QUEUEING BASED RESOURCE ALLOCATION IN COGNITIVE RADIO NETWORKS

by

Hilary Mutsawashe Tsimba

Submitted in partial fulfillment of the requirements for the degree
Master of Engineering (Electronic Engineering)

in the

Department of Electrical, Electronic and Computer Engineering
Faculty of Engineering, Built Environment and Information Technology

UNIVERSITY OF PRETORIA

June 2017

SUMMARY

QUEUEING BASED RESOURCE ALLOCATION IN COGNITIVE RADIO NETWORKS

by

Hilary Mutsawashe Tsimba

Supervisor(s):	Prof. B.T. Maharaj
Co-Supervisor:	Prof. A.S. Alfa
Department:	Electrical, Electronic and Computer Engineering
University:	University of Pretoria
Degree:	Master of Engineering (Electronic Engineering)
Keywords:	Cognitive Radio Networks, Queueing Theory, Resource Allocation, Optimisation

With the increase in wireless technology devices and mobile users, wireless radio spectrum is coming under strain. Networks are becoming more and more congested and free usable spectrum is running out. This creates a resource allocation problem. The resource, wireless spectrum, needs to be allocated to users in a manner such that it is utilised efficiently and fairly.

The objective of this research is to find a solution to the resource allocation problem in radio networks, i.e to increase the efficiency of spectrum utilisation by making maximum use of the spectrum that is currently available through taking advantage of co-existence and exploiting interference limits. The solution proposed entails adding more secondary users (SU) on a cognitive radio network (CRN) and having them transmit simultaneously with the primary user. A typical network layout was defined for the scenario.

The interference temperature limit (ITL) was exploited to allow multiple SUs to share capacity. Weighting was applied to the SUs and was based on allowable transmission power under the ITL. Thus a more highly weighted SU will be allowed to transmit at more power. The weighting can be determined by some network-defined rule. Specific models that define the behaviour of the network were then

developed using queuing theory, specifically weighted processor sharing techniques. Optimisation was finally applied to the models to maximize system performance. Convex optimization was deployed to minimize the length of the queue through the power allocation ratio.

The system was simulated and results for the system performance obtained. Firstly, the performance of the proposed models under the processor-sharing techniques was determined and discussed, with explanations given. Then optimisation was applied to the processor-sharing results and the performance was measured. In addition, the system performance was compared to other existing solutions that were deemed closest to the proposed models.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to the following for their invaluable assistance during the course of this research:

- My supervisor, Prof Maharaj for his guidance, insight and excellent mentorship.
- My co-supervisor, Prof Alfa for the technical leadership, insight and indispensable advise.
- The Sentech Chair in Broadband Wireless Multimedia Communications (BWMC) and Mastercard for both the resources and financial support required.
- All my friends in the BWMC group who provided much needed discussions and solutions from the beginning.
- My parents, Christopher and Irene, my brother Clive for without their never ending support I could not have achieved this.

LIST OF ABBREVIATIONS

CRN	Cognitive Radio Network
D2D	Device to Device
DSA	Dynamic Spectrum Access
ECC	Electronic Communications Code
FCC	Federal Communications Commission
FDD	Frequency Division Duplex
FSA	Fixed Spectrum Access
GSM	Global System for Mobile Communication
IoT	Internet of Things
ILP	Integer Linear Programming
ISM	Industrial Scientific and Medical band
ITU	International Telecommunications Union
MILP	Mixed Integer Linear Programming
NLP	Non-linear Programming
Ofcom	Office of Communications
PS	Processor Sharing
PU	Primary User
QoS	Quality of Service
QBD	Quasi-birth-death
SDR	Software Defined Radio
SU	Secondary User
TDD	Time Division Duplex
UHF	Ultra-high Frequency
VHF	Very High Frequency
WS	White Space
WSD	White Space Device

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

1.1 PROBLEM STATEMENT

1.1.1 Context of the problem

Wireless mobile technology has over the past years experienced massive growth in terms of use and innovation. The world's population continues to grow and so does the demand for information. Every day people seek to be connected, to communicate and to absorb information. People want to do this fast and reliably and also while mobile. Radio frequency spectrum, the medium that carries all wireless communication, is finite in its usable region. The increase in the number of users is leading to congestion and dropped connections in the networks. This is only the beginning, as the number of mobile internet users continues to grow worldwide. According to the international telecommunications union (ITU), there was a 12.5% percent increase in active mobile broadband users from 2014 to 2016 [1].

Research has shown that regardless of the high number of users and services requiring spectrum access but being refused, spectrum usage is still under-utilised. This is due to the licensing structure of radio spectrum. The challenge that confronts the wireless communication sector is to find an efficient method of managing spectrum already in use, that is, increasing the network utilisation to allow more users to connect and communicate with no dropped connections and little to no congestion. However, this is easier said than done, as there are many aspects to be considered. Some of these are transmission power and interference. For example, two transmitters and two receivers operating on the same frequency channel will interfere with each other unless there are measures in place to prevent this.

Certain users, called primary users (PUs), pay for the right to have exclusive use of spectrum. This means that even though the PUs are not using the particular channel at a particular time no-one else is allowed to use it. The result is that there will be idle channels while there are users waiting with no access.

A solution must be found to allow users to use radio spectrum efficiently through increasing spectrum utilisation.

1.1.2 Research gap

Cognitive radio networks (CRNs) have been proposed as a solution to the increase in network users [2]. CRNs add a secondary user (SU) to the network that can and may opportunistically access the network when it is not in use or when access is not harmful to the incumbent user. Therefore, when a PU is not using its licensed channel, an SU may occupy that channel and transmit on it until the PU returns. Research has been done on many ways this can be achieved. One such way is the use of queueing theory. Queueing theory has seen wide use in CRN as a technique to implement and evaluate the performance of many schemes targeted at increasing spectrum utilisation. However, not all methods are successful and some provide better results than others given a particular situation. Co-existence has been studied before, even in queueing theory through the use of priority systems. This paves the way for more spectrum use but improvements can still be made.

One aspect of queueing theory that has not yet been exploited in CRN extensively is processor-sharing queues. These queues provide the potential for improving on already existing methods. Co-existence can be implemented by adapting the processor sharing and through certain rules, high spectrum efficiency can be achieved. A gap therefore exists in this area, in terms of processor sharing and it is the purpose of this research to propose, develop and evaluate the benefits of such queues in terms of solving the resource allocation problem at hand.

1.2 RESEARCH OBJECTIVE AND QUESTIONS

A network with two more SUs is going to experience interference problems. The following questions were investigated:

- Is it possible for two or more secondary users to transmit efficiently on the same channel simultaneously with the primary users?
- Which is the appropriate domain on which to model the queue?

