# Design of a Load-balancing Algorithm for Multipath Video Streaming in Heterogeneous Wireless Networks


Presented by:
Mutikedzi Mudzanani
MDZMUT003

Prepared for:
Professor Olabisi Falowo Dept. of Electrical and Electronics Engineering
University of Cape Town

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

**October 30, 2023**

# Declaration

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

Signature:... M.Mudzanani .....

Date:... **October 30, 2023** ...

Word Count(9772)

# Terms of Reference

| | | |
|---|---|---|
| **Student proposed?** | *Y/N* ***N*** | *If Y, student name* |
| **ID:** | **OF-3** | |
| **SUPERVISOR:** | Olabisi Falowo | |
| **TITLE:** | Design of a Load-balancing Algorithm for Multipath Video Streaming in Heterogeneous Wireless Networks | |
| **DESCRIPTION:** | The next generation wireless network is heterogeneous consisting of multiple radio access networks coexisting in the same geographical location. Some of the mobile terminals designed for next generation wireless networks have multi-homing capability, with the ability to connect to multiple networks simultaneously. The objective of this project is to design a load-balancing algorithm for video steaming in heterogeneous wireless network. The load balancing decision could be made from the mobile terminal side or from the network side. | |
| **DELIVERABLES:** | A review of load balancing schemes, a load-balancing algorithm, simulation results, and report. | |
| **SKILLS/REQUIREMENTS:** | MATLAB or any other programming language, EEE4121F. | |
| **GA1: Problem solving:** *Identify, formulate, analyse and solve complex* engineering problems creatively and innovatively* | The student is expected to (1) design an algorithm for load balancing in a heterogeneous wireless network supporting multi-homing, and (2) implement the algorithm. | |
| **GA 4**: Investigations, experiments and analysis:** *Demonstrate competence to design and conduct investigations and experiments.* | The student is expected to investigate the performance of the designed load balancing algorithm through simulations. | |

| | |
|---|---|
| **EXTRA INFORMATION:** | For a student interested in pursuing a master's degree, the project can be expanded to an MSc dissertation. |
| **BROAD Research Area:** | Wireless Networks |
| **Project suitable for ME/ ECE/EE/ All programmes?** | **EE/ECE** students who have taken EEE4121F course. |

# Acknowledgments

I would like to take this opportunity to express my deepest gratitude to my supervisor, Prof Olabisi Falowo for his guidance and support throughout the whole project.

I am also thankful to my family for their continuous encouragement and belief in my abilities.

I am grateful to my friends and colleagues who assisted me in making this academic journey enjoyable.

# Abstract

Watching videos online takes up most of the data on the internet and over the past ten years, internet traffic due to streaming videos has increased massively. The increase in mobile devices and the demand for high quality video streaming poses a significant challenge due to congestion created. Simultaneously the rise in streaming services such as Netflix, YouTube, Amazon Prime Video, and user generated content facilitated by platforms such as TikTok, continuously overwhelm the internet with video data.

Even though there have been improvements in video compression methods, enhancing computing capabilities, expanding internet speed and capacity, and increasing bandwidth, there is still a struggle to achieve the best load balancing in video streaming through multiple paths in heterogeneous wireless networks. The concept of Dual Connectivity(DC) was introduced which allows two network connections to occur simultaneously, increasing throughput and reducing service interruption. This DC can also be used in achieving load balancing.

In this study, a load-balancing algorithm suited to LTE and 5G DC was designed with the aim of optimizing multipath video streaming in heterogeneous wireless networks. Leveraging both resources from LTE and 5G simultaneously, allows this algorithm to effectively distribute traffic to LTE and 5G networks, balancing the load on the networks. Traffic splitting mechanisms proposed were based on the capacities and residual bandwidths of the two networks.

The service splitting mechanism designed in this paper allows the incoming calls to be split such that portions of the bandwidth of each call can be serviced by the residual bandwidths in the two networks simultaneously. This is to combat the drawbacks of not utilizing the residual bandwidth in each network if they are not enough to admit a call alone, thereby reducing the blocking and dropping probability.

An analytical model of the splitting mechanism is formed and its performance is evaluated using the blocking probability of new calls, dropping probability of handoff calls, and average utilization of the network as performance metrics. The simulation outcomes indicate that the developed algorithm effectively reduces the occurrence of blocks and drops of calls while achieving better resource utilization in networks where mobile devices have the ability to make multiple connections.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Background to the study

In today's world, the consumption of video content through video streaming over wireless networks has become popular and it is now an essential part of our daily lives. According to [1], it was estimated for video streaming to account for 82% of global Internet traffic by the end of 2022. In the initial months of 2023, video streaming has recorded an audience reach of 92% among internet users globally [2]. Video streaming has dominated the internet traffic and it seems to be increasing every year.

From streaming high-definition movies or videos to live video streaming, the demand for seamless and high-quality video streaming experience is rising. Even though there have been improvements in video compression methods, enhancing computing capabilities, expanding internet speed and capacity, and increasing bandwidth, the study in [4] stated that coping with the unprecedented user demands for Quality of Experience (QoE) remains a significant challenge. This is driven by a struggle to achieve the best load balancing in multi-path video streaming in heterogeneous wireless networks.

To ensure acceptable QoE to clients, it is crucial to address the complex issue of balancing the load in networks, especially within the context of LTE-5G Dual Connectivity(DC). DC allows the utilization of both LTE and 5G resources simultaneously leveraging the benefits of both LTE and 5G networks [5]. In video load balancing, DC involves distributing video traffic across the two networks, taking into consideration network factors such as the available bandwidth, network congestion, and device capabilities.

The dynamics of network conditions make it hard to achieve effective load balancing between networks. Designing a load-balancing algorithm for multi-path video streaming in heterogeneous wireless networks is a complex problem that requires advanced algorithms that are capable of adapting to dynamic network conditions to optimize video streaming performance.

## 1.2 Objectives of this study

### 1.2.1 Problems to be investigated

The increasing network traffic due to video streaming and the demand for better QoE in video streaming make it essential to address the issue of balancing the network load. Given that certain mobile devices engineered for next-generation wireless networks possess multi-homing capabilities, this study aims to address the issue of load balancing in multipath video streaming specifically tailored for LTE-5G Dual Connectivity. To achieve this goal, this report investigates the problem by:

- A review on Dual Connectivity in non-standalone 5G network
- A review on video splitting mechanisms
- Design of a video splitting mechanism to balance the load between LTE and 5G networks.
- Simulate the designed splitting mechanism with different network parameters
- Analyse the performance of the splitting mechanism by comparing the blocking and dropping probabilities with a scenario where there is no splitting of calls.
- Analyse the network utilization achieved by the splitting mechanism.
- Suggest what can be done to increase the efficiency of the splitting mechanism

### 1.2.2 Purpose of the study

The aim of this research is to design and assess the effectiveness of a load-balancing algorithm intended for enhancing the performance of multipath video streaming in diverse

wireless networks. By achieving this purpose, this paper aims to enhance the quality of video streaming experienced by users while optimizing the resource utilization in the network. Video streaming quality is assessed by evaluating the effects of changing the network parameters in blocking and dropping of video calls. For this algorithm to be meaningful, it should minimize the blocking and dropping probabilities while maximizing resource utilization.

## 1.3 Scope and Limitations

The primary emphasis of this project lies in the development of a load-balancing mechanism for multipath video streaming specifically tailored to LTE-5G Dual Connectivity. The effectiveness of this algorithm is assessed using metrics, including the call blocking probability, call dropping probability, and average network utilization. The design and evaluation do not include hardware implementation, and the algorithm's performance evaluation is based only on simulations.

## 1.4 Plan of development

The document is structured into the subsequent segments:

**Chapter 1:** This gives a brief explanation of what the research project is about. It talks about the background, goals, and what the research will and won't cover.

**Chapter 2:** Delves deeper into the research of Dual Connectivity and the work related to video splitting algorithms to attain load balancing in different types of wireless networks.

**Chapter 3:** Describes the network model to be evaluated in this research. It details the Markov models of the splitting mechanisms designed.

**Chapter 4:** Details the simulation designs on MATLAB

**Chapter 5:** Details the performance results obtained from the simulations. Discusses the results and the comparisons of the two splitting mechanisms.

**Chapter 6&7:** Draws conclusions and recommendations from the results obtained.

# Chapter 2

# Literature Review

Most of today's network consists of wireless communications that are continually developing to improve access to information by transmitting signals such as voice, data, and multimedia. In wireless communications information is transmitted without the use of cables which is particularly vital for mobile communications. However, this comes with an increased demand for high performance especially in multimedia. With video streaming being one of the most popular network services, this has led to many proposed solutions to improve the quality of video streaming.

This project focuses on designing an algorithm for balancing the load during multipath video streaming in diverse wireless networks and also evaluating its performance. Hence the aim of this literature review is to review the designs that have been proposed concerning this load-balancing algorithm. The structure of the literature review is as follows, begin by explaining the keywords, review of different algorithms proposed based on LTE-5G Dual cConnectivity, research gaps, and conclusion.

## 2.1 Heterogeneous Wireless Networks

Heterogeneous Wireless Network refers to network architecture that consists of devices using different radio access technologies (RATs) to provide better network performance, coverage, and capacity. These networks offer enhanced user experience as they provide better data rates, improved signal quality, and reduced latency which contributes to improved user experience for services including voice calls and video streaming. A heterogeneous wireless network example in 5G is shown in figure 2.1.

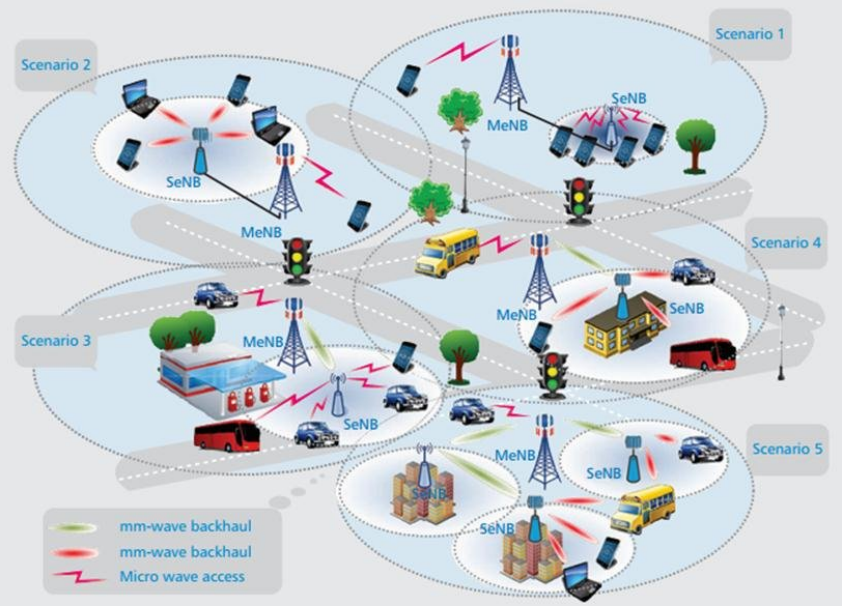


Figure 2.1: Illustration of Heterogeneous Wireless Network in 5G [3]

## 2.2 Multipath Video Streaming

Multipath video streaming refers to segmenting a video into smaller segments or packets that are sent simultaneously through different paths in a network instead of relying on a single network path. The receiving device will then combine and synchronize the segments to reconstruct the video. This increases the complexity of video streaming, hence this will need algorithms to manage the distribution of segments and reassembling to form a video. Multipath offers several benefits, including enhanced reliability, sustained connections, improved apparent data transfer rate, and load distribution [4].

## 2.3 Load Balancing Algorithms in Video Streaming

As discussed above, efficiency in multi-path video streaming is achieved by the use of algorithms. This includes load-balancing algorithms that play a vital role in enhancing the performance of video streaming by efficiently distributing video segments across multiple paths. This minimizes the latency, reduces congestion in the paths, increases reliability, and enhances the video quality experienced by the user [4]. The user needs to have a multi-homed device that allows for connecting to multiple network interfaces to increase reliability, resilience, and performance. Due to changing network conditions and streaming demands, load-balancing algorithms are frequently adjusted to adapt to these changes.

One effective approach to achieving load balancing of video streaming in multi-path heterogeneous networks is by implementing dual connectivity. This Dual Connectivity enables dynamic load balancing, as the network can distribute video traffic to two independent networks based on their states. This concept is reviewed in the following subsections.

## 2.4 4G-5G Dual Connectivity

The concept of Dual Connectivity was initially proposed for LTE in Release 12 of 3rd Generation Partner Project. In Release 15 of the 3GPP, the idea of Dual Connectivity was intended to enhance the performance of 5G systems [5]. Dual Connectivity enables users to concurrently utilize two RATs simultaneously i.e LTE and 5G operating at different frequencies.

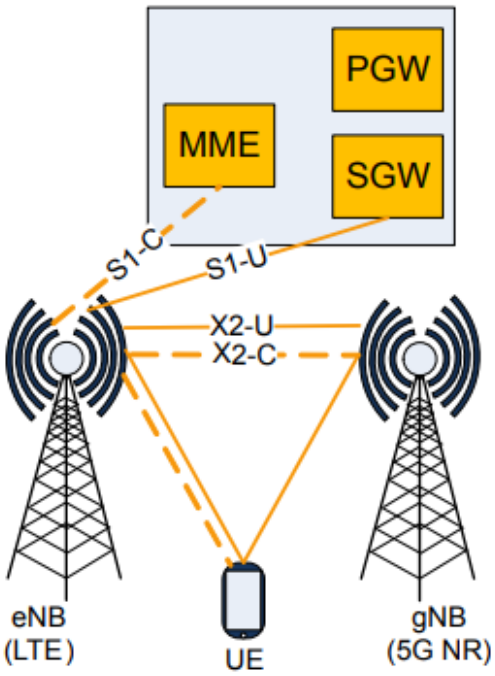


Figure 2.2: Demonstration of 5G Non-Stand Alone (NSA) [5].

Figure 2.2 shows the physical architecture of 5G NSA. This ensures that the user leverages the benefits of both LTE and 5G systems which enhance the performance of the network experienced by the user. A lot of work has been done on the benefits of dual connectivity that helps in improving load balancing. Some of these benefits has been discussed in the following subsections.

### 2.4.1 Mobility and Handover Management

Small cells deployment in macro base stations of dual connectivity has been considered to be the most practical approach to efficiently handle a large amount of mobile data traffic cost-effectively. However, this deployment of small cells has caused high handover,co-channel interference, high link failure rate in high mobility scenarios and also increases the complexity of the network infrastructure [6] [7].

A survey article [7] shows that the concept of dual connectivity can solve the problem of co-channel interference. In load balancing these small cells create a problem of frequent handovers as a result of the low coverage area of small cells and their high frequency which covers short distances as they are unable to pass through walls and other obstructions.

Because of the user's mobility, it is very common for handover scenarios to occur. This can be either when the User Equipment (UE) transitions from an old LTE eNB to a new 5G gNB or switches between two small cells, moving from an existing 5G gNB to a new gNB. In the survey, the pivotal metric concerning handovers is the Reference Signal Received Power (RSRP). For instance, when the RSRP value of the new gNB exceeds that of the connected gNB, the UE undergoes a handover. Conversely, if the RSRP value of the new gNB is lower, the connection remains unchanged.

