V Bhalla, “Generations of mobile wireless technology,” International Journal of computer applications,5 (4),26-32,2010.

25) S.M Ahsan Kazmi, L.U Khan, H.T Nguyen and C. Seon Hong, “Network Slicing for 5G and Beyond Networks,” Switzerland, Springer Nature ,2019.

26) C. Casetti, “The 5G Ecosystem Forges Ahead as Remote Services Take Centre Stage,” March 2021, IEEE Vehicular Technology Magazine, 2021, pp 7 -13.

27) I.C. Chochilouros, A.S. Spiliopoulou, P. Lazaridis, A. Dardamanis, Z. Zaharis, and A. Kostopoulous, “Dynamic Network Slicing: Challenges and Opportunities’, IFIP Advances in Information and Communication Technology,” (IFIPAICT), Vol. 585, 29 May 2020, pp 47-60.

28) D. Alotaibi, “Survey on Network Slice Isolation in 5G Networks: Fundamental Challenges,” Procedia Computer Science, Vol. 182, 2021, pp 38-45.

29) K. Mebarkia, and Z. Zsoka, “Analysis of Network’s QoS in Service Chains,” International Symposium on Performance Evaluation of Computer and Telecommunication Systems, 20 – 22 July 2020, Madrid Spain, pp 1-9.

30) M.F. Pervej, L.T. Tan, and R.Q. Hu, “User Preference Learning – Aided Collaborative Edge Caching for Small Cell Networks,” IEEE Globecom, 7 – 11 Dec 2020, Taipei, Taiwan, pp 1-6.

31) S. Sarmah and S. K. Sarma, "A Novel Approach to Prioritized Bandwidth Management in 802.11e WLAN," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033871.

32) M.Mamman and M.H. Zurina, “An Efficient Dynamic Call Admission Control for 4G and 5G Networks,” 9th International Conference on Signal, Image Processing and Pattern Recognition (SPPR 2020), pp 55-60.

33) H. Hermanns, J.P Katoen, J.M Kayser and M. Siegle, “A Markov Chain Model Checker,” conference on tools 2000.

34) M.A Laundteigen and M. Rausand, “Markov Chain,” Norwegian University of science and Technology.

35) R. A. Addad, M. Bagaa, T. Taleb, D. L. C. Dutra and H. Flinck, “Optimization model for Cross-Domain Network Slices in 5G Networks,” *IEEE Transactions on Mobile Computing,* vol. 19, no. 5, pp. 1156-1169, 2019.

36) R. M. Sohaib, O. Onireti, Y. Sambo and M. A. Imran, “Network Slicing for Beyond 5G Systems: An Overview of the Smart Port Use Case,” *Electronics,* vol. 10, no. 9, p. 1090, 2021.

37) S. Geetha and R. Jayaparvathy, “Dynamic bandwidth allocation for multiple traffic classes in ieee 802.16 e wimax networks: A petrinet approach,” *Journal of Computer Science,* vol. 7, no. 11, p. 1717, 2011.

38) M. Vincenzi, E. Lopez-Aguilera and E. Garcia-Villegas, “Timely Admission Control for Network Slicing in 5G With Machine Learning,” *IEEE access,* vol. 9, pp. 127595-127610, 2021.

39) N. Yarkina, Y. Gaidamaka, L. M. Correia and K. Samouylov, “An analytical model for 5G network resource sharing with flexible SLA-oriented slice isolation,” *Mathematics,* vol. 8, no. 7, p. 1177, 2020.

40) N. Van Huynh, D. T. Hoang, D. N. Nguyen and E. Dutkiewicz, “Optimal and Fast Real-time Resources Slicing with Deep Dueling Neural Networks,” *IEEE Journal on Selected Areas in Communications,* vol. 37, no. 6, pp. 1455-1470, 2019.

41) B. Han, “A Markov Model of Slice Admission control,” 30 August 2018.

42) S. Bakri, “Towards enforcing network slicing in 5G Networks,” 28 January 2021.

43) J.D. Mallapur, S. Abidhusain, S. Soumya Vastrad and A.C Katageri, “Fuzzy Based Bandwidth Management for Wireless Multimedia Networks,” International Conference on Business Administration and Information Processing, 81-90,2021.

44) S. Sarmah and S.K Sarma, “A Novel Approach to Prioritized Bandwidth Management,” 5th International Management Conference for Convergence in Technology 2019.

45) K. Sharma, N. Dhir, “IJCSIT International Journal of Computer Science and Information Technologies,” Vol. 5 (6), 2014, 7810-7813.