For the first question, transmission power limit is the concern here. If too many SUs are present there might be too much interference with the licensed user. The SUs also have to meet a minimum transmission power level to ensure the quality of their communication is adequate. Transmission power based resource allocation has been extensively researched, therefore it can be extended to the scenario of multiple SUs in the CRN [3, 4]. The CRN will have to determine how to allocate power to the SUs adequately. Priority queue models have been proposed to enable more than one SUs to transmit on the same channel [5, 6]. Two types of priority queues that have been proposed are pre-emptive priority discipline queues and non-pre-emptive priority discipline queues. For the pre-emptive discipline, the SUs are divided into several classes with different priorities. A higher class can interrupt the transmission of a lower queue. The lower priority queue will not be allowed to transmit anything until there are no more high-priority packets in the CRN [7]. For the non-pre-emptive discipline, the higher-priority queue will not pre-empt the service of the lower priority but rather wait until the current service has been completed before beginning transmission. The use of priority queues makes it possible for multiple requirements to be accommodated when allocating spectrum. A two-part call level queue that implements a direct Markov model has also been proposed [8]. Here, the SU queue is separated into two parts. In the first part SUs will, on interruption, save packets in a buffer until the licensed user leaves. In the second part, the SUs will discard the current packets on interruption. The results have shown that, for the delay queue, the throughput and packet loss rate increases as the delay increases and the length of the queue increases. The delay queue length does not have any effect on spectrum utilisation. The solution increases the number of SUs in the system, but, spectrum is still under-utilised.

The SUs' transmission power is limited by the interference limit of the PU, according to equation 1.1 [9],

$$P_s = \frac{Q}{|h_p|^2} \tag{1.1}$$

where h_p is the channel coefficient of the SU to PU link and Q is the power constraint imposed by the PU. The transmit power of the SU is given by:

$$P_s = min\{P_{max}, \frac{Q}{|h_p|^2}\} \tag{1.2}$$

where P_{max} is the maximum transmission power of the SU device.

For the second question, discrete time and continuous time models have been used extensively to model queue behaviour. Discrete models are also being used today in telecommunication systems. The analysis is done per time slot and therefore discrete. However, communication happens in real time and time is continuous. In some cases, and for some models, discrete time models are easier to analyse than continuous time models and vice versa.

1.3 APPROACH

If the SUs' requirements can be reclassified as a transmission power problem, then the SUs can be divided into separate classes according to their requirements. A queue model can then be developed to facilitate spectrum sharing. One widely used queue system in CRNs is the M/G/1 queue [9]. The system was constrained by a maximum power limit defined by the interference on the PU. The service time was also defined to be equal to the channel capacity according to Shannon's theorem. The channel chosen was a Nakagami-m fading channel. The resulting model had an embedded discrete time Markov chain and following this, performance measures such as packet transmission time, blocking probability, throughput, channel utilisation, mean number of packets and system time were deduced analytically.

The PUs are commonly associated with On/Off behaviour meaning the PU is either present or absent. There is no other state for the PU. Performance measures for the queues can be evaluated. Therefore with some modification an adequate queue model can be developed to fit at least two SUs into an underlay network.

The following steps were taken during the research:

- Extensive literature study was done on queue modelling and priority queueing.
- Mathematical derivations of the system constraints were obtained.
- Queue model dynamics and rules were defined.
- A numerical Markov chain model was developed.
- Sufficient software simulations were carried out.

- Lastly, system performance was measured from the simulations and comparisons made, enabling conclusions to be drawn.

1.4 RESEARCH GOALS

The research goals are as follows:

- To reclassify the SUs requirements as a transmission power problem then divide the SUs' separate classes according to their requirements.
- To make use of queueing theory, specifically, weighted processor sharing, to develop rules and model a CRN that allows at least two SU to occupy a single channel together with the PU and transmit simultaneously.
- To optimise the system in order achieve the highest possible performance.
- To obtain results and compare them with existing solutions in order to establish if the developed system does indeed improve spectrum efficiency.

1.5 RESEARCH CONTRIBUTION

1.5.1 Technical contributions

The following are the major technical contributions of the research work:

1. Pre-emptive processor-sharing model

   A model was developed that allows a higher priority queue to pre-empt the service of a lower priority queue. In this work, the state space model was developed and the transition matrix was generated. The Rate matrix structure was determined and compared to the closest already existing models.

2. Non-pre-emptive processor-sharing model

   Here, a model was developed that denies a higher priority queue the ability to pre-empt the service of a lower priority queue. Again, state space model was developed and the transition matrix was generated. The Rate matrix structure was determined and compared to already existing models.

3. Optimisation

   Optimisation was undertaken by first studying the pseudo-convexity of the objective function. From there, convex programming was applied and an optimal point was could be found that achieved the best network performance results.

Overall, a model was developed that shows the potential to increase spectrum utilisation by allowing more than one SU to be active in an underlay CRN. The results indicate that the performance measures have been improved significantly. This alone indicates the possibility that no more extra spectrum is required to cater for the rapid growth in communications technology and devices.

1.5.2 Paper contributions

The following paper, based on the work presented here, has been published in peer-reviewed international conference proceedings:

H.M. Tsimba, B.T. Maharaj and A.S. Alfa, "Increased Spectrum Utilisation in a Cognitive Radio Network: An M/M/1-PS Queue Approach", *in Proc. IEEE 2017 Wireless Communications and Networking Conference*, March 2017.

The following paper, based on the work presented here, has been submitted to a peer-reviewed journal:

H.M. Tsimba, B.T. Maharaj and A.S. Alfa, "Optimal Spectrum Utilisation in Cognitive Radio Networks Based on Processor Sharing Techniques", *IEEE Transactions on Wireless Communications*, May 2017, in review.

1.6 OVERVIEW OF STUDY

This study deals extensively with aspects of radio wireless technology. Firstly, extensive background literature on radio communication and networking is given in Chapter 2. In this chapter past, present and future prospects of wireless radio communication are discussed in minute detail. The problems currently faced by wireless networks are identified and the challenges faced by the proposed solutions are discussed in detail. CRNs are discussed in terms of their origin, developments and possible future advancements. CRN technology is the backbone of this research and a fair amount of Chapter 2 is dedicated to its discussion. The second major topic of this study is queueing theory. Chapter 2 also includes an introduction to queue theory and its link from mathematics to telecommunications. Specifically discussed are models involving two or more queues that are assigned some priority or are at least sharing one resource (radio channel). The elementary queue with Poisson arrival, exponential service and a single server is taken advantage of because it provides enough theoretical background to compare to the simulation results. M/M/1 priority queue models have also been extensively studied and therefore are also compared to the proposed model in order to determine the performance improvements.

Chapter 3 contains the various methods that were used to develop the proposed model. This includes the model specifics and explanation. The mathematical background to the quality of service requirement is given and assumptions made are given and explained. The power allocation method is also discussed here. The proposed queueing models are also presented. The queue behaviour is analysed and the transition matrices developed. The rate matrices are developed and the special structure due to the channel share scheme is presented. The chapter also explains the optimisation process. It gives a view of convex optimisation, why and how convexity was determined.

Chapter 4 presents the simulation results obtained from the proposed system. These are then compared with some theoretical results where possible. The results are extensive and try to showcase the different scenarios and variables of which the proposed models are capable. An example is the effect of changing the power allocation ratio. This can in some circumstances lead to drastic deterioration of network quality and in some cases it has little effect. Optimisation results are also presented here.

Chapter 5 presents the discussion of the results of the proposed model. The discussion seeks to explain why and how the results are the way they are and also any future improvements that can be made.

Chapter 6 is a conclusion of the work done. It presents the overall contribution of the work. Ideas on how to further this research are also discussed briefly.