The proposed mobility and handover management solution in the literature consists of self-adjustment and self-management of the network, trying to minimize failures occurring during handovers in mobile situations. Self-healing approaches to detect damaged and out-of-service paths in a multi-path network can improve handovers since the faulty paths can be avoided in future handovers. With these solutions combined, seamless handover can be achieved in dual connectivity and as a result, an optimized load balancing algorithm can be designed.

### 2.4.2 Data Splitting

The concept of dual connectivity can enhance the overall throughput but there are many issues relating to this. This will need packet reordering and resource management since data is divided into two portions, with one directed toward the macro base station and the other toward the small base station, while UEs simultaneously receive data from both sources. UEs can receive packets that are out of order and the user is required to store the out-of-sequence packet until the old packets in the pipeline arrive.

To mitigate the above problem, a fountain code was introduced which works together with dual connectivity, the original data can be encoded using fountain codes prior to transmission, and the user collects the packets from both base stations and then proceeds to decode them. The transmitter always codes and transmits the fountain code until the receiver accumulates a sufficient number of packets to reconstruct the data that was sent and after that the receiver sends back a finish indicator[6].

In NSA dual connectivity scheme proposed by [8], two methods for splitting traffic, specifically the UE feedback-based and the BS observation-based mechanisms were discussed. These mechanisms enable the adjustment of the sequence in which packets arrive at the UE side to prevent the problem of receiving packets in a sequence that is not in the correct order in the Packet Data Convergence Protocol layer. This solves the problem of out-of-order packets.

In [9] they proposed a traffic offloading system based on video quality that adopts the fountain code to deal with packet reordering challenges faced by video streaming services when operating over 5G networks that use dual connectivity. They also employed SDN technology to monitor the dynamic network status, while achieving efficient radio resource allocation managed by an intelligent Software-Defined Networking (SDN) controller with a global outlook on the internet.

The SDN controller collects network status details such as bandwidth, latency, and loss of packets from both master BSs and small cell APs or secondary BSs, along with video content characteristics. When network congestion is anticipated the SDN controller initiates the resource allocation system which calculates the target video quality achievable with macro-cell and small-cell resources. This information is used to establish criteria for distributing radio resources and the resource allocating module assigns radio resources to users based on these criteria [9].

Data splitting together with encoding in dual connectivity enable optimal resource utilization as user data traffic is distributed between the base stations based on their current load and capacity. This helps in mitigating congestion as data can be offloaded from the congestion base station to the other which prevents congestion and maintain acceptable performance levels. This is an advantage to load balancing algorithms as it improves load balancing accuracy by making more accurate decisions on when to split data between base stations.

### 2.4.3 Throughput Improvement

A throughput simulation project was done by [10] where they implemented non-standalone architecture by adapting the dual connectivity of User Equipment to both LTE and 5G New Radio networks. The utilization of millimeter-wave (mmWave) frequency bands offers a huge spectrum but these bands are more susceptible to blockages and obstacles. The concept of this modified dual connectivity was introduced to overcome these drawbacks of mmWave.

Within the protocol stack of the adapted Dual Connectivity in [10], the UE accommodates the protocol stack of two RATs using Packet Data Convergence Protocol as the layer of integration of dual connectivity. The User Equipment establishes an enduring Radio Resource Control connection with an LTE evolved Node B and designates it as the coordinator. The LTE eNB is responsible for selecting a suitable 5G New Radio gNodeB and establishing a connection with it. This arrangement allows the LTE base station to manage the control plane functions for the gNodeB, while the gNodeB solely manages user plane tasks, making a non-standalone design for the execution of 5G New Radio networks.

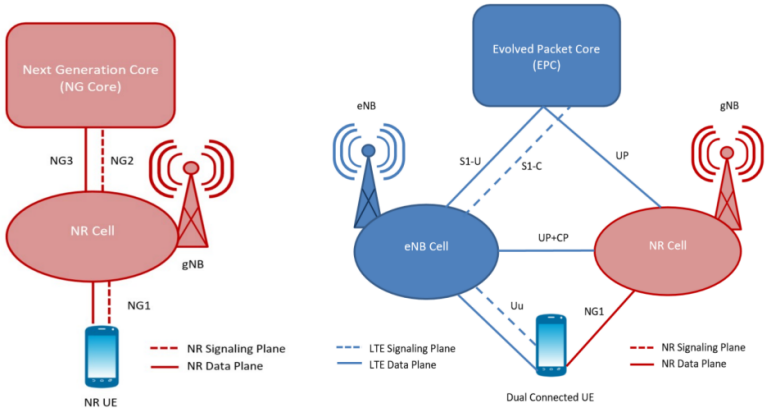


Figure 2.3: Illustration of SA and NSA modes of 5G that were simulated [10].

The standalone mode and non-standalone mode of 5G shown in figure 2.3 were simulated, expanding the separation between the User Equipment and the gNodeB, and updating the distance at intervals of 100 milliseconds. As the User Equipment continually moved away from the gNB, the overall throughput in standalone mode gradually decreased over time. Also, signal-to-interference-plus-noise-ratio reduces gradually as the distance between UE and gNodeB increases. This occurred because there was no backup connection to the User Equipment when the serving gNodeB experienced an outage.

However, for dual connectivity in non-standalone, the overall throughput stayed constant

for a certain period of time. This accomplishment was made possible by maintaining the connection between the User Equipment and the LTE module even when the gNodeB experienced an outage. This continuous connection prevents any interruption in data services and also facilitates gNodeB switching to enhance data transfer rates.

In [5] it was also shown that the throughput for dual connectivity is double the throughput for single connectivity. From the throughput results presented by [5], it was clear that Dual Connectivity, as a feature, improves user experience by increasing throughput. This is also cost-effective as the 5G NR is being added to an already existing LTE site. Since dual connectivity allows for the utilization of large bandwidth, it can offer increased data speeds in addition to the total throughput achieved by combining dual connectivity.

This increased throughput shows that the network can handle a larger volume of traffic data simultaneously. Therefore in this case the network can accommodate more users who are streaming videos. Load balancing can take advantage of this increased throughput to distribute traffic more evenly among the available resources in multiple paths.

### 2.4.4 Proportional Fair Scheduling

In [11], they discussed an algorithm for scheduling heterogeneous radio access technologies to satisfy user requirements while ensuring proportional fairness and also allocating resources among RATs to mitigate inter-cell interference.

They used an interference graph model to represent their network topology where the vertex represents the base stations and the links represent the available paths. To mitigate inter-cell interference in base stations, they establish clusters. The cluster is employed to schedule base station access in a manner that minimizes the probability of collisions. Users that are in the same cluster can perform the same scheduling algorithm simultaneously but users in different clusters cannot.

They discussed proportional fair scheduling that takes channel conditions into consideration for single connectivity. The users that are serviced first are the ones with the highest priority parameters. They calculated the priority parameter of a user by dividing the traffic rate requirement by the cumulative average transmission rate for that user in the same slot. If the communication of the user succeeds, then they increase the corresponding cumulative average rate of transmission so that it can have less chance of being chosen for the next round of scheduling. Conversely, when the requirements are more demanding, the numerator becomes larger, thus enhancing the likelihood of being scheduled in the

upcoming round.

In dual connectivity, the proportional fair scheduling scheme has to be designed with double service lines to serve LTE and 5G NR. Therefore the wireless spectrum is segmented into resource blocks. Resource blocks are supplied to users according to their priority parameters. Inter-cell interference is a frequent occurrence in Dual Connectivity due to the close proximity of base stations. To combat this they use the signal-to-interference-and-noise-ratio (SINR) on each resource block. With this SINR they compute different interference environments for various combinations of base stations and users. They stored this information in different states and choosing carefully which state should be served first reduces the inter-cell interference [11].

Proportional Fair scheduling aims to achieve fairness in resource allocation and efficient use of resources by ensuring that all users receive a fair share of network resources. This also allows traffic prioritization which allows load balancing mechanisms to allocate resources based on the specific requirements of each type of traffic. This contributes to effective load balancing.

## 2.5 Other Related Works on load balancing

### 2.5.1 Joint call admission control algorithm

The current Joint Call Admission Control (JCAC) algorithms developed for heterogeneous wireless networks block or drop incoming calls when there isn't sufficient bandwidth available to accept them in each of the Radio Access Technologies. This increases the blocking/dropping probabilities in cases where the bandwidth required by the call does not fit in one RAT. In [12], Olabisi introduced a Joint Call Admission Control algorithm that opts for multiple Radio Access Technologies when none of the accessible Radio Access Technologies possesses adequate remaining bandwidth to admit the incoming call. This is shown in figure 2.4

This algorithm considered session splitting so that portions of a call can be distributed to multiple networks to achieve load balancing. This JCAC algorithm gives preference to handoff calls over new calls by implementing distinct rejection thresholds for each. The choice to distribute the call is determined by the call category, the call's bandwidth requirements, and the present load in each of the accessible RATs. The RATs with

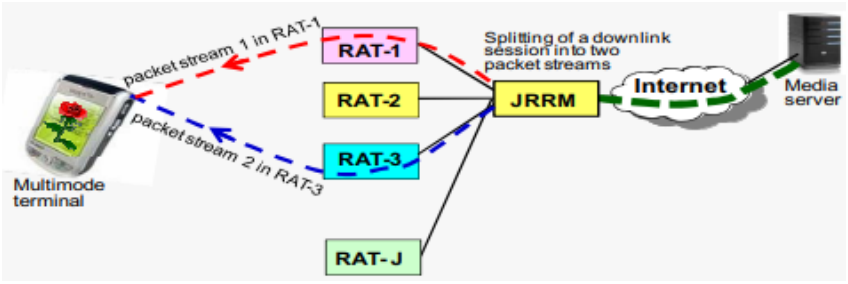


Figure 2.4: Session splitting between RATs [12].

the greatest remaining bandwidth capable of accommodating the call's service class are chosen for the call.

The performance of the suggested JCAC approach was assessed against a JCAC strategy that lacks the capability for multiple RAT selections and session splitting. In the proposed scheme, the likelihood of blocking new calls was consistently lower than that of the JCAC scheme which does not involve session splitting and RAT selection. The same trend applied to dropping probability. This study demonstrated that the proposed JCAC strategy minimizes the overall call blocking and dropping probability in heterogeneous wireless networks [12]. However, this study did not explore scenarios where all the accessible RATs could be utilized to attain higher bandwidth utilization.

Another efficient algorithm for multipath selection of video transmission called FITPATH was introduced in [13]. FITPATH simultaneously transmits video flows with different bitrates. Like other algorithms, this approach takes into consideration the dynamic network conditions in real-time and bitrate requirements in order to achieve load balancing, minimize packet losses, and reduce delays.

FITPATH allows for multiple video sources at the same time, hence this approach employs an efficient algorithm to identify a set of potential paths and then ranks these paths based on the flow requirements. Figure 2.5 illustrates an overview of how the FITPATH works in selecting the best path.

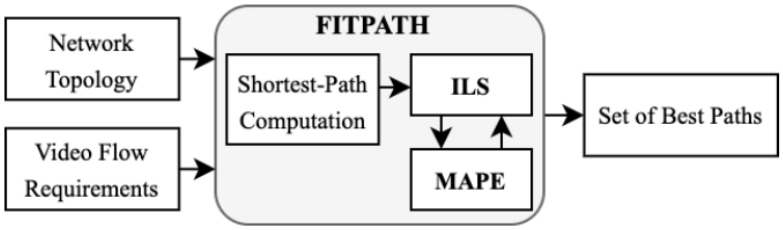


Figure 2.5: Overview of FITPATH scheme [13].

As shown in Figure 2.5, FITPATH takes Network topology and video flow requirements from different sources as input into the system. Then it calculates the shortest path to create a collection of potential paths. This set is used by the Iterated Local Search(ILS) as input into this subsystem. ILS utilizes an iterated search algorithm and uses a Multimedia-Aware Performance Estimator (MAPE) to evaluate each solution. MAPE is responsible for evaluating the network performance of potential paths in terms of throughput, delay, and loss, which are the metrics used by ILS to determine the optimal path for each video flow.

After simulations on network simulator 3, FITPATH outperformed prior splitting mechanisms by having a higher end-user quality of experience which symbolizes better network performance. I had the lowest end-to-end delay, lowest video frame losses, and highest throughput. However, this algorithm did not consider a case of session splitting as a way to improve performance.

### 2.5.2 Congestion-aware multi-path load balancing

The launch of 5G has led to increased user demand for improved QoS and network performance [14]. Among the various strategies for enhancing network performance, load balancing stands out as one of the most frequently employed policies and it improves the performance by resolving load imbalances, poor resource utilization, and data transmission latency.

The authors in [14] gave an insight into some of the algorithms that have been proposed over the past years. These are Equal Cost Multipath (ECMP), Weighted ECMP (W-ECMP), Dynamic and Adaptive multi-path routing (DAMR), Dijkstra-Repeat algorithm, as well as SDN-based Dynamic Flowentry-Saving Multipath (DFSM) algorithm. These algorithms improve link utilization rate as well as network reliability but they lower the bandwidth utilization. This led the authors in [14] to propose an SDN-based dynamic load-balancing method in multipath routing to meet the constraint of bandwidth.

The Depth First Search (DFS) algorithm implements the SDN controller to acquire the status of the network's topology. Then the system starts calculating the link costs in the topology and assigning weights to each path as the sum of the link costs in that particular path. The weight of a link is defined as the ratio of the reference bandwidth to the remaining available bandwidth which is not used at the moment. The path with the lowest sum of link weight is chosen for data transmission as a means to prevent path

congestion.

In [15] a multipath-aware TCP (MA-TCP) is introduced, which can detect the movement of TCP flows between paths. The objective of this algorithm is to migrate the congested flows to the other paths that are less congested in order to balance the traffic loads. MA-TCP identifies and gathers data about congested and non-congested paths at the end host, and it re-routes flows according to the collected information. The path selection is triggered whenever congestion is detected on the current flow.

Another congestion-aware load-balancing algorithm that is appropriate for 5G base stations is the RavenFlow algorithm[16]. Similar to other congestion-aware algorithms, RavenFlow makes forwarding decisions based on the link congestion value. For congestion monitoring, it uses the Discounting Rate Estimator(DRE) in order to identify congestion at a local level and for global congestion, it uses feedback packets that will be sent to switches when it is necessary to get the overall congestion.

### 2.5.3 Multi-RAT load balancing for LTE and 5G

The 5G-ALLSTAR project, which is an acronym for "5G AgiLe and flexible integration of SaTellite And cellular", has proposed 4 algorithms that were being developed for multi-connectivity [17]. The fourth one is dealing with load balancing where the traffic level of every cell is assessed within each Radio Access Technology. This load balancing algorithm consists of two steps that occur in the cells namely intra-RAT load balancing together with the inter-RAT load balancing.

This algorithm begins with intra-RAT load balancing, in which it transfers the suitable edge User Equipments from an overloaded cell to neighboring cells with lower utilization. Subsequently, it re-evaluates the load, and if it still exceeds the threshold, it proceeds to inter-RAT load balancing. In inter-RAT load balancing, data that can tolerate delays is offloaded to a satellite link.

The physical architecture that they used contains two types of RATs which are 5G NR RATs and satellite RATs which makes it a multi-RAT network. The UE requests fro connection which may be for video streaming and specifies the bitrate that it desires from the AP and the AP itself will compute the physical resources to be assigned to User Equipment in to fulfill the bitrate required in the request. These computations are done considering the path loss between the User Equipment and the access point, inter-RAT and intra-RAT interference, and thermal noise.