CHAPTER 2 LITERATURE STUDY

2.1 CHAPTER OBJECTIVES

The aim of this chapter is to present a thorough background to the knowledge required for this research. An introduction into general wireless communication and where the research problem lies is given. Cognitive radio technology is also introduced. Existing solutions are given to identify the research gaps that exist and to find possible improvements. The background to queueing theory is extensively discussed. In brief, the objective of the chapter is to show that wireless spectrum is running out and that although some solutions have been proposed, there is much room for improvement in those solutions.

2.2 WIRELESS RADIO SPECTRUM

The radio frequency spectrum is the medium on which all wireless communication is carried. It is a natural resource and it is finite in terms of usable regions. Usable regions of spectrum for communication range from 3.0 kHz to 300 GHz. This is because of the limitations of technology such as transmitters, modulation techniques and antennas. Physical limitations on antennas will prevent higher frequency transmissions from being realised [10]. In the usable region, the spectrum is divided into blocks. The blocks are of varying size and each serves a particular purpose. Within each block, the spectrum is further divided into channels. Channels will have a different width for each block depending on the use of that block. To prevent interference, there can only be one user per channel. This is the fixed spectrum access (FSA) policy introduced to manage spectrum use.

2.2.1 Spectrum management

The FSA policy was developed by the federal communication commission (FCC) to reduce interference and for security purposes [11]. Under the FSA only certain transmissions are allowed in specific bands. For instance analogue TV communication may only be allowed in the 400 MHz to 700 MHz block of spectrum. This will help ensure that all transmissions are interference-free because the transmission power, modulation techniques and channel allocation can be regulated.

The FSA policy is implemented and enforced by a government agency (the regulator). Therefore spectrum block allocation may vary from country to country. However, as a guide, the ITU has set international standards to help with the allocation. One such standard is the global communication system for mobile communication (GSM). GSM was developed as the second generation of mobile communication after the analogue telecommunication standards [12]. GSM is allocated to the spectrum bands of 800, 900, 1600, 2300 and 2600 MHz. Therefore, to ease international communication all government agencies will set aside these bands for GSM. Cellphone manufacturers are then mandated by law to ensure that their products connect to all of these bands, thus ensuring the device will work anywhere there is GSM coverage. Recently, as a result of growing technologies and reshaping spectrum use, these bands have been reassigned for other uses such as long-term evolution (LTE) communication [13].

In South Africa the local authority, the independent communications authority of South Africa (ICASA), is the regulator in terms of spectrum allocation and management. ICASA has for instance procedures to award spectrum on a competitive basis where spectrum is insufficient to meet demand. The regulator will publish an invitation to apply (ITA), after which users will apply for use of the frequency advertised in the ITA. ICASA also allows spectrum sharing. The regulator allows two or more licensees to be granted radio frequency spectrum access. Therefore spectrum sharing is allowed and is regulated by the responsible authority [14]. Furthermore, ICASA is committed to digital migration of TV signals. This will free up spectrum that can be redirected to mobile communications.

Other factors to be considered during spectrum band allocation are physical limitations. For example, TV signals are generally required to cover a large area on a single transmission antenna. This keeps the costs down and infrastructure to a minimum. To achieve this the transmission frequency must be kept as low as possible. This is because the lower the frequency, the further the signal travels. The

same principle applies to mobile communication. To keep infrastructure minimal, the operators prefer to use the lower frequency bands.

Since the lower frequency bands are now in high demand, the regulators have taken to auctioning off and licensing these lower frequency bands. This spectrum is in such high demand that an auction held in the United States in 2017 raised 19.6 billion dollars. However, only so many bands are available and therefore a limited number of channels can be used. This leads to high demand and competition for the channels among users.

2.2.2 Increase in users and 5G

Over recent years rapid development of the mobile and wireless communication sector has resulted in a drastic increase in the number of users. The ITU reports that the number of mobile users increased by 12.5% percent [1]. At this level of growth, the networks are hard pressed to accommodate every user. Not only the number of users is growing, but user data demands are increasing rapidly as well [15]. This has led to a need for faster and more efficient networks.

Emerging technologies such as 5G aim to alleviate some of these challenges [16]. 5G promises faster uninterrupted connectivity, which will require higher efficiency in terms of spectrum management. 5G also promises low latency values of 1 ms and minimum data rates of 30 and 50 Mbps respectively for 50% of the population by 2020 [17]. The technology will still mostly use the existing frequency bands hence the need to increase spectrum utilisation. Some 5G applications, such as Smart Grid, have been proposed on the 60 GHz band. The results indicate that the band is as reliable as fibre optic communication for overhead high-voltage grids [18]. However, the technology is still faced by some challenges such as frequent hand-off in small-cell networks [19]. Interference is also a major concern in dense small cell 5G networks in which a large number of small cells are densely distributed in a telecommunications network. This allows the spectrum to be reused because the small cells have low transmission power and limited coverage. However, the large cell is still providing cover over the entire region and there will be some interference between the small cell communications and the large cell communications. A Stackelberg game model is proposed to determine the channel allocations with the intent of alleviating the interference. Simulation results have shown some benefit of the proposed scheme [20]. The speed and the latency specifications for 5G make it ideal for mobile healthcare

applications. Numerical results indicate that the architecture based on 5G outperforms the existing 4G architecture. 5G technology is on the way and has already shown great potential [21]. Device-to-device (D2D) communication is widely regarded as one of the cornerstones of 5G communication. D2D devices in close proximity to each other can communicate via a direct path using licensed LTE spectrum. This will provide some increase in spectrum utilisation but a significant problem is encountered if there many devices and applications in close proximity. Results show that D2D can achieve high data rates and reduced discovery times [22].

A new term that is emerging is called the internet of things (IoT). It describes a world of smart devices all linked together to create an interconnected environment. Because of the high level of interconnectivity, interference will have to be managed effectively. Power consumption and communication priorities will also need to be properly set up. Through that, a path has to be cleared to allow 5G networks to thrive.

2.2.3 Congestion and spectrum utilisation

The FCC recently conducted research to determine the state of spectrum usage [23]. They discovered that although there was little to no unlicensed spectrum, spectrum was being under-utilised. They found that at any given time there were channels not being used by the licensed operators and yet new operators were struggling to find free spectrum to use. Some operators only licensed a small amount of spectrum that was insufficient for their user base. In this case the network would be congested and there would dropped and blocked communications. The overall assessment was that the spectrum was congested in some parts and greatly under-utilised in some parts. Some reasons for the under-utilisation were that regulators gave geographical licenses. Operators would license spectrum even in regions were they had a very low subscriber base.

Locally, ICASA has conducted its own surveys and has determined that mobile telecommunication is the primary mode of communication in the country, with 85.5% of all household relying on only cellphone communication. Data collected by the regulator show that mobile service revenue increased by 4.3% [24] whereas fixed internet and data revenue decreased by 7.5% from 2015 to 2016. This shows that the market demands of wireless access is thus putting more pressure on spectrum utilisation. Data collected by Statistics South Africa show that almost all the country's data traffic comes from

urban areas [25]. This means that in rural areas where there is coverage in some parts, spectrum is under-utilised and SUs may be allowed an opportunity to give the rural population affordable access.

Spectrum improvement techniques have been applied in the past. Initially communication was only half duplex. That is, it only went one way. During communication a user would either be transmitting or receiving, but never both. This was later improved to full duplex, whereby a user could receive and transmit simultaneously. Frequency division duplexing (FDD) was another a technique that was introduced. In FDD, two channels form a single duplex radio channel. The two channels will have fixed frequency separation known by the system [26]. Another technique, time division duplexing (TDD) has also been adopted. In TDD, information is transmitted in time slots to the base station from the mobile and vice versa. Given enough transmission rate over the user data rate, it is possible to store a portion of the transmission and give the illusion of full duplex communication. Other non-conventional techniques have been proposed, such as using drones to mitigate congestion in 5G networks by employing them as relays to allow for rapid offload of data traffic [27]. The point is that user demands change and technology must always be improved to meet them.

For certain bands, regions exist where there are large unused gaps between the frequency channels. This is especially common in the ultra-high frequency (UHF) and very high frequency (VHF) bands. These gaps are known as TV white spaces [28]. Furthermore, white space may become more freely available after migration from analogue to digital broadcasting. The ITU set a deadline of June 2016 for digital migration. Use of TV white space is only allowed under strict interference adherence policies. White space devices (WSDs) are required to use sensing or geo-location databases to determine the acceptable transmission power levels to prevent interference. Sensing is still a major challenge in terms of accuracy and it adds cost and complexity to the system [29]. Geo-location databases are maintained by the regulator and keep track of white spaces in certain areas. Databases are currently the approach to determining white spaces recommended by the FCC, the Office of Communications (Ofcom) and the Electronic Communication Code (ECC) [28]. Ofcom has conducted a white space pilot study to determine the viability of using white spaces. Trials done under the pilot study had to test specific aspects set by Ofcom [30]. The pilot study aimed to test:

- WSD operation and conformance.
- Interference management.

- Geo-location database operation and calculations.
- Coexistence.

The key findings of the trials were that:

- TV white spaces have most potential in below-rooftop and underground deployments.
- There is a lot of white space available in the UK, particularly in London although it is a highly metropolitan area.
- Reassignment of the 700 MHz spectrum can affect the capacity in some white space usage areas.
- High capacities can be obtained by aggregation of white space channels.

A study done in China proposed and showed that white spaces can be used for LTE communications [31]. The study considered the 470 - 806 MHz band and the researchers found that 67% of channels are used over cities.

TV white space trials have also been conducted in Africa where the population is dispersed and most people still live in rural areas [32]. TV white spaces on this continent present an opportunity with great potential. The white spaces can be used to provide coverage to large parts at low cost and with minimum to zero interference with the incumbent broadcaster. Trials conducted in Africa set simple and clear goals to increase internet penetration in the region. Some of these objectives were:

- To show that white spaces can deliver broadband at low cost;
- To discover and observe frequencies that have the greatest potential for white space implementation; and
- To measure levels of interference between WSDs and the licensed broadcasters.

The benefits of white space broadband from an African perspective are immense. Since white space implementation requires low regulation, the cost of running such system is minimal. Operators can quickly provide service in remote parts without much difficulty. Trials conducted in Cape Town have shown promising results of 12 Mbps on downlink. This was achieved with zero co-channel interference [33].

While white space is not the focus of this research, the results of trials and research done in that field provide much valuable insight into the behaviour of transmission power and its effect on interference with neighbouring users. The trials also focused on interference effects and on limiting harmful transmission. The results can be applied in any network and may be implemented in CRNs and hence provide enough evidence for this research to base assumptions on transmission power.

2.2.4 Dynamic spectrum access

Dynamic spectrum access (DSA) is an alternate to the earlier FSA policy. This can be achieved through software defined radio (SDR). SDRs can be programmed to read and adjust to different network conditions. They were first developed in 1984 [34].

2.2.4.1 Dynamic exclusive use model

Two approaches to DSA have been proposed. The first is spectrum property rights. Adopting this approach, licensees are free to trade or sell spectrum and also have the right to choose the technology they use on the spectrum. This approach encourages a market-driven implementation in which only the most profitable route will be chosen. The regulator will no longer mandate how the license is used.

The second approach is called dynamic spectrum allocation. This approach aims to improve spectrum utilisation by making use of the traffic statistics of different services, i.e. spectrum is allocated to specific services in a given region and a particular time.

2.2.4.2 Open sharing model

The open sharing model is also sometimes known as spectrum commons. The success of WiFi and Bluetooth technology on the industrial, scientific and medical (ISM) band has proven that an unlicensed, unmanaged approach to spectrum allocation can work. However, the model will be limited to peers in a network and ideally applied per region.

2.2.4.3 Hierarchical access model

This model is the foundation of the work of this research. It entails two the classes of users, PUs and SUs thus forming a hierarchy. Two approaches to this are overlay and underlay spectrum.

The overlay approach allows the SU to transmit at full power and only when the PU is absent. Therefore the SU is allowed to identify and use any available (free) spectrum. The SU is required to vacate the spectrum should the PU require to use that spectrum. When this happens the SU can find other available spectrum or stop transmitting.

The underlay approach on the other hand is not limited to free spectrum only; the SU can transmit together with the PU. The consequence of this however, is that there will have to be a severe limit on the transmission power of SU. This is required in order to avoid interference on the PU transmission.

A term that has been coined to describe such underlay and overlay networks which employ DSA, is cognitive radio technology. Cognitive radio is an SDR that has the ability to learn from and adapt to its environment by autonomously reconfiguring its network parameters. Thus the hierarchical access model, specifically cognitive radio technology networks are implemented in this research. The chosen access model allows for two classes of users with a power constraint imposed.

2.3 COGNITIVE RADIO NETWORKS

Cognitive radio techniques allow the use and sharing of spectrum in an opportunistic manner [35]. Fig. 2.1 shows a basic model of a CRN.