There is a Connection Admission Control that is implemented for the UE to measure the receiving power from APs and select the best AP to connect with. Because of the limited availability of resources on the Access Point side, it is possible for UE to get resources for a lower bitrate than the desired one as a way to balance the load on the AP side.

Another solution to enhance the resource distribution for video transmission in 5G was proposed by [18]. This was done to mitigate the following gaps: 1) The need for a precise and straightforward approach for estimating delays in real time. 2) a system for transmitting scalable videos that leverage multipath cooperation so that more than one mobile edge computing server can transmit video to a client. They designed a scheme with software-defined networking architecture for collaborative video transmission through multiple paths which also analyses the real-time delay with the objective of optimizing the video qualities of all clients.

The module for multi-path cooperative transmission optimization in the micro-cell base station (MBS) incorporates information about the transmission rate quality, physical channel condition, and packet delay which is used in determining the best cooperative transmission strategy. Therefore after finding the best strategy, the MBS transmits control information regarding resource allocation and packet scheduling to the mobile edge computers(MECSs). MECs will send their cached videos simultaneously to the DASH client through a small base station(SBS).

## 2.6 Research Gaps

There is a need for research that will combine the simple effective load balancing algorithms that already exist in order to develop a more robust system of load balancing in multi-path video streaming. Achieving efficient seamless handover while optimizing resource allocation poses a significant research challenge for video streaming in Dual Connectivity.

Simultaneously there is a higher demand for dynamic resource allocation strategies that consider both user preferences and network conditions, ensuring better network resource utilization. Furthermore, research on how load balancing through Dual Connectivity affects the dropping and blocking of calls in a heterogeneous network is needed. This is to access the ability of dual connectivity to handle the growing demands for video streaming.

In conclusion, this research has delved into the use of Dual Connectivity in LTE-5G

heterogeneous network which enhances the performance of the network experienced by the user. Throughout the reviewed studies, it is evident that Dual Connectivity increases the network capacity, data rates, and throughput. This shows that Dual Connectivity in LTE-5G network has the potential to enhance load balancing in the network. This project focuses on enhancing load balancing in LTE-5G multipath video streaming.

# Chapter 3

# SYSTEM MODEL

## 3.1 PROPOSED SPLITTING MECHANISMS

The Dual Connectivity network considered in this model consists of LTE and 5G networks. There are two types of splitting mechanisms proposed namely arrival splitting and service splitting. Arrival splitting is shown in figure 3.1, where the quantity of calls arriving is split among the two networks. Service splitting is shown in figure 3.2, where the bandwidth of each call is split and distributed to the two networks simultaneously, this means that both LTE and 5G will have portions of each call at a time.

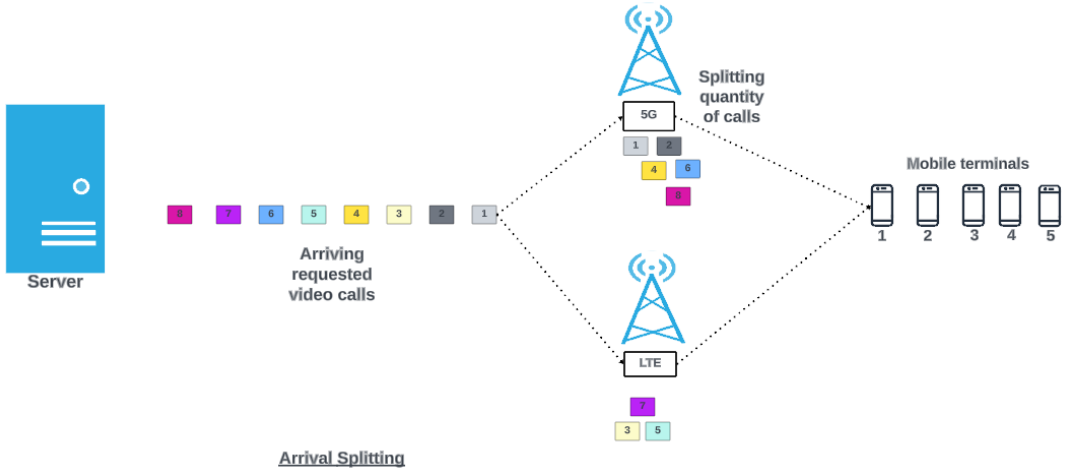


Figure 3.1: Illustration of arrival splitting

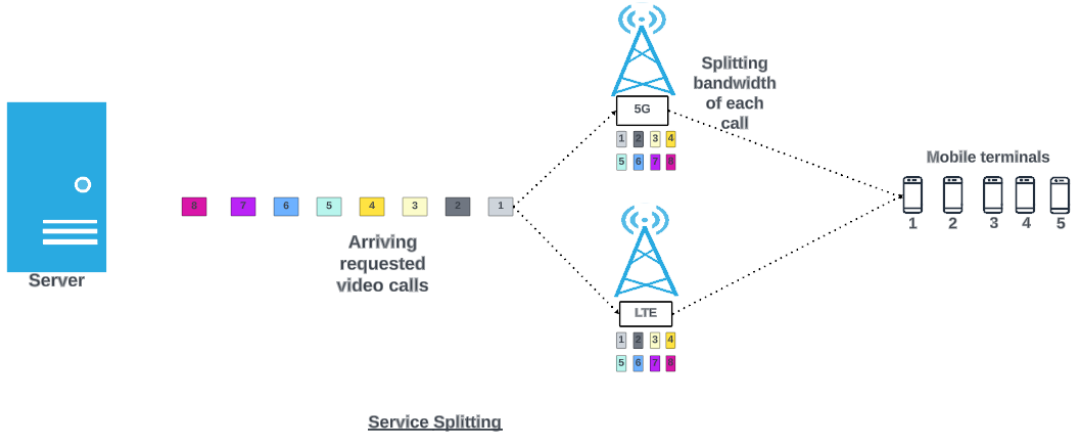


Figure 3.2: Illustration of service splitting

### 3.1.1 Arrival splitting

In this mechanism when new calls arrive in the network, the system will use the capacities of LTE and 5G to determine how many calls will be assigned to LTE and how many will be assigned to 5G. For example, when there are six incoming new calls and the ratio of LTE capacity to 5G capacity is 1:2, LTE will receive two calls and 5G will receive four calls. The same applies to handoff calls. This ensures that the network with higher capacity will receive more calls than the network with lower capacity in order to prevent overloading the network with low capacity while the network with high capacity has more bandwidth that is unused.

If this splitting of calls is not allowed, i.e., the residual bandwidth in each of the networks is not enough to accommodate the assigned calls, the calls will be blocked or dropped depending on whether they are new calls or handoff calls. The flow chart in Figure 3.3 shows how this arrival splitting mechanism works. To check if the proposed splitting is allowed for new calls, the system checks if the bandwidth that will be used by all the calls in the network is less than the threshold for rejecting new calls in the network. For handoff calls, the system checks if the bandwidth that will be used by all the calls in the network is less than the capacity of the network.

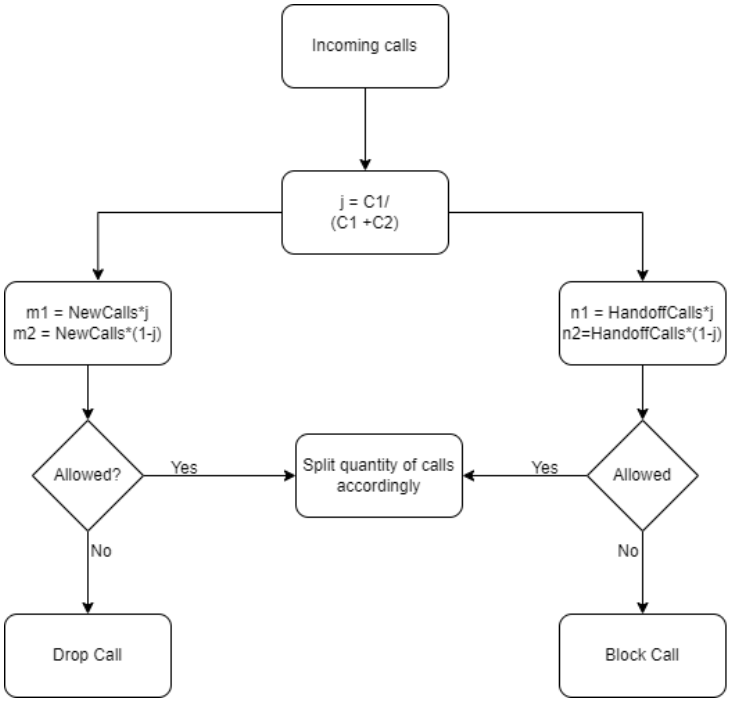


Figure 3.3: Flowchart of arrival splitting algorithm

### 3.1.2 Service splitting

In this mechanism, the bandwidth of each call gets split depending on the ratio of the capacities. Each network receives a portion of each call simultaneously. If the ratio of LTE capacity to 5G capacity is less or equal to 1/3, each call will be split such that LTE will receive 1/3 of the total bandwidth per call and 5G will receive 2/3 of the total bandwidth. If the ratio is between 1/3 and 2/3, equal splitting will be considered. If the ratio is greater or equal to 2/3, then LTE will receive 2/3 of the required bandwidth per call and 5G will receive 1/3 of the bandwidth.

If the above splitting combinations cause calls to be blocked or dropped in one of the networks, then the splitting is adjusted such that this particular network receives less bandwidth, i.e, 1/3 of the total bandwidth per call. Again if this splitting causes a block/drop then the whole call is blocked or dropped in both networks. This ensures that the network with higher capacity will receive larger portions of calls and the network with lower capacity will receive smaller portions of calls in order to balance the load. The flowchart in Figure 3.4 illustrate this service splitting mechanism.

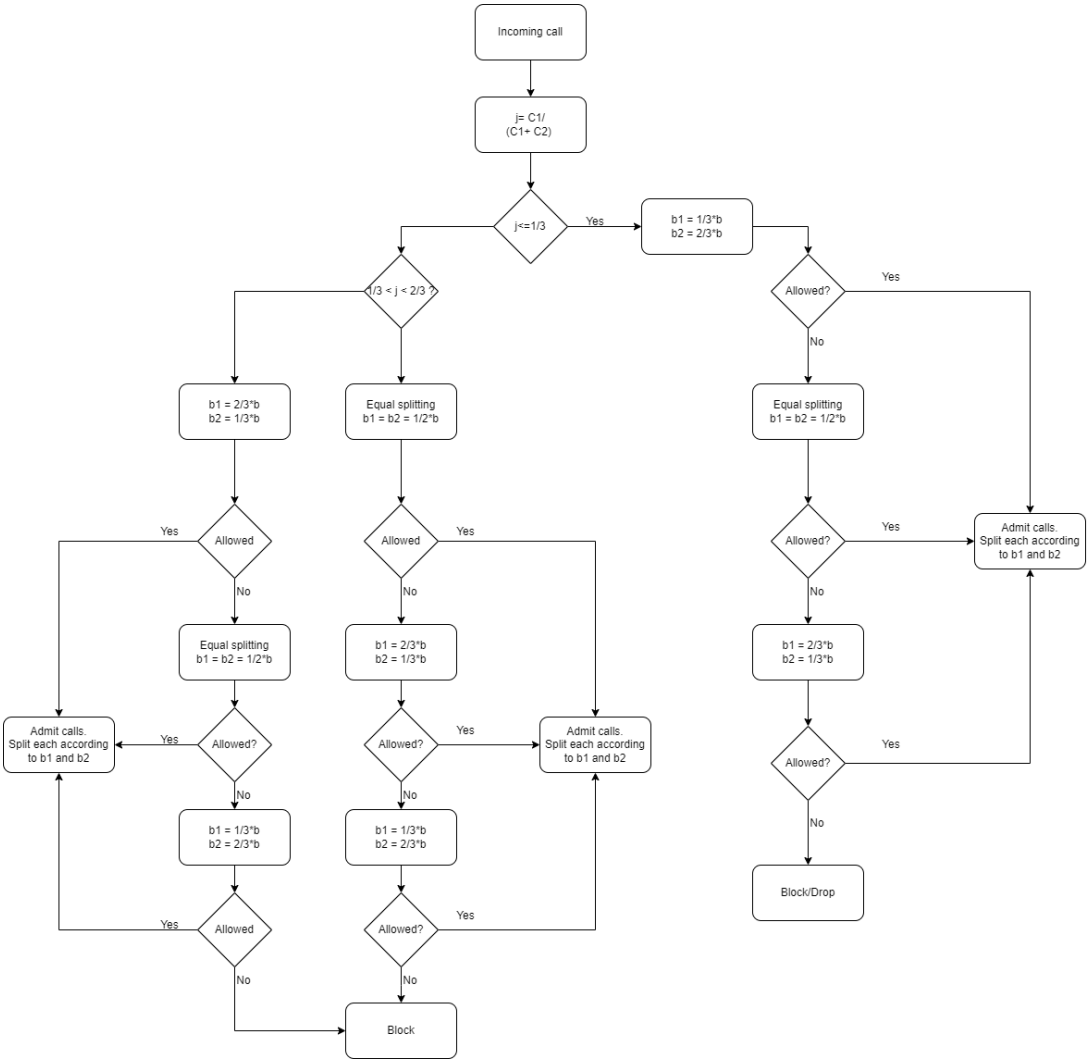


Figure 3.4: Flow chat of service splitting algorithm

### 3.1.3 Network Model Assumptions

In all the above calls the bandwidth is given in basic bandwidth units(bbu).The capacity of networks represented by $C_i$ is equal to the total available bandwidth in a network. The threshold of rejecting new calls in LTE and 5G is represented by $T_i$. New calls are rejected when the current bbu being used has exceeded the threshold, whereas handoff calls are rejected when all the available bbu are being used.

## 3.2 MARKOV MODEL FOR ARRIVAL SPLITTING

The call distribution model for arrival splitting described above can be modeled as a Markov chain where the current state of the heterogeneous system can be represented as follows

$$\Omega = (m_i, n_i : i = 1, 2) \tag{3.1}$$

The non-negative integers $m_i$ and $n_i$ represent the number of new and handoff calls in the network, respectively, where $i = 1$ represents the LTE network, and $i = 2$ represents the 5G network.

Let S denote the state space of all admissible calls in the network as it evolves over time. This state space S shows a combination of the number of new and handoff calls that can be admitted into the network without dispersing the bandwidth constraints in the network.

$$S = \left\{ \Omega(m_i, n_i : i = 1, 2) \,\middle|\, (m_1 + n_1) \cdot b_1 \leq C_1 \wedge (m_1) \cdot b_1 \leq T_1 \wedge (m_2 + n_2) \cdot b_2 \leq C_2 \wedge (m_2) \cdot b_2 \leq T_2 \right\} \tag{3.2}$$

Let $\rho_{\text{new}_i}$ and $\rho_{\text{han}_i}$ denote the load generated by new and handoff calls, respectively, in the LTE and 5G networks.

$$\rho_{\text{new}_i} = \frac{\lambda_{\text{new}_i}}{\mu_{\text{new}_i}} \tag{3.3}$$

$$\rho_{\text{han}_i} = \frac{\lambda_{\text{han}_i}}{\mu_{\text{han}_i}} \tag{3.4}$$

where the non-negative $\lambda_{\text{new}_i}$ and $\mu_{\text{new}_i}$ denotes the new call arrival rate and the new call departure rate in LTE or 5G depending on the value of $i$.