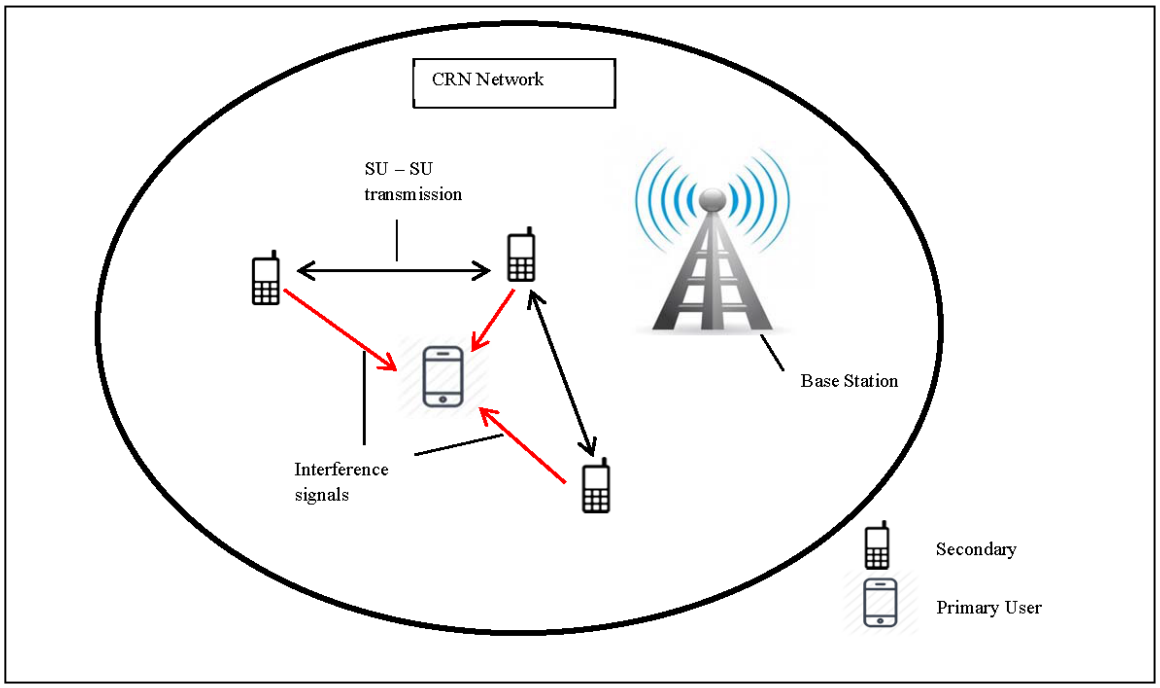

Figure 2.1. Typical CRN Network

In Fig. 2.1 a typical network cell with a base station is shown. There are three SUs and one PUs. SUs are allowed to transmit their data as long as there is no harmful interference to the PU. The PU will connect outside the cell via the base station. The SUs may also use the base station to communicate with one another. If the communication is on the same channel there may be interference with the PU communication, as shown by the arrows. The SUs may also communicate via peer-to-peer transmission. Either way interference with the PU must still be mitigated.

2.3.1 Main functions of a CRN

The following are the main functions of a CRN

1. Spectrum sensing: Refers to the ability of the network to detect unused spectrum without interfering with the PU or other SUs. Sensing also detects the arrival of the licensed user in the spectrum [36].

2. Spectrum management/decision: This refers to the identification of the best possible spectrum to meet user requirements. Requirements may be data rate, delay sensitivity and packet error rate [37].
3. Spectrum mobility: Refers to the ability of the network to maintain user communication requirements seamlessly during transition to better spectrum.
4. Spectrum sharing: This refers to the fair distribution of spectrum among SUs in the network and may also entail sharing a particular channel between two or more SUs based on their user requirement. This function is necessary to minimise interference among users [36].

2.3.2 Types of CRN

As mentioned before a CRN can be classified into types namely underlay and overlay. However, some texts describe a third type known as hybrid.

1. Underlay: In this type of system the PU and the SU can occupy one channel simultaneously. This is achieved by ensuring that the SUs transmit at power below a threshold power such that they do not interfere with the transmission of the PUs [38]. The SU would need instantaneous knowledge of the interference threshold for every channel and every PU.
2. Hybrid: Here, the SUs have knowledge of the data sequence from the PU and this is used to assist the PU's transmission. The SUs use part of their power to transmit the PU data, thus reducing interference [39]. SUs also occupy a channel with a PU.
3. Overlay: In this type of system the spectrum users only transmit when the PU is absent. The SUs detect when there is free spectrum and use the free channel to transmit. Power constraints are still in effect so as not to interfere with neighbouring users but are much less restrictive than in the other two types. The secondary and PUs do not occupy a channel simultaneously [40].

2.3.3 CRN spectrum allocation approaches

A key part of CRNs is spectrum allocation. Spectrum must be assigned to users according to some criteria that are defined in the network. Ideally these criteria will be known to the SUs such that rules can be developed by which the SUs must abide. The main approaches are centralised, distributed and cluster-based.

1. Centralised: This entails, as the name implies, a central node that takes decisions on channels assignments. The central node collects radio information and requirements from all SUs either periodically or when requested. A separate entity called a spectrum server may serve as the central node [35]. The centralised scheme has many advantages, such as ease of throughput maximisation and reduction of interference between SUs. This is due to the global view that the central node has of the network, which also helps to maintain connectivity. Fairness can also be easily achieved in terms of spectrum allocation or by reducing the number of greedy users. This scheme can furthermore integrate topology control using conflict graphs. Centralised systems are also advantageous in that the central node can make use of priorities to SUs with constrained interfaces to maximise their throughput [3]. A disadvantage of the scheme, however, is that the need for SUs to exchange information with the central node induces signalling overhead in the CRN. A serious disadvantage is that if the central node fails, spectrum allocation will not be achievable and SUs will begin selecting channels independently, resulting in unfairness [3].
2. Distributed: In a distributed scheme, SUs take decisions on their own and by cooperating with neighbouring users. There is no central node to oversee the network, therefore each SU calculates a metric and sends the information to nearby SUs. The SU will then determine the traffic load on each channel and ultimately select the channel with the minimum interference or load [37]. The benefits of a distributed scheme are lower signalling overhead, faster decision time, an incentive for SUs to participate in information sharing and better handling of outages over a centralised scheme. The distributed scheme has some drawbacks, for example that global fairness is difficult to optimise, decisions are not optimal as information is from neighbouring SUs only and inaccurate and false information can destabilise the network as well as allowing malicious users to exploit free spectrum [3].
3. Cluster-based: This is more of a hybrid between the centralised and distributed schemes. The network has mesh routers that are static and clients that are mobile. The network is divided into clusters, each with a cluster head. The cluster collects and combines sensing information from all the SUs in the cluster and generates a final spectrum allocation vector [41]. A cluster head will then make a decision on which spectrum band to use, based on allocation vectors from all the cluster heads. In the event of a cluster head failure, the SUs can subscribe to the closet cluster head. The benefits of this scheme are robustness against cluster head failures, efficient bandwidth utilisation, overall decrease in communication overhead and the possibility of bandwidth reuse. The drawback of a cluster-based scheme is that the mesh can become congested quickly, but this can be mitigated by structuring clusters to adapt the load dynamically [3].

There are other less common methods of spectrum allocation, which are briefly described:

1. Multi-channel selection: Here the SUs split their data into multiple pieces and transmit on multiple channels. The advantage is that their data rate and spectral efficiency are increased. The disadvantage, however, is that there is high switching overhead (it takes longer for the network to reconfigure) and there is more risk of interference.
2. Common control channel: An overall channel is dedicated to handle all spectrum assignments. The advantage is that information exchange is guaranteed and the drawback is that the common channel is susceptible to jamming.
3. Segment-based: Segments are employed such that each SU has at least one common channel in a segment. This approach results in much lower overhead switching but the trade-off is that it can be easily congested.

2.3.4 Spectrum allocation techniques

A three-step process is generally followed to spectrum allocation. The first step is to determine the criteria of the user requirement. The second, is to define an approach that best models the target criteria and the final step is to identify the technique that will solve the spectrum allocation problem [3].

2.3.4.1 Criteria

Within a CRN itself there is competition for free spectrum among the SUs to transmit when allowed. SUs have different requirements to maintain quality of service (QoS). An example is applications such as video chatting. Therefore, spectrum allocation to the SUs must be based on these needs. The requirements are as follows:

1. Interference power: The constraint of CRNs is that the SUs must not transmit at any power high enough to cause harmful interference to PUs. Moreover, for optimum performance, the interference power between the SUs themselves must be kept to a minimum. Most of the work thus far has been based on interference temperature limit (ITL) at the PU and spectrum is assigned to SUs with the aim to keep the ITL under a predetermined threshold [42, 43]. Several methods for spectrum allocation are given in [4]. One such method is to reduce the

transmission power given different data rates. This is achieved by allocating an SU with a large data rate to a channel with low interference and large bandwidth. However, the method makes the assumption that the SUs exchange information at the required data rate. A method of determining interference based on the path loss model has also been analysed [44]. The work done on interference power minimisation has relied on assumptions about the model. Some common assumptions are given below [3]:

- SUs can accurately sense the ITL at the PU, since SUs may only transmit at power that is below a predefined threshold.
- SUs access the same set of channels.
- The SUs are cooperative and share information such as data rate and transmission power.
- Channel gain information is available instantaneously.
- PU information such as location and bandwidth is known to the SUs.

2. SU data rate: The transmission rate of the SUs should be as high as possible. However, many constraints have to be considered before maximising the data rate, such as:

- The SU's transmission power, which is limited because of possible interference.
- The transmission channel capacity (Shannon theorem).
- Harmful impact on the PUs.
- Effect on neighbouring SUs.

The data rates of all SUs in CRN have been considered as a whole and the sum of the data rates set as the criterion to be considered [44]. This approach, however, can lead to starvation where some SUs data rates are neglected. A different approach was considered where an assumption is made that the SUs use uplink sub-carriers from a primary network. The objective is given as [45]:

$$\sum_{k=0}^{|k|} g_{i,k} \tag{2.1}$$

where k is the sub-carrier, $g_(i,k)$ is the allocated bits of SU i on sub-carrier k. General assumptions made in many works are [3]:

- The SUs are static such that the network topology does not change.
- Noise only comes from co-channel interference.

- Bandwidth is available to support more than one SU transmission.
- Channel conditions are stable.

3. Delay: Some SUs in the CRN may be delay-sensitive and hence spectrum allocation to these users must account for this requirement. Delay can be classified as either end-to-end delay or switching delay. End-to-end delay refers to the total time for the transmission of a packet from its source to its destination whereas switching delay is the amount of time taken by an SU to move from one spectrum frequency to another. During switching, transmission is briefly halted, thus causing extra delay. The switching delay may be as high as 10 ms for a 10 MHz change in frequency up to 3 GHz [46]. The total delay can also be calculated as a sum of the delay of the existing flow and the new flow [47]. Common assumptions about the delay requirement are [3]:

   - Spectrum assignment is usually combined with routing.
   - Switching delay is constant.
   - Channel widths are constant.

4. Energy efficiency: As in any system efficient utilisation of energy is an important requirement in CRN. A distributed approach to spectrum access has been considered [48]. The system operates in time slots where in each slot, the SUs that wish to access the system sense the entire spectrum and locate free channels. The work takes advantage of the SUs' ability to select multiple sub-carriers to distribute spectrum selection and power allocation to minimise energy consumption. The objective is then to find the optimal number of channels that guarantees data rate requirements while transmitting with minimum power. Common assumptions considered for energy efficiency are:

   - SUs cooperate fully in exchanging information on transmission power.
   - Transmission power is the main attribute to be considered for energy management.

5. Network connectivity: Connectivity is affected by the distance to nodes and transmission power. The impact of spectrum assignment on connectivity has been studied where the CRN was modelled using graph colouring [49]. Interference between SUs was shown to have a high impact on network connectivity. General assumptions about connectivity are:

   - The network has a fixed communication graph.
   - Only co-channel interference is present.

- There is a centralised approach to the CRN.
- The channel is stable.

6. Fairness: Fair distribution of spectrum among SUs is a very important requirement. All SUs in the CRN must be treated equally and assigned to spectrum fairly. Several methods have been proposed, such as considering average throughput per SU and using a fairness factor [50]. Fairness has mostly been analysed using a centralised approach. While this solves the starvation problem, the approach does not consider minimum throughput of SUs [3]. The fairness problem has also been analysed together with spectral efficiency [50]. The objective was to optimise the spectrum used by each SU. Game theory was used to allocate the spectrum fairly, observing sensor priorities. The results indicate that fairness was indeed achieved. Another method to achieve fairness is to use traditional max-min fairness. Here, the objective is to maximise the minimum share of resource among the SUs [51]. Max-min fairness was used to maximise the minimum average flow bandwidth for all users while minimising spectrum waste by idle SUs [52]. Proportional fairness was used, with several objectives being proposed, such as assigning each SU a data rate that is inversely proportional to predicted spectrum usage and maximizing the logarithmic utility functions [53]. Fairness measurements such as the Jain fairness index can be used to analyse the spectrum allocation. The index is a quantitative measure of spectrum sharing or allocation and is independent of the amount of spectrum available. The index lies between 0 and 1, with higher values indicating a more fair allocation [54]. A number of studies have proceeded to use the index to evaluate their fairness techniques [55, 56]. Some general assumptions made to solve the fairness problem are:

   - All channels have the same capacity.
   - SUs are single radio devices.

7. Profit/Cost: The trade-off between the cost of transmitting over a channel and the reward obtained by the SU is also a requirement to consider in spectrum allocation. Profit functions have been developed in an attempt to determine the reward for SUs [57]. Spectrum is split into two types: 1) spectrum is shared between SUs and PUs and 2) spectrum is shared between SUs only. Using the profit functions and determining which type of spectrum they want users can decide how much they need to pay for spectrum. Common assumptions made on cost of spectrum are:

- SUs pay the price of the spectrum they want and the price is dependent on the amount of spectrum.
- There are different classes among SUs and they should not all pay the same price.
- Spectrum use should earn a reward.

8. Risk: In an overlay CRN, risk is calculated as the probability of an SU's transmission being blocked by the emergence of a PU on the spectrum. A risk calculation method is proposed and used to access the optimal channel assignment that minimises the blocking probability in the network [58]. Common assumptions made on network risk are:

    - Occupancy of different channels is mutually independent.
    - The SUs can sense many channels but can transmit on only one channel.
    - An SU has sole use of a channel at a specific time.
    - The SUs have knowledge of the channel availability probabilities.

2.3.4.2 Techniques

Various methods have been proposed to solve the spectrum assignment problem. The techniques must take into consideration the requirements of the user as well as spectral efficiency.

1. Heuristics: The solution to a spectrum allocation problem bound by certain criteria is found through iterative algorithms. The provides the challenge that the solution is mostly particular to that specific problem. Should the criteria change or be updated then a new heuristic will need to be developed. However, the technique is easy to implement [59].
2. Game theory: The spectrum assignment problem is modelled as a game with some criteria set as the rules. The Nash equilibrium is then used to solve the game. It is defined as a stable system in which no participant stands to benefit from a universal shift in strategy as long as the strategies of other participants remain the same. The advantage of game theory is that proper decision-making can be applied to cooperative and non-cooperative CRNs. A drawback however, is that equilibrium is not always achieved [60, 61].
3. Graph theory: Conflict graphs are used to model aspects such interference allowing the network to be analysed visually. However some parameters cannot be modelled, hence the technique has

a few shortcomings. The benefit is that solutions already exist and they can be used in spectrum allocation [62, 63].