The probability of a network being in any particular state depends on the network load obtained by equation 3.3 and 3.4 which depends on the arrival and departure rates of calls

in the network. The performance measures of interest can be determined by summing up the appropriate state probabilities. Let $P(S)$ denote the probability of being in any state $s \in S$.

$$P(S) = \frac{1}{G} \prod_{i=1}^{2} \frac{(\rho_{\text{new}_i})^{m_i}}{m_i!} \frac{(\rho_{\text{han}_i})^{n_i}}{n_i!} \quad \forall s \in S \tag{3.5}$$

Where $G$ is the normalization constant given by:

$$G = \sum_{s \in S} \prod_{i=1}^{2} \frac{(\rho_{\text{new}_i})^{m_i}}{m_i!} \frac{(\rho_{\text{han}_i})^{n_i}}{n_i!} \tag{3.6}$$

### 3.2.1 Performance Evaluation

In the proposed model, the performance is analyzed by using the new call-blocking probability, handoff call-dropping probability, and the normalized average network utilization. These performance matrices are discussed in the following subsections.

#### New Call Blocking Probability

Since handoff calls are prioritized over new calls, in arrival splitting new calls are blocked when the networks do not have enough resources to accommodate new calls and when the resources used have exceeded the threshold for rejecting new calls in each network. Let $S_{b_i}$ denote the set of states in which new calls are blocked in both LTE and 5G in arrival splitting:

$$S_{b_i} = \{s \in S : (b_1 + (m_1 + n_1) \cdot b_1 > C_1) \vee ((b_1 + (m_1) \cdot b_1 > T_1) \wedge (b_2 + (m_2 + n_2) \cdot b_2 > C_2) \vee (b_2 + (m_2) \cdot b_2 > T_2))\} \tag{3.7}$$

Thus, the blocking probability is given by summing all the probabilities of being in any of the states described in equation 3.7. Therefore the equation for finding the new call blocking probability is given by:

$$p_{b_i} = \sum_{s \in S_{b_i}} P(s) \tag{3.8}$$

**Handoff Call Dropping Probability**

Handoff calls are dropped when all the resources or bandwidths used exceed the total capacity of each network. This means that the networks do not have enough bandwidth to accommodate handoff calls. Let $S_{d_i}$ denote the set of states in which the networks do not have enough bandwidth to accommodate incoming handoff calls in arrival splitting

$$S_{d_i} = \{s \in S : (b_1 + (m_1 + n_1) \cdot b_1 > C_1) \vee (b_2 + (m_2 + n_2) \cdot b_2 > C_2)\} \tag{3.9}$$

Thus, the dropping probability is given by summing all the probabilities of being in any of the states described in equation 3.9. Therefore the equation for finding the handoff call dropping probability is given by:

$$p_{d_i} = \sum_{s \in S_{d_i}} P(s) \tag{3.10}$$

**Network Utilization**

The average network utilization is the sum of the product of network utilization in a particular state and the probability of the network being in that particular state for all the admissible states in the network $s(s \in S)$. Thus the average network utilization(U) in arrival splitting is obtained by the following equation

$$U = \sum_{s \in S} U_s \cdot P(s) \tag{3.11}$$

where $U_s$ is the total amount of resources allocated to all calls admitted into the network at a particular state $S$. The normalized utilization is obtained by dividing the average utilization (U) by the total resources available in the networks.

## 3.3 MARKOV MODEL FOR SERVICE SPLITTING

The call distribution model for service splitting described above can be modeled as a Markov chain where current state of the heterogeneous system can be represented as follows

$$\Omega = ({m^j}_i, {n^j}_i : i = 1, 2; 1 \leq j \leq k-1) \tag{3.12}$$

The non-negative integers ${m^j}_i$ and ${n^j}_i$ represent the number of new and handoff calls in the network, respectively, where $i = 1$ represents the LTE network, $i = 2$ represents the 5G network and $j$ represents the basic bandwidth unit(bbu) for that call. k represents the basic bandwidth units required by each call.

Let S denote the state space of all admissible calls in the network as it evolves over time. Since the required bandwidth per call is split before distributing to LTE and 5G, this state space shows a combination of the portions of new and handoff calls that can be admitted into the networks without dispersing the bandwidth constraints.

$$\begin{aligned} S = \Bigg\{ & \Omega(m_i^j, n_i^j : i = 1, 2; 1 \leq j \leq k-1) \Bigg| \\ & \sum_{j=1}^{k-1} (m_1^j + n_1^j) b_1^j \leq C_1 \wedge \sum_{j=1}^{k-1} m_1^j b_1^j \leq T_1 \\ & \wedge \sum_{j=1}^{k-1} (m_2^{k-j} + n_2^{k-j}) b_2^{k-j} \leq C_2 \wedge \sum_{j=1}^{k-1} m_2^{k-j} b_2^{k-j} \leq T_2 \\ & \wedge \forall j : \left( m_1^j = m_2^{k-j} \wedge n_1^j = n_2^{k-j} \right) \Bigg\} \end{aligned} \tag{3.13}$$

Where ${b^j}_i$ and ${b^j}_2$ denotes portions of the bbu allocated to LTE and 5G respectively. $C_1$ and $T_1$ denotes the capacity and threshold for accepting new calls in LTE respectively. $C_2$ and $T_2$ denotes the capacity and threshold for accepting new calls in 5G respectively.

Let $\rho_{\text{new}_i}$ and $\rho_{\text{han}_i}$ denote the load generated by new and handoff calls, respectively, in the LTE and 5G networks

$$\rho_{new^j{}_i} = \frac{\lambda_{new^j{}_i}}{\mu_{new^j{}_i}} \quad \forall j = 1 \ldots k-1 \tag{3.14}$$

$$\rho_{han^j{}_i} = \frac{\lambda_{han^j{}_i}}{\mu_{han^j{}_i}} \quad \forall j = 1 \ldots k-1 \tag{3.15}$$

where $\lambda_{new^j{}_i}$ and $\mu_{new^j{}_i}$ are the new call arrival and departure rates for network $i$ with a bandwidth of $j$. $\lambda_{han^j{}_i}$ and $\mu_{han^j{}_i}$ are the handoff call arrival and departure rates for network $i$ with a bandwidth of $j$.

The probability of a network being in any particular state depends on the network load obtained by equation 3.14 and 3.15 which depends on the arrival and departure rates of calls in the network. The performance measures of interest can be determined by summing up the appropriate state probabilities. Let $P(S)$ denote the probability of being in any state $s \in S$.

$$P(S) = \frac{1}{G}\left(\prod_{j=1}^{k-1} \frac{(\rho_{new^j{}_i})^{m_i^j}}{m_i^j!}\frac{(\rho_{han^j{}_i})^{n_i^j}}{n_i^j!}\right)\left(\prod_{j=1}^{k-1} \frac{(\rho_{new^{k-j}{}_i})^{m_i^{k-j}}}{m_i^{k-j}!}\frac{(\rho_{han^{k-j}{}_i})^{n_i^{k-j}}}{n_i^{k-j}!}\right) \quad \forall s \in S, j = 1 \ldots k-1 \tag{3.16}$$

Where $G$ is the normalization constant given by:

$$G = \sum_{s \in S}\left(\prod_{j=1}^{k-1} \frac{(\rho_{new^j{}_i})^{m_i^j}}{m_i^j!}\frac{(\rho_{han^j{}_i})^{n_i^j}}{n_i^j!}\right)\left(\prod_{j=1}^{k-1} \frac{(\rho_{new^{k-j}{}_i})^{m_i^{k-j}}}{m_i^{k-j}!}\frac{(\rho_{han^{k-j}{}_i})^{n_i^{k-j}}}{n_i^{k-j}!}\right) \quad \forall j = 1 \ldots k-1 \tag{3.17}$$

### 3.3.1 Performance Evaluation

In the proposed model, the performance is analyzed by using the new call-blocking probability, handoff call-dropping probability, and average network utilization. These performance matrices are discussed in the following subsections.

**New Call Blocking Probability**

In service splitting new calls are blocked when the networks do not have enough resources to accommodate portions of new calls and when all the resources used have exceeded the threshold for rejecting new calls in each network. Let $S_{b_i}$ denote the set of states in which new calls are blocked in both LTE and 5G in service splitting:

$$S_{b_i} = \{s \in S \mid \Bigg( (b_1^v + \sum_{j=1}^{k-1}(m_1^j + n_1^j)b_1^j > C_1 \vee b_1^v + \sum_{j=1}^{k-1} m_1^j b_1^j > T_1) \\ \vee (b_2^{k-v} + \sum_{j=1}^{k-1}(m_2^{k-v} + n_2^{k-j})b_2^{k-j} > C_2 \vee b_2^{k-v} + \sum_{j=1}^{k-1} m_2^{k-j} b_2^{k-j} > T_2) \Bigg) \quad \forall v = 1 \dots k-1 \tag{3.18}$$

Thus, the blocking probability is given by summing all the probabilities of being in any of the states described in equation 3.18. Therefore the equation for finding the new call blocking probability is given by:

$$p_{b_i} = \sum_{s \in S_{b_i}} P(s) \tag{3.19}$$

**Handoff Call Dropping Probability**

Handoff calls are dropped when all the resources or bandwidths used have exceeded the total capacity of each network. This means that the networks do not have enough bandwidth to accommodate incoming portions of calls. Let $S_{d_i}$ denote the set of states in which handoff calls are dropped in both LTE and 5G in service splitting:

$$S_{d_i} = \{s \in S \mid \Bigg( b_1^v + \sum_{j=1}^{k-1}(m_1^j + n_1^j)b_1^j > C_1 \\ \vee b_2^{k-v} + \sum_{j=1}^{k-1}(m_2^{k-v} + n_2^{k-j})b_2^{k-j} > C_2 \Bigg) \quad \forall v = 1 \dots k-1\} \tag{3.20}$$

Thus, the dropping probability is given by summing all the probabilities of being in any

of the states described in equation 3.20. Therefore the equation for finding the handoff call dropping probability is given by:

$$p_{d_i} = \sum_{s \in S_{d_i}} P(s) \tag{3.21}$$

**Network Utilization**

The average network utilization is the sum of the product of network utilization in a particular state and the probability of the network being in that particular state for all the admissible states in the network $s(s \in S)$. Thus the average network utilization(U) in service splitting is obtained by the following equation

$$U = \sum_{s \in S} U_s \cdot P(s) \tag{3.22}$$

# Chapter 4

# Simulation Design

The algorithms discussed in the previous chapter were simulated on MATLAB Livescript to view their performance. In this chapter the performance of arrival splitting and service splitting are evaluated with respect to the probability of blocking new calls, the probability of dropping handoff calls, and the normalized network utilization. These performance matrices were simulated on different network conditions which will be detailed together with the discussion of results in the next subsections.

## 4.1 Overview of Codes

Each of the splitting mechanisms has its own code to simulate how it works and its performance. The code in Appendix A named "Arrival Splitting Simulation Code" executes the the performance of arrival splitting mechanism. The code named "Service Splitting Simulation Code" executes the performance of the service splitting mechanism. Both codes begin by setting network parameters and then comes the simulation loop which is the main execution loop that calls the function called "probability" for computing the probabilities and the average network utilization .

The code for service splitting differs from the code for arrival splitting in that in service splitting a function called "serviceSplit" that returns the divided bandwidths is implemented. The functions for calculating the probabilities are also different in the two codes as they execute different Markov models discussed in Chapter 3.

## 4.2 Running Codes

The MATLAB code for simulating the results when the new call arrival rate is changing can be found in Appendix A. Running the "Arrival Splitting simulation Code" gives the performance of arrival splitting and similarly running the "Service Splitting Code" gives the performance of service splitting. To get the plot with dropping and blocking probabilities for both splitting mechanisms, the "Probability Plotting Code" was run on the command window after running the above splitting codes. Similarly to get the combined utilization, the "Utilization Plotting Code" was run in the command window.

### 4.2.1 Effect of New Call Arrival Rate Simulation

The codes in Appendix A simulate the scenario where the new call arrival rate is changing. This is shown in Figure 4.1 where the new call arrival rate(ar_new) is being changed in the simulation loop and updating the arrays to store probabilities and utilization at each time.

```
%Probability
[PB, PD, Utilization] = probability();

PB_values_service = [PB_values_service, PB];
PD_values_service = [PD_values_service, PD];
ar_new_values = [ar_new_values, ar_new];

%Utilization of network
utilization_values_service = [utilization_values_service, Utilization];

% Changing parameters
ar_new = ar_new + 0.2;
```

Figure 4.1: Changing new call arrival rate in code

### 4.2.2 Effect of New Call Departure Rate Simulation

For simulating the performance when the new call departure rate is changing, the codes in Appendix A were altered in the simulation loops such as shown in Figure 4.2, where the new call departure rate(dr_new) is being changed in the simulation loop and updating the arrays to store probabilities and utilization at each time.

```
% Probabilities
[PB, PD, Utilization] = probability();

PB_values_arrival_changing_dr = [PB_values_arrival_changing_dr, PB];
PD_values_arrival_changing_dr = [PD_values_arrival_changing_dr, PD];
dr_new_values = [dr_new_values, dr_new];

% Normalized Network Utilization
utilization_values_arrival_changing_dr = [utilization_values_arrival_changing_dr, Utilization];

% Changing parameters with time
dr_new = dr_new + 0.2;
```

Figure 4.2: Changing new call departure rate in code

### 4.2.3 Effect Bandwidth Simulation

For simulating the performance when the required bandwidth per call is changing, the codes in Appendix A were altered in the simulation loops such as shown in Figure 4.3, where the bandwidth per call(b) is being changed in the simulation loop and updating the arrays to store probabilities and utilization at each time.

```
% Probabilities
[PB, PD, Utilization] = probability();

PB_values_arrival_bbu = [PB_values_arrival_bbu, PB];
PD_values_arrival_bbu = [PD_values_arrival_bbu, PD];
bbu_values = [bbu_values, b];

% Normalized Network Utilization
utilization_values_arrival_bbu = [utilization_values_arrival_bbu, Utilization];

% Changing parameters with time
b = b + 0.5;
```

Figure 4.3: Changing call bandwidth in code

### 4.2.4 Effect Threshold Simulation

For simulating the performance when the thresholds for rejecting new calls are changing, the codes in Appendix A were altered in the simulation loops such as shown in Figure 4.4, where the thresholds(T1 and T2) are being changed in the simulation loop and updating the arrays to store probabilities and utilization at each time.

```
% Probabilities
[PB, PD, Utilization] = probability();

PB_values_arrival_threshold = [PB_values_arrival_threshold, PB];
PD_values_arrival_threshold = [PD_values_arrival_threshold, PD];
T1_values = [T1_values, T1];

% Normalized Network Utilization
utilization_values_threshold = [utilization_values_threshold, Utilization];

% Changing parameters with time
T1 = T1 + 1;
T2 = T2 + 1;
```

Figure 4.4: Changing threshold in code

### 4.2.5 Effect Capacity Simulation

For simulating the performance when the capacities for rejecting new calls are changing, the codes in Appendix A were altered in the simulation loops such as shown in Figure 4.5, where the capacities(C1 and C2) are being changed in the simulation loop and updating the arrays to store probabilities and utilization at each time.

```
% Probabilities
[PB, PD, Utilization] = probability();

PB_values_arrival_capacity = [PB_values_arrival_capacity, PB];
PD_values_arrival_capacity = [PD_values_arrival_capacity, PD];
C1_values = [C1_values, C1];

% Normalized Network Utilization
utilization_values_arrival_capacity = [utilization_values_arrival_capacity, Utilization];

% Changing parameters with time
C1 = C1 + 0.5;
C2 = C2 + 0.5;
```

Figure 4.5: Changing capacity in code

# Chapter 5

# Simulation Results and Discussion

The above-mentioned codes in Chapter 5 were used in all simulation scenarios but with different network parameters that will be indicated in the following subsections. In the figures of results, $PB_A$, $PD_A$, $PB_S$, and $PD_S$ are the blocking probability in arrival splitting, dropping probability in arrival splitting, blocking probability in service splitting, and dropping probability in service splitting respectively. $U_A$ and $U_S$ represent the network utilization in arrival splitting and service splitting respectively.

## 5.1 Effect of New Call Arrival Rate

To determine how the algorithm of service splitting performs in comparison with the arrival splitting, the blocking and dropping probabilities as well as network utilization were plotted with respect to varying arrival rates of new calls. The new call arrival rate is varied from 1 call per second to 5 calls per second. Table 5.1 shows the network parameters that were assumed and used in this simulation. It is assumed that the bandwidth required per call is 4 bbu and it does not change as there is only one type of call (video call) that is being evaluated.

Table 5.1: Network parameters when new call arrival rate changes

| Parameter | LTE | 5G |
|---|---|---|
| Capacity(bbu) | 15 | 30 |
| Threshold(bbu) | 10 | 20 |
| New call arrival rate | 1 to 5 | 1 to 5 |
| New call departure rate | 0.5 | 0.5 |
| Handoff call arrival rate | 1 | 1 |
| Handoff call departure rate | 0.5 | 0.5 |

### 5.1.1 Call Blocking and Dropping Probability

Figure 5.1 illustrates how the blocking and dropping probability change when the new call arrival rate is increased in both arrival splitting and service splitting. When the call arrival rate increases in the network and the departure rate is kept constant, the probability of blocking and dropping calls increases due to an increase in network load. As shown in Figure 5.1, the dropping and dropping probabilities in service splitting are always lower than the corresponding probabilities in arrival splitting for network parameters specified in Table 5.1.

When calls arrive in arrival splitting, their bandwidth does not get split such as what happens in service splitting, only the quantity gets split among the two networks. When the network is highly loaded and the remaining resources in each network cannot accommodate a call alone, it will be blocked or dropped and this is evidenced by the blocking probability which is approaching 1 in a case of arrival splitting, and the dropping probability which is also high.

Since calls are split in service splitting, the chances for a call to arrive in a network and not get the resources are low as the residual bbu are combined to accommodate a call. By strategically splitting the call bandwidth and assigning them to two networks simultaneously, the chances of blocking or dropping a call are lower and this is evidenced by Figure 5.1 which shows lower probabilities for service splitting and higher probabilities for arrival splitting.

Moreover, the dropping probabilities are always lower than the blocking probabilities because of the priority that is given to handoff calls over new calls by setting rejection thresholds for accepting new calls.

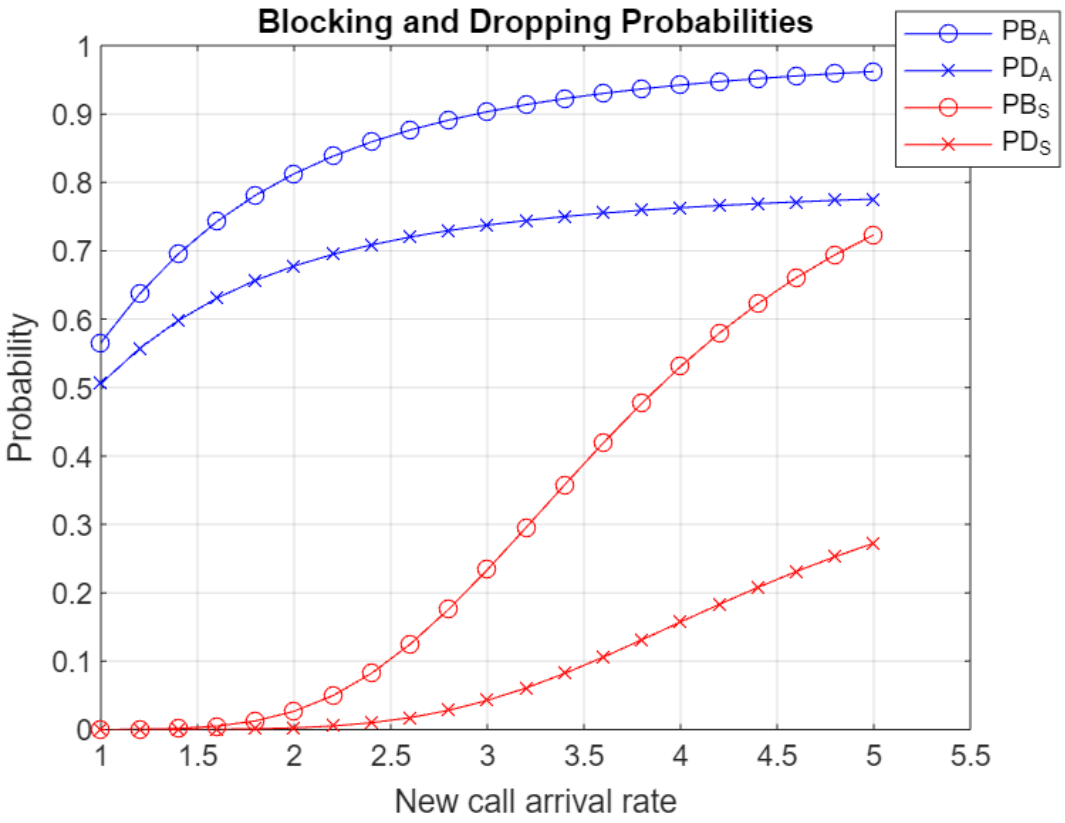


Figure 5.1: Blocking and dropping probability when changing the new call arrival rate

### 5.1.2 Network Utilization

In Figure 5.2 $U_A$ abd and $U_S$ illustrate the normalized network utilization in arrival splitting and service splitting respectively. Initially, when the new call arrival rate is low, the network utilization in arrival splitting is higher than the network utilization in service splitting. This is because when calls arrive in a network where there is no splitting of call bandwidth, they tend to utilize more of the resources in each network than when there is splitting of call bandwidths.

As the new call arrival rate increases, the network utilization in case of service splitting increases faster than in case of arrival splitting. This is because of the increase in network load, which is causing some of the calls to be blocked/dropped when there is not enough residual bbu to accommodate them in arrival splitting, leaving the remaining bbu in each network not being utilized as they are not enough to handle a call. On the other side, in service splitting, the residual bbus of each network are combined to accommodate the calls which ensure that most resources are utilized.

For higher new call arrival rates, the network utilization in service splitting is higher than the network utilization in arrival splitting. This is shown in figure 5.2.

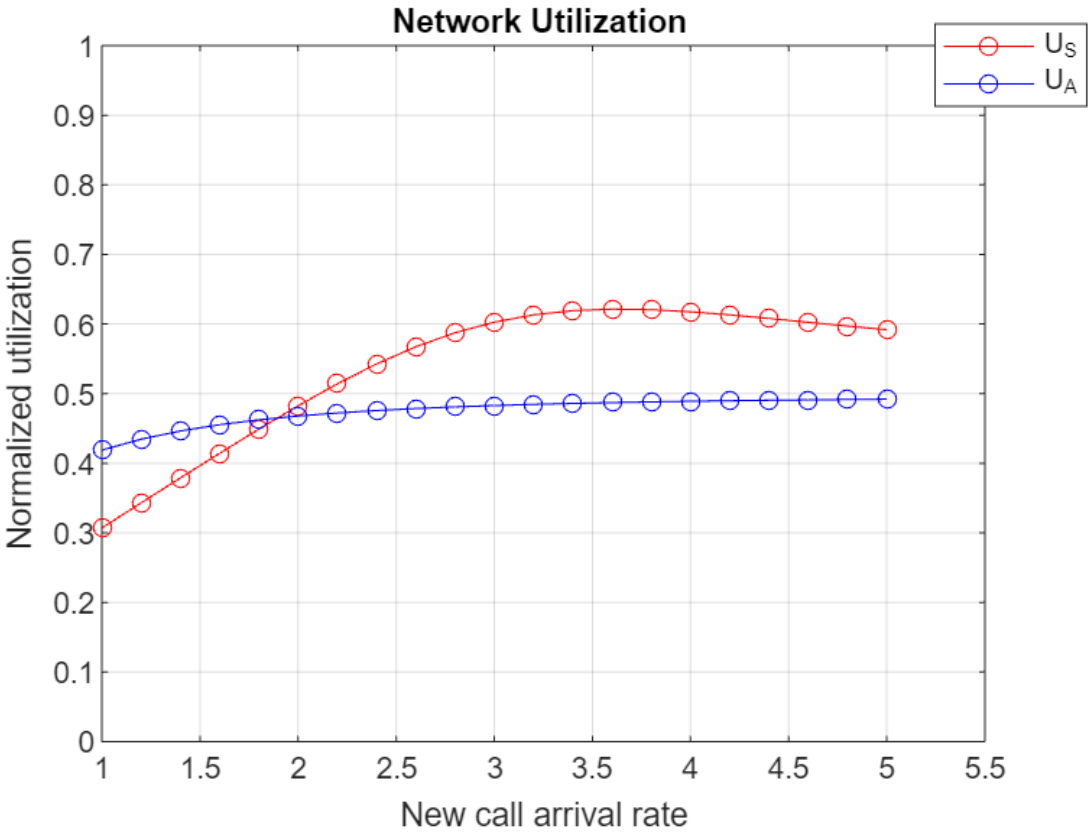


Figure 5.2: Network utilization when changing the new call arrival rate

## 5.2 Effect of New Call Departure Rate

In this scenario, the new call departure rate is varied from 0.5 to 4.5 to see the performance of splitting mechanisms when the new call departure rate is increasing. The performance matrices are still the blocking probability of new calls, dropping probability of handoff calls, and network utilization but now they are plotted with respect to the changing new call departure rate. The bandwidth required per call is 4 bbu and it is the same for both new and handoff calls. Table 5.2 shows the network parameters that were assumed in this simulation.

Table 5.2: Network parameters when new call departure rate changes

| Parameter | LTE | 5G |
|---|---|---|
| Capacity(bbu) | 15 | 30 |
| Threshold(bbu) | 10 | 20 |
| New call arrival rate | 12 | 12 |
| New call departure rate | 0.5 to 4.5 | 0.5 to 4.5 |
| Handoff call arrival rate | 1 | 1 |
| Handoff call departure rate | 0.5 | 0.5 |

### 5.2.1 Call Blocking and Dropping Probability

Figure 5.3 illustrates the blocking and dropping probability of calls when changing the new call departure rate. When the new call departure rate increases it means that more calls will be leaving the network or getting serviced more quickly. As a result, this reduces the probability of blocking new calls in both arrival and service splitting. Similarly, the probability of dropping handoff calls decreases in both arrival and service splitting. This is shown in Figure 5.3.

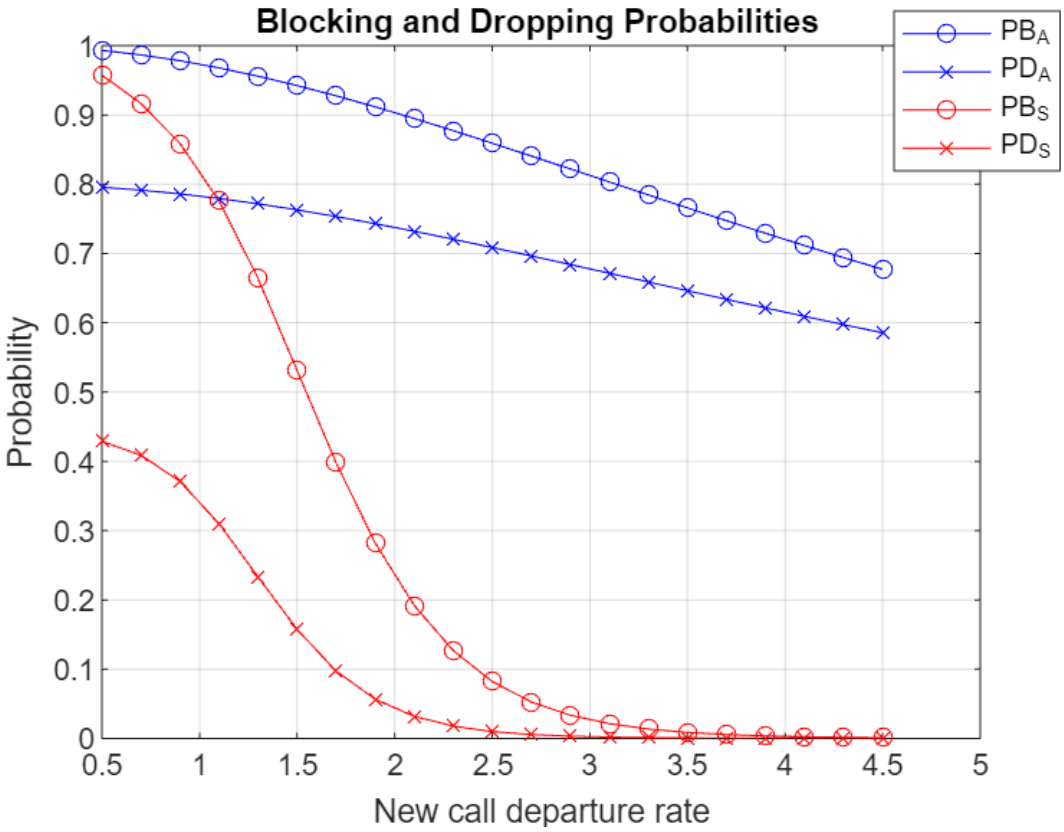


Figure 5.3: Blocking and dropping probability when changing the new call departure rate

For the whole range of new call departure rates evaluated, the blocking and dropping probabilities in service splitting remain lower than the corresponding probabilities in arrival splitting and as the departure rate increases, the probabilities in service splitting decrease faster than in arrival splitting. For higher departure rates, the blocking and dropping probabilities in service splitting tend to be zero, this is because in service splitting, the load is split among the LTE and 5G which makes it faster for the network to process the calls as they have small bandwidth. Moreover, the handoff dropping probabilities remain lower than the new call blocking probabilities.

### 5.2.2 Network Utilization

Figure 5.4 illustrates the normalized network utilization when the new call departure rate is changing. The overall network utilization is decreasing as the new call departure rate increases for all splitting mechanisms. This is because the rate at which calls leave the network is increasing which leaves some of the resources unused when no new or handoff call has arrived.