4. Linear programming: Mixed integer linear programming is used to formulate the joint power and data rate problem. This problem is common in spectrum allocation because of the challenge in maintaining high data rates while maintaining low power. The advantage is that there are many existing linear programming techniques that can be used. However, the drawback is that several assumptions have to be made in the modelling [64].
5. Fuzzy logic: A set of rules and weighting functions are used to achieve the optimal point with speed and quality being the greatest benefit. The trade-off, however, is inflexibility because the rules are predefined and cannot simply solve a different problem [65].
6. Evolutionary algorithms: A stochastic approach is taken and objective functions are used to re-analyse the problem at every iteration. The diversity of the approach allows it to handle various constraints and objectives but the technique is slow and does not always achieve the optimal solution [66, 67].

A centralised approach is eventually chosen for the proposed model. The interference power limit is the criteria chosen and heuristics are used to obtain the solutions to performance parameters.

2.4 QUEUEING THEORY

Queueing theory is one of the more recent approaches to be used in an attempt to solve the CRN spectrum allocation problem. Queueing theory has been used to solve and analyse many of real world problems such as telephone exchange systems, air traffic control, hospital queues, etc [68]. Queue analysis provides useful QoS information such as waiting time and blocking probability. Knowledge of such parameters in a CRN can provide better service to all users, involved since the network can be optimised according to the analysis. SUs, for example, can use knowledge of the waiting time to decide whether to join the service queue or not.

2.4.1 Types of queues

Queues can be divided into several types each differing in application and available resources [7].

- Single-node queue: Packets are served at one location and then leave the system. The packets do not proceed to another server for further processing. Packets can however re-join the system at the first location and be served again as a new packet. A single-node queue system can have multiple servers operating at its single location. This is called a single-node multiple parallel servers queue. Another variation is to have multiple queues at the location but with only one server. Here, polling is used to serve the multiple queues. This is a called a polling system. Another example is a feedback queue whereby after completion a packet immediately re-joins the queue to be served.
- Tandem queues: A tandem queue combines several single-node queues in series so that after a packet exits one queue it immediately joins another queue at a second location. However, the route is sequential, so a packet cannot skip a location before exiting the entire system.
- Network of queues: In this version, various queues are involved and they may be in no particular order. A packet can enter the system at any location, proceed to any location and exit the system at any time.

2.4.2 Markov process

Queueing theory has always gone hand in hand with Markovian processes. In fact, nearly all queueing theory problems can be set up as a Markov chain [7]. Thus, analysis of Markov chains is essential in understanding queueing theory. A Markov process is a memoryless process that is independent of past states in that the future state is only dependent on the present state [69]. Markov chains can be split into two classes namely 1) discrete time Markov chain and 2) continuous time Markov chain. Discrete time Markov chains have been studied more extensively than continuous time Markov chains.

2.4.3 Characterisation of queues

Queues can be characterised according to their attributes such as arrival pattern, service pattern, number of servers, etc. A queue can be described using a set of letters that state the attributes of the queue. For example queue notation can be written as M/M/1/K/FCFS. Here the first position represents the arrival pattern, the second position represents the service pattern, the third position represents the number of servers, the fourth position represents the buffer size and the fifth position represents the

queue discipline. Table A.1 in the appendix summarises the meaning of the various symbols often used.

In addition, there are various queue performance measures that can be evaluated [7]:

- Queue length: The number of packets that are waiting to be served.
- Blocking/Loss probability: The probability that when a packet arrives it cannot join the queue because there is no space available in the buffer.
- Waiting times: The delay that a packet can expect before it can be served.
- System time: The total time spent from arrival to departure from the system.
- Work load: This is the sum of the service time of the currently being served and the service times of all the packets in the queue.
- Age process: This is a measure of the amount of time that the current packet being served has spent in the system.
- Busy period: This is the time from when a server begins to serve packets after an empty period to when the system is empty again.
- Idle Period: This is the opposite of a busy period. This is a measure of the time from when the system is empty to when a packet arrives.
- Departure times: This is the amount of time it takes for a packet to complete service in a single-node system.

2.4.4 Queues in CRN

The most widely used queue system in CRN is the M/G/1 queue [9, 70]. The PUs are commonly associated with On/Off behaviour. Queue performance measures for the M/G/1 model were analytically evaluated with the service times being equal to the channel capacity [71]. It was found that with decreasing PU outage probability the mean service time of the SUs would increase. The increase in mean service time also had a detrimental effect on the waiting time and the mean number of packets in the system.

Another M/G/1 system was modelled and analysed for an underlay CRN [9]. A finite buffer length was added to simulation to model a more realistic situation. The system was put under a maximum

power constraint defined by the interference on the PU. The service time was also defined to be equal to the channel capacity according to Shannon's theorem. The channel chosen was a Nakagami-m fading channel. The resulting model had an embedded Markov chain and following this, performance measures such as packet transmission time, blocking probability, throughput, channel utilisation, mean number of packets and system time could be deduced analytically. The results indicated that blocking probability will decrease if the buffer size is increased and mean waiting time will increase with an increase in buffer size. The mean number of packets in the system was found to increase with an increase in the arrival rate and lastly the throughput decreased with an increase in fading severity. All the numerical results were verified with simulations.

A multi-service queue with server failure being modelled as a channel being occupied by a PU is another possible scenario [72]. A centralised CRN was considered with the central node fully aware of the SU and PU information. The service times were variable and there was no priority amongst the SUs themselves. Analytical expressions for the remaining SUs and mean number of SUs in the system were derived. The results showed that the system would be stable as long as the arrival rate times the service demand is less than the capacity of the CRN. The number of SUs in the queue would increase, however, if the SU arrival rate increased. An M/G/1 queue was used to model the PU traffic.

Multiple SUs, on PU vacation, can compete for the same channel, thus causing collision and unfairness. An approach that allows the SUs to share the channel has been proposed [5]. When the SUs access the same channel they will time-share the channel. The SUs are assumed to be transmitter-receiver pairs with co-operation. The network is also considered to be static. In order to achieve fairness and spectral efficiency the SUs were assigned to priority classes with a higher class able to pre-empt a lower class. PUs were assigned the highest priority and SUs were divided according to their requirements. The objective of the SUs was to improve QoS parameters such as reducing end-to-end delay and blocking probability. Simulation results indicate that the proposed approach performs much better than conventional methods in terms of video quality.

A less known queue model is the Mt/Mt/1 model, in which the arrival and service rate are regarded as time-varying. The mean rates will change with time. A closed form solution for such type of queues was obtained by using Volterra type integrals [73].

2.4.4.1 M/M/1 priority queues

Priority based queuing: This is based on an M/M/1 system whereby when the PU arrives in system two schemes can be used to handle the SU [74].

1. Pre-emptive resume priority: The PU has the highest priority and when the PU arrives in the system the SU's transmission will immediately be interrupted to accommodate the PU [6]. The interrupted SU can then continue transmitting the PU has left the system. Expressions for the queue length and system were derived with the help of Little's Law [68, 72]. The analysis was also expanded to a multichannel priority system. The results indicate that the system time of the SUs will always be greater than that of the PUs, regardless of whether the SU has less processing time or not. The performance of the system is also dependent on the number of users and channels that are available. It is important to note that other priority systems exist in which when interrupted, the SU will immediately exit the system and will have to re-join the queue. Such a system is known only as pre-emptive priority [7].