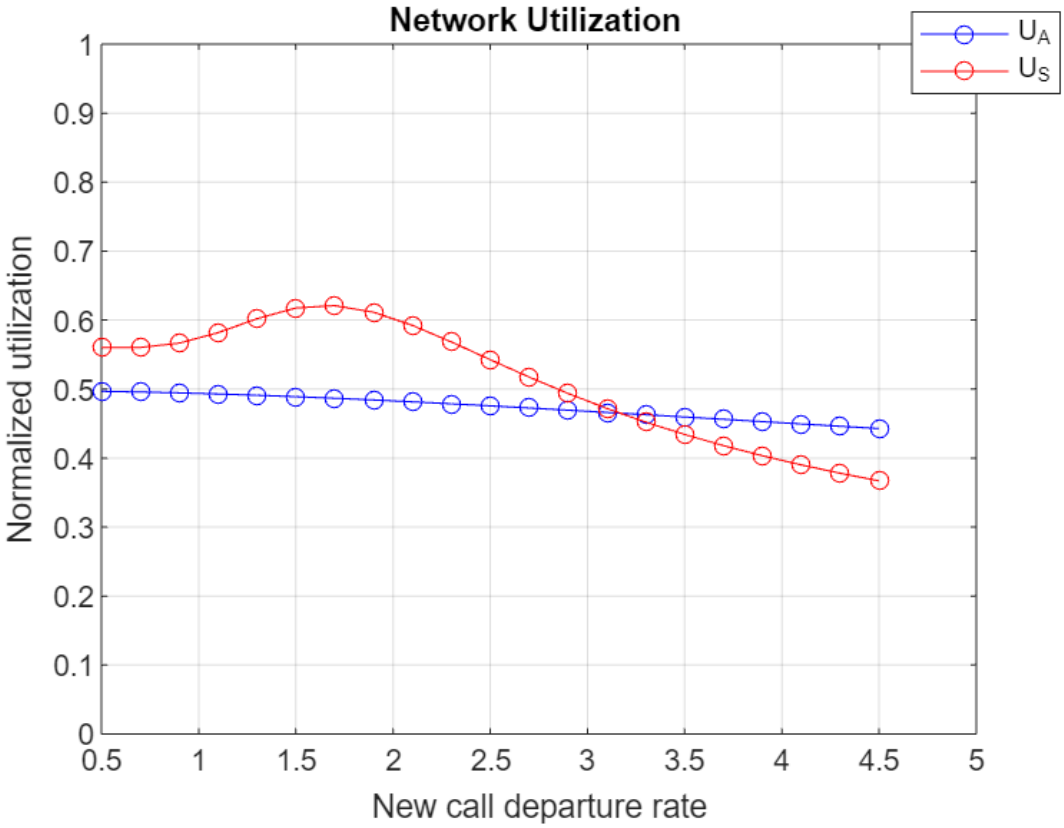


Figure 5.4: Network utilization when changing the new call departure rate

Initially, at a low departure rate, the network utilization in service splitting is higher than the network utilization in arrival splitting. As the departure rate increases, the network utilization in service splitting decreases faster than in arrival splitting. As shown in Figure 5.4, when the departure rate is at 4.5, the network utilization in service splitting is lower than the network utilization in arrival splitting, this is because it takes less time to transmit calls that have low divided bandwidth in service splitting than calls that have larger bandwidth in arrival splitting. Leaving more resources unused in service splitting.

## 5.3 Effect of Threshold

In this scenario, the thresholds of both LTE and 5G are varied from 6 to 15 bbu to evaluate how effective the splitting mechanisms are when the threshold for rejecting new calls is increasing. Similarly, the performance is evaluated by plotting the probability of blocking new calls, the probability of dropping handoff calls, and network utilization but now with respect to changing the threshold of rejecting new calls. The bandwidth rewuired per call is 4 bbu. Table 5.3 shows the network parameters that were assumed in this simulation.

Table 5.3: Network parameters when the threshold changes

| Parameter | LTE | 5G |
|---|---|---|
| Capacity(bbu) | 20 | 30 |
| Threshold(bbu) | 6 to 15 | 6 to 15 |
| New call arrival rate | 2 | 2 |
| New call departure rate | 0.5 | 0.5 |
| Handoff call arrival rate | 4 | 4 |
| Handoff call departure rate | 0.5 | 0.5 |

### 5.3.1 Call Blocking and Dropping Probability

Figure 5.5 shows how the blocking and dropping probability changes when the threshold for rejecting new calls is increasing in both arrival and service splitting. An increased threshold means that the network will accept more new calls and that is reducing the probability of blocking new calls. This increases the handoff call dropping probability since more bandwidth will be occupied by new calls. This is evidenced by Figure 5.5. The dropping and blocking probabilities in service splitting remain lower than the corresponding probabilities in arrival splitting. Furthermore, the handoff call dropping probability is always lower than the new call blocking probability.

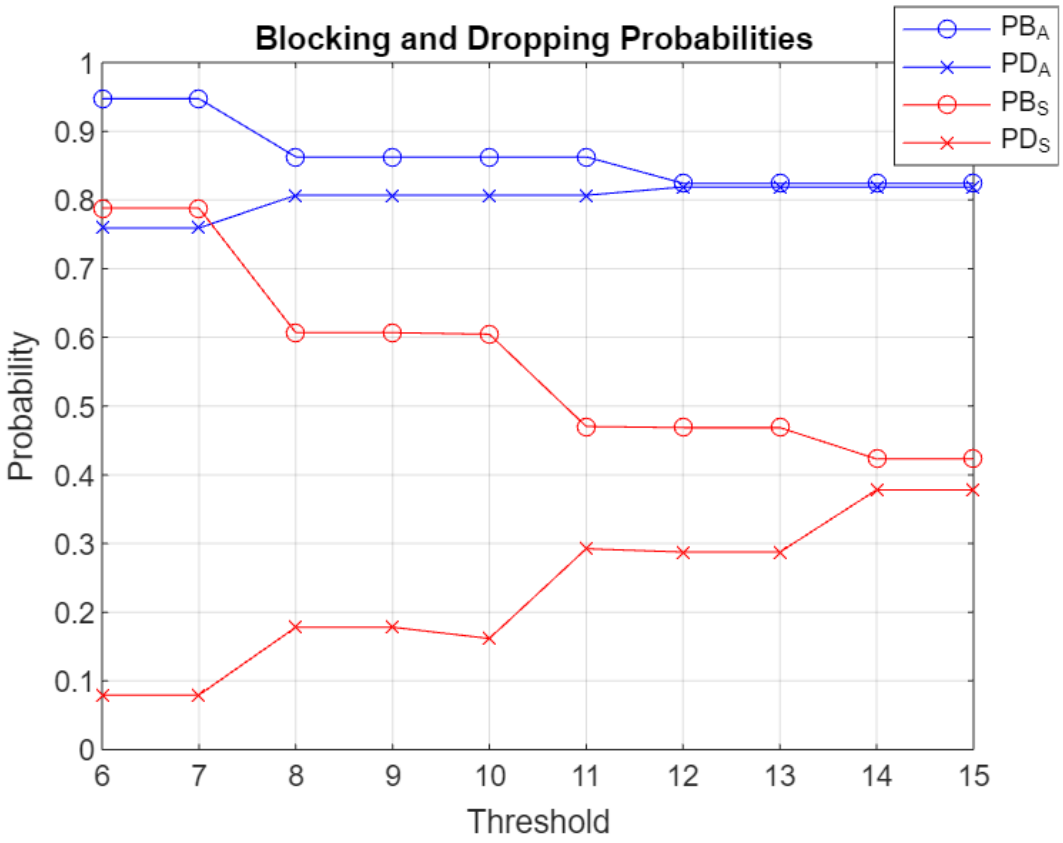


Figure 5.5: Blocking and dropping probability when changing the threshold

### 5.3.2 Network Utilization

Figure 5.6 shows the normalized network utilization in both arrival and service splitting when changing the threshold for accepting new calls. Initially, when the threshold is low, the network utilization for service splitting is lower than the network utilization in arrival splitting. As the threshold increases, the utilization in service splitting increases but in arrival splitting it remains constant, this is because the bandwidth left in the arrival splitting networks cannot accommodate incoming calls, but in service splitting it can accommodate incoming calls. When the threshold gets higher, the network in service splitting can accommodate incoming calls which causes the utilization to be higher than in arrival splitting.

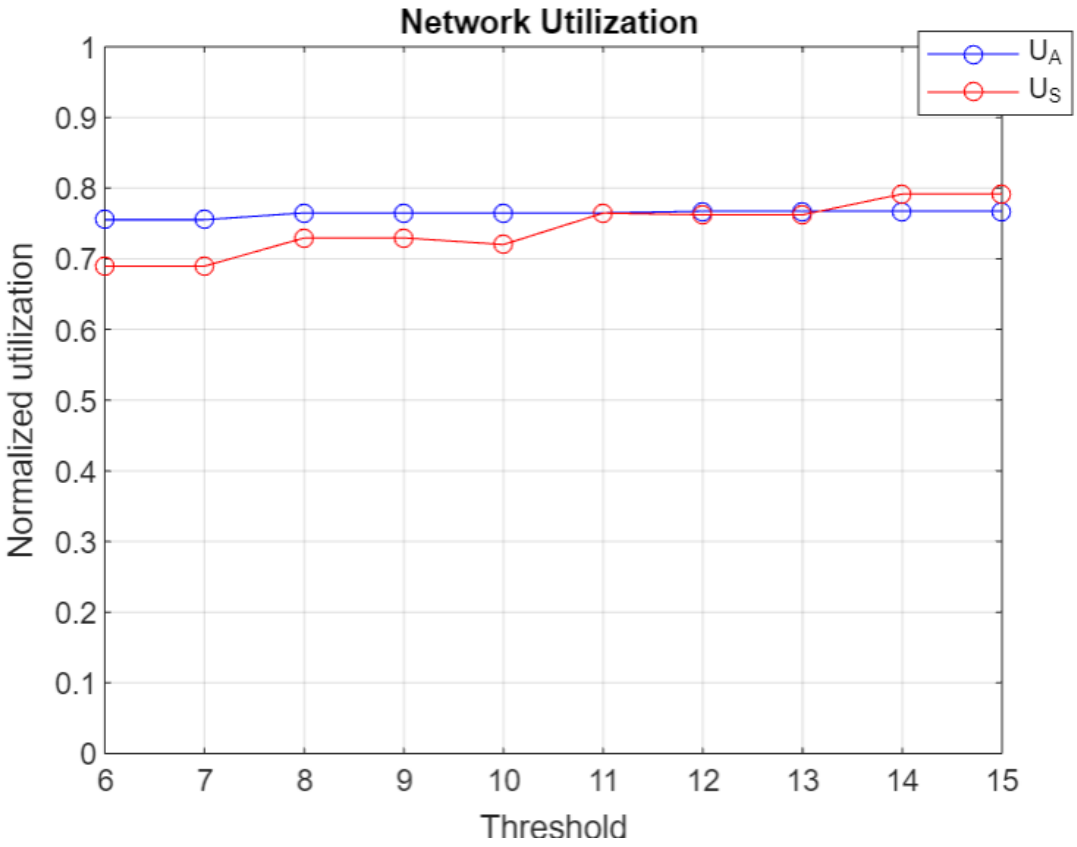


Figure 5.6: Network utilization when changing the threshold

## 5.4 Effect of Bandwidth Per Call

In this scenario, the bandwidth required per call is varied from 1 bbu to 11 bbu to see how effective the splitting mechanisms are when the required call bandwidth is increasing. Similarly, the performance is evaluated by plotting the probability of blocking new calls, the probability of dropping handoff calls, and network utilization but now with respect to changing bandwidth per call. Table 5.4 shows the network parameters that were assumed in this simulation.

Table 5.4: Network parameters when the bandwidth per call changes

| Parameter | LTE | 5G |
|---|---|---|
| Capacity(bbu) | 15 | 30 |
| Threshold(bbu) | 10 | 20 |
| New call arrival rate | 1 | 1 |
| New call departure rate | 0.5 | 0.5 |
| Handoff call arrival rate | 1 | 1 |
| Handoff call departure rate | 0.5 | 0.5 |

### 5.4.1 Call Blocking and Dropping Probability

Figure 5.7 shows how the blocking and dropping probability changes when the bandwidth per call is increased in both arrival splitting and service splitting. When the bandwidth per call increases, it means calls will require more bandwidth in the network. The network becomes resource-scarce and unable to allocate bandwidth to incoming calls which increases the blocking and dropping probability.

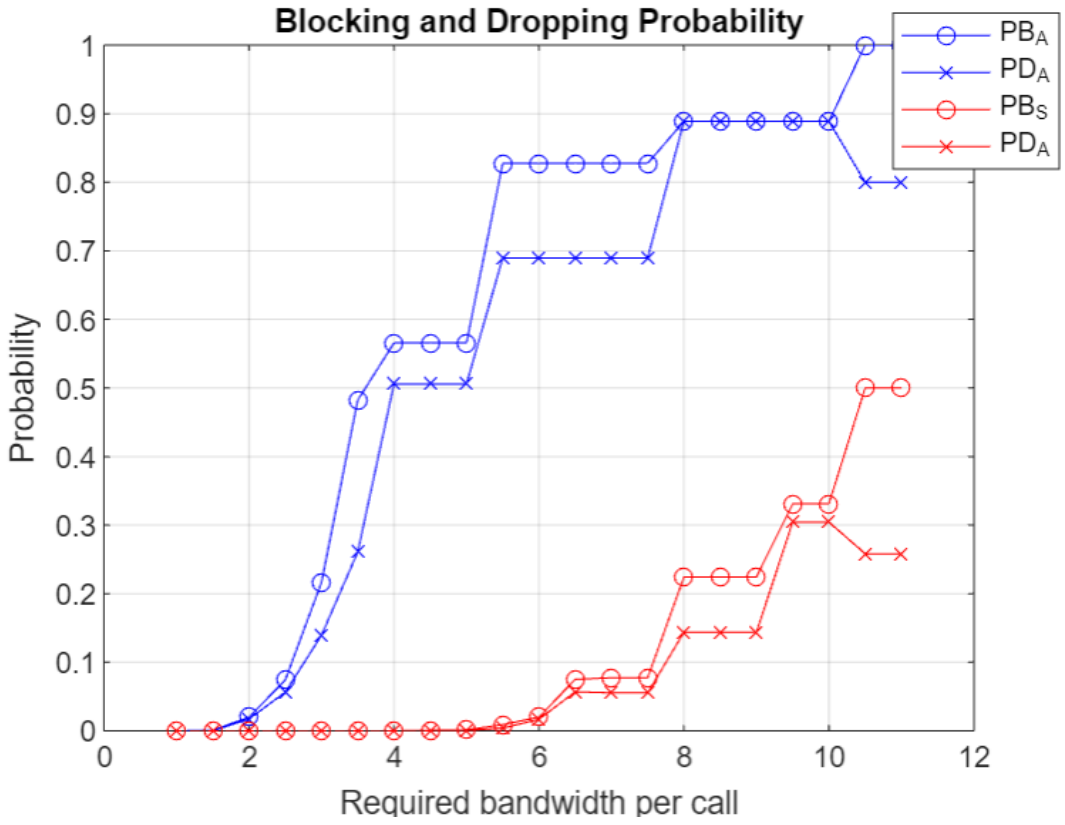


Figure 5.7: Blocking and dropping probability when changing the bandwidth per call

Both blocking and dropping probabilities in service splitting are lower than the corresponding probabilities in arrival splitting. When they increase with an increase in the call bandwidth, the probabilities in service splitting increase at a slower rate than in arrival splitting.

### 5.4.2 Network Utilization

Figure 5.8 shows the normalized network utilization in arrival splitting and service splitting when changing the required bandwidth per call. Initially, at low bandwidth per call, the network utilization in service splitting is lower than network utilization in arrival splitting. This is because when calls arrive in a network where there is arrival splitting,

the bandwidths of each call do not get split and they tend to utilize more bandwidth in both LTE and 5G. On the other side, the utilization in arrival splitting is initially low due to the splitting of bandwidths in each call.

As the bandwidth per call increases, the network utilization in arrival splitting reaches a point where it can not increase further but rather fluctuates. This is a result of when calls require a very large amount of bandwidth than the bandwidth left in LTE and 5G. There will be more bandwidth left in the networks but they will not be used as they cannot accommodate large calls and thus reduce the network utilization. In contrast, service splitting shows better resource utilization.

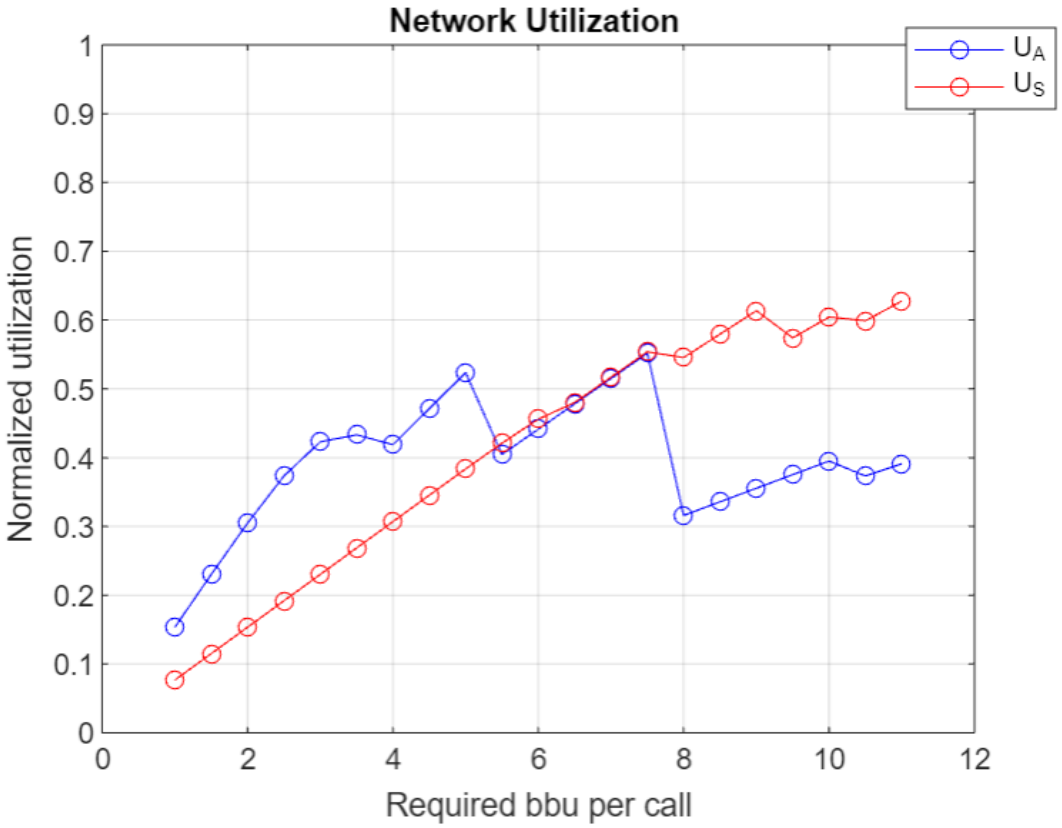


Figure 5.8: Network utilization when changing the bandwidth per call

## 5.5 Effect of Capacity

In this scenario, the capacities of both LTE and 5G are varied from 20 to 30 bbus to evaluate how the splitting mechanisms perform when the capacity is increasing. The performance is evaluated by plotting the probability of blocking new calls, and the probability of dropping handoff calls but now with respect to changing the capacity of the networks. Table 5.5 shows the network parameters that were assumed in this simulation.

Table 5.5: Network parameters when the capacity changes

| Parameter | LTE | 5G |
|---|---|---|
| Capacity(bbu) | 20 to 30 | 20 to 30 |
| Threshold(bbu) | 10 | 10 |
| New call arrival rate | 1 | 1 |
| New call departure rate | 0.5 | 0.5 |
| Handoff call arrival rate | 2 | 2 |
| Handoff call departure rate | 0.5 | 0.5 |

### 5.5.1 Call Blocking and Dropping Probability

Figure 5.9 shows how the blocking and dropping probability changes when the capacity is increasing in both arrival splitting and service splitting. Increasing capacity means there will be more bandwidth to accommodate incoming calls, therefore reducing traffic in the networks.

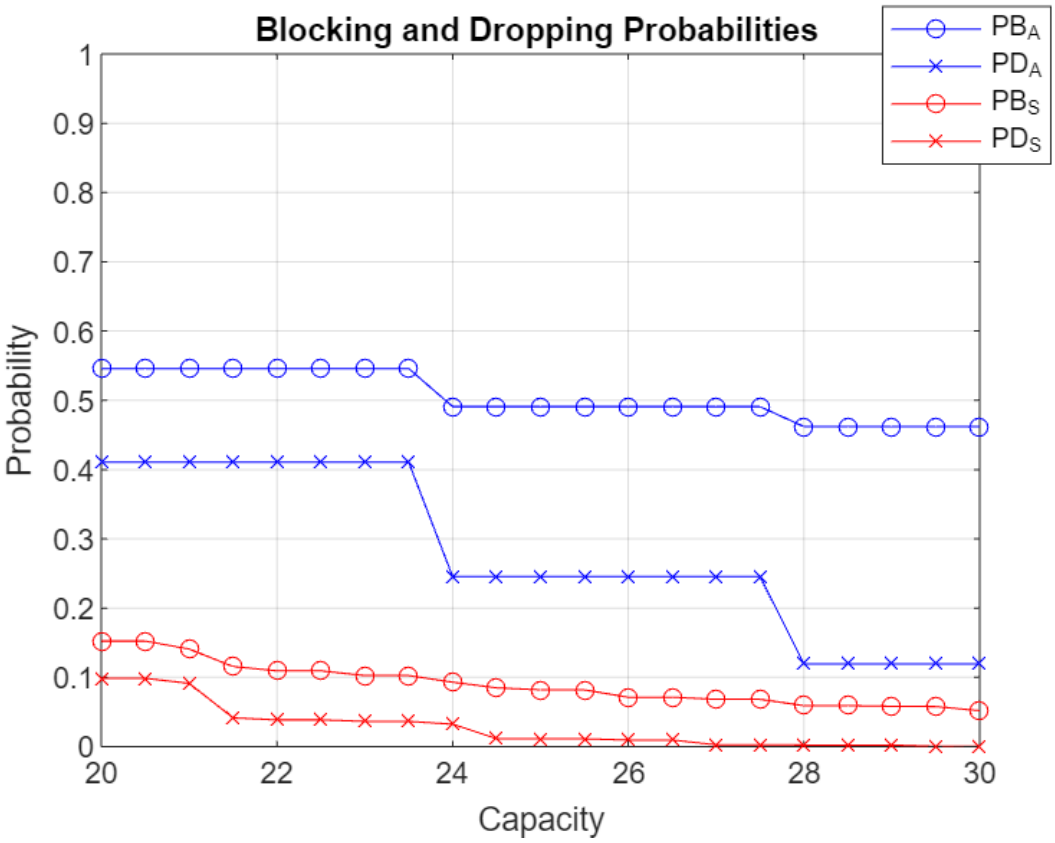


Figure 5.9: Blocking and dropping probability when changing the capacity

As shown in Figure 5.9, when increasing the capacity, both blocking and dropping probabilities decrease in both arrival and service splitting mechanisms. It can be seen that the blocking and dropping probabilities for service splitting are always less than the corresponding probabilities in arrival splitting. Moreover dropping probabilities are always less than blocking probabilities.

# Chapter 6

# Conclusions

The primary aim of this study was to design a load-balancing algorithm for multipath video streaming in heterogeneous wireless networks. This was based on the mobile terminal side of the network since on the network side, service providers can transmit as much streaming data as possible depending on how they are financed. This was also motivated by the availability of multi-homed mobile terminals with the ability to connect to multiple networks simultaneously, hence this study was specifically tailored for LTE-5G Dual Connectivity.

To achieve this goal, two kinds of video splitting mechanisms were implemented, namely arrival splitting and service(session) splitting. In arrival splitting the quantity of incoming calls is split among the two networks. In service splitting, the bandwidth of each call is split so that one portion of the call is transmitted over LTE and the other over 5G. Arrival splitting was used as a benchmark in this study to compare the performance of service splitting which is the main mechanism to achieve load balancing.

To measure the performance of these mechanisms, Markov models have been developed for each mechanism to investigate load balancing on the probability of blocking new calls, the probability of dropping handoff calls, and average resource utilization in the heterogeneous network. This was simulated on MATLAB, changing the network parameters to evaluate how the implemented service splitting mechanism would perform in comparison with the benchmarking arrival splitting mechanism.

Simulation results show that when increasing the new call arrival rate, both probabilities of dropping and blocking calls increase, however, the probabilities in service splitting are lower than the corresponding probabilities in arrival splitting. When increasing the new

call departure rate, the blocking and dropping probabilities decrease with the probabilities of service splitting remaining lower than the probabilities of arrival splitting.

In the scenario of increasing the threshold for accepting new calls, the blocking probabilities were decreasing with increasing threshold and the dropping probabilities were increasing with increasing threshold. When increasing the required bandwidth per call, the blocking and dropping probabilities were also increasing. However, the dropping and blocking probabilities in service splitting were still lower than the corresponding probabilities in arrival splitting in both scenarios.

Furthermore, service splitting showed better network utilization than arrival splitting, especially when the new call arrival rate and the required bandwidth per call were high which resembles a highly loaded network. The network utilization results evidenced that in service splitting, calls are serviced faster than in arrival splitting as more calls leave the network in service splitting than in arrival splitting.

The designed service splitting mechanism outperforms the benchmarking arrival splitting mechanism in every scenario of varying the network parameters. Therefore it can be concluded that the service splitting mechanism achieves better load balancing in heterogeneous networks. consequently, it can be used in balancing load in multipath video streaming.

# Chapter 7

# Recommendations

In the course of this research project, a load balancing algorithm for multipath video streaming has been designed and its performance has been analyzed. It showed a better performance when compared to a benchmark scenario however there were assumptions made which can be explored in future works. Hence, the following are recommendations made for future work:

- This thesis assumes that incoming video calls have the same bandwidth requirement, future works should investigate a network where calls have different bandwidths as it symbolizes different kinds of videos.
- Future work should consider an algorithm that takes into account the residual capacities in the networks so that in cases where one network is fully loaded, splitting will not allowed, and the total bandwidth of a call will be serviced by a network that has enough remaining capacity.

# Bibliography


[1] M. Mohsin, “10 video marketing statistics you should know in 2023 [infographic],” Oberlo, https://www.oberlo.com/blog/video-marketing-statistics (accessed Oct. 3, 2023).

[2] L. Ceci and S. 5, “Top video content type by Global Reach 2023,” Statista, https://www.statista.com/statistics/1254810/top-video-content-type-by-global-reach/ (accessed Oct. 3, 2023).

[3] S. Popoola, N. Faruk, A. Atayero, M. Adeyeye Oshin, O. Bello, and E. Mutafungwa, ”Radio Access Technologies for Sustainable Deployment of 5G Networks in Emerging Markets,” International Journal of Applied Engineering Research, vol. 12, pp. 14154-14172, 2017.

[4] A. Hodroj, M. Ibrahim and Y. Hadjadj-Aoul, ”A Survey on Video Streaming in Multipath and Multihomed Overlay Networks,” in IEEE Access, vol. 9, pp. 66816-66828, 2021, doi: 10.1109/ACCESS.2021.3076464.

[5] M. U. Sheikh, M. Z. Asghar and R. Jäntti, ”Dual Connectivity in Non-Stand Alone Deployment mode of 5G in Manhattan Environment,” 2020 International Conference on Electronics, Information, and Communication (ICEIC), Barcelona, Spain, 2020, pp. 1-4, doi: 10.1109/ICEIC49074.2020.9051202.

[6] W. Ding, P. Ren and Q. Du, ”Fountain Code Transmission in Dual Connectivity Based on Partial Overlapped Data,” 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), Xi’an, China, 2019, pp. 1-6, doi: 10.1109/WCSP.2019.8928084.

[7] S. Khan, I. Shayea, M. Ergen, and H. Mohamad, ”Handover management over dual connectivity in 5G technology with future ultra-dense mobile heterogeneous networks: A review,” Engineering Science and Technology, an International Journal, vol. 35, pp. 101172, 2022. DOI: 10.1016/j.jestch.2022.101172.

[8] J. Sun, S. Zhang, S. Xu and S. Cao, "High Throughput and Low Complexity Traffic Splitting Mechanism for 5G Non-Stand Alone Dual Connectivity Transmission," in IEEE Access, vol. 9, pp. 65162-65172, 2021, doi: 10.1109/ACCESS.2021.3076301.

[9] G. S. Park and H. Song, "Video Quality-Aware Traffic Offloading System for Video Streaming Services Over 5G Networks With Dual Connectivity," in IEEE Transactions on Vehicular Technology, vol. 68, no. 6, pp. 5928-5943, June 2019, doi: 10.1109/TVT.2019.2909547.

[10] S. Dubey and J. Meena, "Improvement of Throughput using Dual Connectivity in Non-Standalone 5G NR Networks," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 6-12, doi: 10.1109/ICSSIT48917.2020.9214179.

[11] H. Cui and Y. Du, "Joint user association and fair scheduling for dual connectivity heterogeneous networks," in China Communications, vol. 20, no. 1, pp. 171-183, Jan. 2023, doi: 10.23919/JCC.2023.01.014.

[12] E. F. Olabisi, "Joint call admission control algorithm for reducing call blocking/dropping probability in heterogeneous wireless networks supporting multihoming," 2010 IEEE Globecom Workshops, Miami, FL, USA, 2010, pp. 611-615, doi: 10.1109/GLOCOMW.2010.5700393.

[13] F. Bhering, D. Passos, C. Albuquerque and K. Obraczka, "Efficient Multipath Selection for IoT Video Transmission," 2022 IEEE 11th International Conference on Cloud Networking (CloudNet), Paris, France, 2022, pp. 73-78, doi: 10.1109/CloudNet55617.2022.9978786.

[14] F. Chahlaoui, H. Dahmouni and H. El Alami, "Multipath-routing based load-balancing in SDN networks," 2022 5th Conference on Cloud and Internet of Things (CIoT), Marrakech, Morocco, 2022, pp. 180-185, doi: 10.1109/CIoT53061.2022.9766801.

[15] Y. Xia, J. Wu, J. Xia, T. Wang and S. Mao, "Multipath-aware TCP for Data Center Traffic Load-balancing," 2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS), Tokyo, Japan, 2021, pp. 1-6, doi: 10.1109/IWQOS52092.2021.9521276.

[16] W. Wu, L. Jiang, C. He, D. He and J. Zhang, "RavenFlow: Congestion-Aware Load Balancing in 5G Base Station Network," 2020 IEEE International Symposium on Circuits and Systems (ISCAS), Seville, Spain, 2020, pp. 1-5, doi: 10.1109/ISCAS45731.2020.9180892.

[17] A. Giuseppi, S. Maaz Shahid, E. De Santis, S. Ho Won, S. Kwon and T. Choi, "Design and Simulation of the Multi-RAT Load-balancing Algorithms for 5G-ALLSTAR Systems," 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea (South), 2020, pp. 594-596, doi: 10.1109/ICTC49870.2020.9289485.

[18] R. Deng, "Resource Allocation for Multipath Cooperative Video Transmission over 5G Networks," 2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Xi'an, China, 2021, pp. 1-6, doi: 10.1109/ICSPCC52875.2021.9564508.

# Appendix A

# Simulation Codes

# Arrival Splitting Simulation code

```
% network parameters and variables
global C1 C2 T1 T2 b;
C1 = 15; T1 = 10;
C2 = 30; T2 = 20;
b = 4; % Required bandwidth for each call
global b1 b2;
b1 = b; b2 =b;
% Simulation parameters
observation_time = 21; % Simulation time in seconds
ar_new = 1; % New call arrival rate per second
dr_new = 0.5; % New call deparure rate per second
ar_han = 1; % Handoff call arrival rate per second
dr_han = 0.5; % Handoff call departure rate per second
global p_new p_han;
% Storing results in arrays for plotting
PB_values_arrival = [];
PD_values_arrival = [];
ar_new_values = [];
utilization_values_arrival = [];
% Main simulation loop
for time = 1:observation_time
 p_new = ar_new/dr_new;
 p_han = ar_han/dr_han;

 % Probabilities
 [PB, PD, Utilization] = probability();
 PB_values_arrival = [PB_values_arrival, PB];
 PD_values_arrival = [PD_values_arrival, PD];
 ar_new_values = [ar_new_values, ar_new];
 % Normalized Network Utilization
 utilization_values_arrival = [utilization_values_arrival, Utilization];
 % Changing parameters with time
 ar_new = ar_new + 0.2;

end
% Visualization of blocking Probability on arrival splitting
figure;
plot(ar_new_values, PB_values_arrival, '-ob');
xlabel('New call arrival rate ');
ylabel('Probability');
title('Probability of blocking new calls in arrival splitting');
grid on;
ylim([0, 1]);
% Visualization of dropping Probability on arrival splitting
figure;
plot(ar_new_values, PD_values_arrival, '-o');
xlabel('New call arrival rate ');
```

```
ylabel('Probability');
title('Probability of dropping handoff calls in arrival splitting');
grid on;
ylim([0, 1]);
% Normalised network utilization on arrival splitting
figure;
plot(ar_new_values, utilization_values_arrival, '-or');
xlabel('New call arrival rate ');
ylabel('Normalized utilization');
title('Normalized network utilization in arrival splitting');
grid on;
ylim([0, 1]);
```

```
function [PB, PD, Utilization] = probability()
global C1 C2 T1 T2 b1 b2 p_new p_han;
SB =0; SD =0; G =0;
U = 0;
for m2=0:T2/b2
 for n2=0:C2/b2
 m1 = round(C1/C2)*m2;
 n1 = round(C1/C2)*n2;
 if (((b1*(m1+n1)<=C1) && (b1*m1<=T1))) && ...
(((b2*(m2+n2)<=C2)&&(b2*m2<=T2)))
 P= (((p_new^m1)*(p_han^n1))/(factorial(m1) * ...
factorial(n1))) * (((p_new^m2)*(p_han^n2))/(factorial(m2) * factorial(n2)));
 G = G + P;
%Sum of the probabilities for blocking states
if ((b1+ (b1*(m1+n1)) > C1) || ((b1+ (b1*m1)) > T1)) ||...
((b2+ (b2*(m2+n2)) > C2) || ((b2+ (b2*m2)) > T2))
 SB=SB+P;
end
%Sum of the probabilities for dropping states
if (b1+ (b1*(m1+n1)) > C1) || (b2+ (b2*(m2+n2)) > C2)
 SD=SD+P;
end
% average Utilization
Us = (m1+n1)*b1 + (m2+n2)*b2;
 U = U + (Us*P);
 states=[(m1),(n1),(m2),(n2)];
 end
 end
end
PB = SB/G;
PD = SD/G;
Utilization = (U/G)/(C1+C2);
end
```

# Service Splitting Simulation Code

```
% Service Splitting Simulation Code


% Network parameters and variables
global C1 C2 T1 T2 b;
C1 = 15;
T1 = 10;
C2 = 30;
T2 = 20;
b = 4; % Required bandwidth for each call
global b11 b12 b13 b21 b22 b23;


% Simulation parameters
observation_time = 21; % Simulation time in seconds
ar_new = 1; % New call arrival rate per second
dr_new = 0.5; % New call departure rate per second
ar_han = 1; % Handoff call arrival rate per second
dr_han = 0.5; % Handoff call departure rate per second
global p_new p_han;


PB_values_service = [];
PD_values_service = [];
ar_new_values = [];
utilization_values_service = [];


% Main simulation loop
for time = 1:observation_time
    p_new = ar_new/dr_new;
    p_han = ar_han/dr_han;

    % Probability
    [PB, PD, Utilization] = probability();


    PB_values_service = [PB_values_service, PB];
    PD_values_service = [PD_values_service, PD];
    ar_new_values = [ar_new_values, ar_new];


    % Utilization of the network
    utilization_values_service = [utilization_values_service, Utilization];


    % Changing parameters
    ar_new = ar_new + 0.2;
```

```
end


% Visualization of blocking Probability on service splitting
figure;
plot(ar_new_values, PB_values_service, '-ob');
xlabel('New call arrival rate ');
ylabel('Probability');
title('Probability of blocking new calls in service splitting');
grid on;
ylim([0, 1]);


% Visualization of dropping Probability on service splitting
figure;
plot(ar_new_values, PD_values_service, '-o');
xlabel('New call arrival rate ');
ylabel('Probability');
title('Probability of dropping handoff calls in service splitting');
grid on;
ylim([0, 1]);


% Normalized network utilization on service splitting
figure;
plot(ar_new_values, utilization_values_service, '-or');
xlabel('New call arrival rate ');
ylabel('Normalized utilization');
title('Normalized network utilization in service splitting');
grid on;
ylim([0, 1]);
```

```
function [b11, b12, b13, b21, b22, b23] = ...
    serviceSplit(m11, m12, m13, n11, n12, n13, m21, m22, m23, n21, n22, n23)
    global C1 C2 T1 T2 b;


    j = C1 / (C1 + C2);
    if (j <= 1/3)
        b11 = (1/3) * b;%Split according to the ratio of residual capacities
        b23 = b - b11;
        b12 = 0;
        b13 = 0;
        b21 = 0;
        b22 = 0;
        if (m21 * b21 + m22 * b22 + m23 * b23) > T2 || (m21 * b21 +...
                m22 * b22+ m23 * b23 + n21 * b21 + n22 * b22 + n23 * b23)>C2
            % Try other splitting combinations
            b12 = 0.5 * b;
```

```
            b22 = 0.5 * b;
            b11 = 0;
            b13 = 0;
            b21 = 0;
            b23 = 0;
            if (m21 * b21 + m22 *b22 + m23 *b23) >T2|| (m21 * b21 + m22...
                    * b22 + m23*b23 + n21*b21 + n22*b22 + n23 * b23) > C2
                % Try other splitting combinations
                b13 = (2/3) * b;
                b21 = b - b13;
                b11 = 0;
                b12 = 0;
                b22 = 0;
                b23 = 0;
            end
        end
    elseif (j > 1/3) && (j < 2/3)
        % Split according to the ratio of residual capacities
        b12 = 0.5 * b;
        b22 = 0.5 * b;
        b11 = 0;
        b13 = 0;
        b21 = 0;
        b23 = 0;
        if (m21 *b21 + m22 *b22 + m23 *b23)>T2 ||(m21 * b21 + m22 * b22...
                + m23 * b23 + n21 * b21 + n22 * b22 + n23 * b23) > C2
            % Try other splitting combinations
            b13 = (2/3) * b;
            b21 = b - b13;
            b12 = 0;
            b11 = 0;
            b23 = 0;
            b22 = 0;
        elseif (m11 *b11 + m12 *b12 + m13 * b13)>T1||(m11 * b11 + m12...
                * b12 + m13 *b13 + n11 *b11 + n12 *b12 + n13 *n13) > C1
            % Try other splitting combinations
            b11 = (1/3) * b;
            b23 = b - b11;
            b12 = 0;
            b13 = 0;
            b21 = 0;
            b22 = 0;
        end
    elseif (j >= 2/3)
        % Split according to the ratio of residual capacities
        b13 = (2/3) * b;
        b21 = b - b13;
        b11 = 0;
        b12 = 0;
        b22 = 0;
        b23 = 0;
```

```
        if (m11 * b11 + m12 * b12 + m13 * b13)>T1||(m11 * b11 + m12 *...
                b12 + m13 *b13 + n11 *b11 + n12 *b12 + n13 * n13) > C1
            % Try other splitting combinations
            b12 = 0.5 * b;
            b22 = 0.5 * b;
            b11 = 0;
            b13 = 0;
            b21 = 0;
            b23 = 0;
            if (m11 * b11 + m12 * b12 + m13 * b13) >T1||(m11 * b11 + m12...
                    * b12 + m13*b13 + n11*b11 + n12*b12 + n13*n13) > C1
                % Try other splitting combinations
                b11 = (1/3) * b;
                b23 = b - b11;
                b12 = 0;
                b13 = 0;
                b21 = 0;
                b22 = 0;
            end
        end
    else
        b11 = 0;
        b12 = 0;
        b13 = 0;
        b21 = 0;
        b22 = 0;
        b23 = 0;
    end
end


function [PB, PD, Utilization] = probability()
global C1 T1 C2 T2;
global p_new p_han;


SB = 0;
SD = 0;
G = 0;
U = 0;
for m11 = 0:T1
    for n11 = 0:C1
        for m12 = 0:T1/2
            for n12 = 0:C1/2
                for m13 = 0:T1/3
                    for n13 = 0:C1/3
                        m23 = m11;
                        m22 = m12;
                        m21 = m13;
                        n23 = n11;
                        n22 = n12;
                        n21 = n13;
```

```
[b11, b12, b13, b21, b22, b23] = ...
    serviceSplit(m11, m12, m13, n11, n12,...
    n13, m21,m22, m23, n21, n22, n23);


condition1 = (((m11 + n11) * b11 + (m12 + n12) *...
    b12 + (m12 + n13) * b13) <= C1) &&...
    ((m11 * b11 + m12 * b12 + m13 * b13) <= T1);
condition2 = (((m21 + n21) * b21 + (m22 + n22) ...
    * b22 + (m22 + n23) * b23) <= C2) && ...
    ((m21 * b21 + m22 * b22 + m23 * b23) <= T2);
if (condition1 && condition2)
    P1 = (((p_new^m11)*(p_han^n11)/(factorial(m11)...
        * factorial(n11)))*((p_new^m12)*(p_han^n12)...
        /(factorial(m12)*factorial(n12)))* ...
        ((p_new^m13)*(p_han^n13)/(factorial(m13)* ...
        factorial(n13))));
    P2 = (((p_new^m21)*(p_han^n21)/(factorial(m21)...
        *factorial(n21)))*((p_new^m22)*(p_han^n22)...
        / (factorial(m22)*factorial(n22))) * ...
        ((p_new^m23)*(p_han^n23)/(factorial(m23) *...
        factorial(n23))));
    P = P1 * P2;
    G = G + P;


    % Sum of the probabilities for blocking states
    b_condition1 =((b11+(m11+n11)*b11+(m12+n12)...
        * b12 + (m13 + n13) * b13) > C1) || ...
        ((b11 +m11*b11+m12 *b12 + m13*b13)>T1)|| ...
        ((b23 +(m21+n21)*b21+(m22+n22) * b22...
        + (m23 + n23) * b23) > C2)|| ...
        ((b23 + m21*b21 +m22 *b22+m23*b23)>T2);


    b_condition2 =((b12+(m11+n11)*...
        b11+(m12+n12)*b12+...
        (m13+n13)*b13)>C1)||...
        ((b12+m11*b11+m12*b12+m13*b13)...
        > T1) |||((b22+(m21+n21)*b21...
        +(m22+n22)*b22+(m23+n23)*b23)>C2)...
        || ((b22+m21*b21 + m22*b22 +...
        m23*b23) > T2);


    b_condition3 = ((b13+(m11+n11)*...
        b11+(m12+n12)*b12 +(m13+n13)...
        *b13)>C1) || ((b13+m11*b11 +...
        m12*b12 + m13*b13) > T1) || ...
        ((b21+(m21+n21)*b21 +(m22+n22)*b22...
        +(m23+n23)*b23)>C2) ...
        || ((b21+m21*b21 + m22*b22 +...
```

```
                                m23*b23) > T2);


                            if (b_condition1||b_condition2 ||...
                                    b_condition3)
                                SB=SB+P;
                            end


                            %Sum of the probabilities for dropping states
                            d_condition1=((b11+(m11+n11)*b11+...
                                (m12+n12)*b12+...
                                (m13+n13)*b13)>C1)||((b23+(m21+n21)*b21 +...
                                (m22+n22)*b22 +(m23+n23)*b23)>C2);
                            d_condition2=((b12+(m11+n11)*b11+...
                                (m12+n12)*b12 +...
                                (m13+n13)*b13)>C1) || ...
                                ((b22+(m21+n21)*b21 +...
                                (m22+n22)*b22 +(m23+n23)*b23)>C2);
                            d_condition3=((b13+(m11+n11)*b11 ...
                                +(m12+n12)*b12 +...
                                (m13+n13)*b13)>C1) || ...
                                ((b21+(m21+n21)*b21 +...
                                (m22+n22)*b22 +(m23+n23)*b23)>C2);


                            if( d_condition1 || d_condition2 || d_condition3)
                                SD=SD+P;
                            end
                            %Utilization
                            Us=(m11+n11)*b11+(m12+n12)*b12+(m13+n13)*b13 +...
                                (m21+n21)*b21+(m22+n22)*b22+(m23+n23)*b23;
                            U = U + (Us*P);


                        end
                    end
                end
            end
        end
    end
end
PB = SB/G;
PD = SD/G;
Utilization =(U/G)/(C1+C2);
end
```

## Probability Plotting Code

```
figure;
plot(ar_new_values, PB_values_arrival, '-ob', 'DisplayName', 'PB_A');
hold on;
plot(ar_new_values, PD_values_arrival, '-xb', 'DisplayName', 'PD_A');
plot(ar_new_values, PB_values_service, '-oc', 'DisplayName', 'PB_S');
plot(ar_new_values, PD_values_service, '-xc', 'DisplayName', 'PD_S');
xlabel('New call arrival rate');
ylabel('Probability');
title('Blocking and Dropping Probabilities');
grid on;
ylim([0, 1]);
legend('Location', 'NorthWest');
```

## Utilization Plotting Code

```
figure;
plot(ar_new_values, utilization_values_service, '-or', 'DisplayName', 'U_S');
hold on;
plot(ar_new_values, utilization_values_arrival, '-ob', 'DisplayName', 'U_A');
xlabel('New call arrival rate');
ylabel('Normalized utilization');
title('Network Utilization ');
grid on;
ylim([0, 1]);
legend('Location', 'NorthWest');
```

# Appendix B

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

# Ethics clearance

## PRE-SCREENING QUESTIONNAIRE OUTCOME LETTER

STU-EBE-2023-PSQ000556

2023/08/06

Dear Mutikedzi Mudzanani,

Your Ethics pre-screening questionnaire (PSQ) has been evaluated by your departmental ethics representative. Based on the information supplied in your PSQ, it has been determined that you do not need to make a full ethics application for the research project in question.

You may proceed with your research project titled:

Design of a Load-balancing Algorithm for Multipath Video Streaming in Heterogeneous Wireless Networks

Please note that should aspect(s) of your current project change, you should submit a new PSQ in order to determine whether the changed aspects increase the ethical risks of your project. It may be the case that project changes could require a full ethics application and review process.

Regards,

Faculty Research Ethics Committee