   A two-part call level queue that implements a direct Markov model has been proposed, in which the SU queue is separated into two parts [8]. The first part consists of SUs that will, on interruption save packets in the buffer until the PU leaves and in the second, the SUs will discard the current packets on interruption. The results have shown that, for the delay queue, the throughput and packet loss rate increase as the delay increases and the queue increases. The delay queue length does not make a difference in terms of spectrum utilisation.
2. Non-pre-emptive priority: In this system the SU is allowed to finish its current transmission when the PU arrives. That is, the PU will have to wait for the server/channel to be idle again before being served [74]. Expressions for mean waiting time and mean throughput time for the PU have been developed. The results show that the SU sojourn time is much greater than PU sojourn time.

2.4.4.2 Processor sharing

The available service time is divided equally among all the packets/customers present in the system. Thus, by definition, there are no waiting customers and everyone present in the system will be served.

Customers that arrive when the system is full will be blocked. The system is defined as full when the service rate can no longer be divided further to sustain reasonable service.

Processor sharing has seen wide use in computer systems where the service rate of the processor is usually fixed. Large and small jobs arrive for service and both begin service immediately. The benefit is that small jobs that arrive when a large job is being served do not have to wait a long time to be served. Hence multiple small jobs can arrive and be completed with little effect on the large job [75].

Weighted processor sharing has also been developed. In this approach, jobs are given weights according to some criteria, for example, importance or the size of the job. The jobs that are weighted higher will receive a bigger chunk of the service rate. There is still no queue and jobs that arrive when the system is full are blocked [76].

Fig. 2.2 shows a typical processor-sharing queue. Here the queues will share the service equally while serving the packets present in their queues. When one queue becomes empty, the remaining queues will share the service rate among only themselves equally again. The total service rate, μ, is shared equally among the n SUs present.

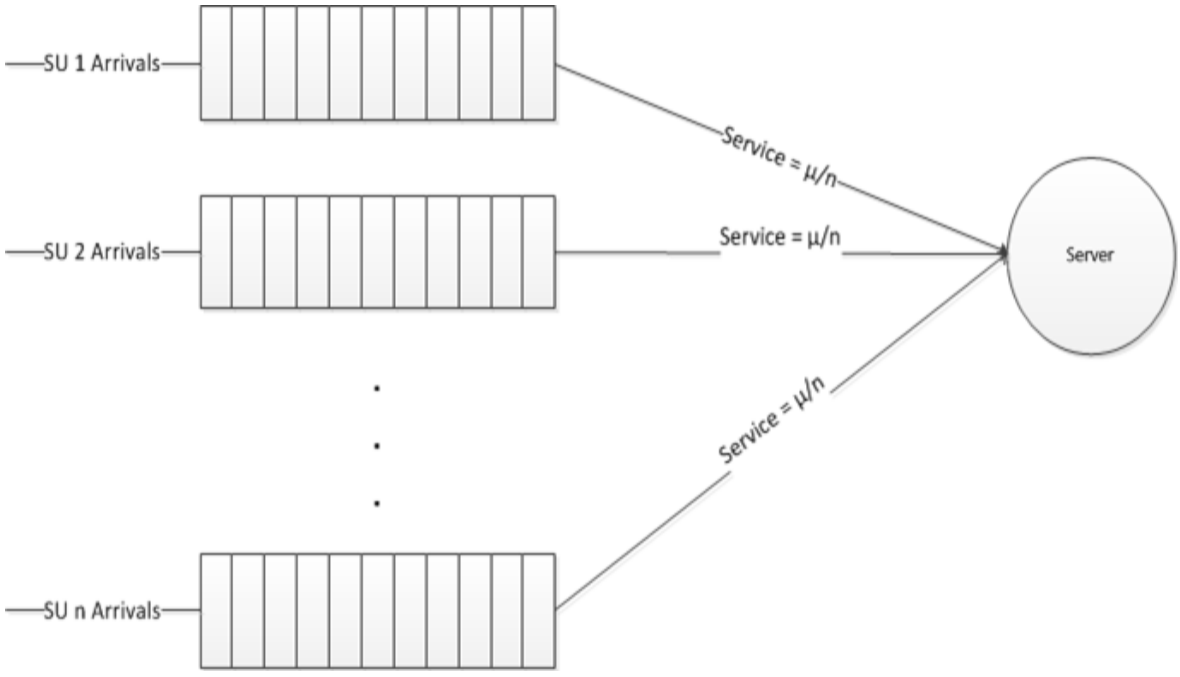

Figure 2.2. Illustration of processor sharing queues.

2.4.4.3 Head of line processor sharing

The processor sharing approach is further developed to allow queues. Multiple queues can be served simultaneously. Each queue will have a similar type of customer that on arrival will join only the relevant queue. The processor will then divide the service rate among the queues equally, but serve only the job at the front of a queue. The benefit is that because of the queues, fewer jobs are blocked [77].

2.4.5 Queue characteristics

Queues are defined by arrival rates and service rates. Since the queues are observed over a long period, arrival and service times are characterised by mean times. The mean arrival rate is denoted by λ and the service rate by μ. The arrival rate indicates the number of units/packets arriving per unit time. Therefore $\frac{1}{\lambda}$ is the amount of time between arrivals, known as the mean inter-arrival time. Similarly, the mean service rate indicates the number of units/packets served per unit time and $\frac{1}{\mu}$ is the amount of time between each service completion and is known as the mean inter-service time. Generally for a stable system, the service rate must be greater than the arrival rate otherwise the queue size will continue to grow. This ratio is known as the traffic intensity and given by the formula:

$$\rho = \frac{\lambda}{\mu}. \tag{2.2}$$

From equation 2.2 it can be seen that for queue stability ρ must be less than 1. The traffic intensity is a measure of how busy the system is. The mean arrival and service time can be used to obtain various queue performance measures, such as system length and waiting time.

Considering an M/M/1 queue system the mean number of packets in the queue (that is, excluding the one in service) is given by [68]:

$$L_q = \frac{\rho^2}{1-\rho}. \tag{2.3}$$

Similarly, the mean waiting time of a packet in a queue (that is, before entering service) is given by [68]:

$$W_q = \frac{\rho}{\mu-\lambda}. \tag{2.4}$$

These equations can also be extended to give the mean length and mean waiting time for the system (that is, including the packet in service and the service time of a particular packet respectively). The mean length for the system is:

$$L = \frac{\rho}{1-\rho} \tag{2.5}$$

and the mean waiting times is:

$$W = \frac{1}{\mu - \lambda}. \tag{2.6}$$

These equations are only valid when the queue is stable and has reached steady state conditions.

2.4.5.1 Little's formula

Little's Formula or Little's Law relates the waiting time and the queue length [78]. Since the waiting time is affected by the position of the packet in the queue, a relationship between the length and waiting time exists. In fact, the queue waiting time is directly proportional to the queue length. It states that:

$$L_q = \lambda W_q \tag{2.7}$$

and similarly:

$$L = \lambda W. \tag{2.8}$$

Finally, an adapted M/M/1 processor sharing queue is adopted to be used in the proposed model. Chapter 3 highlights the specifics of the adoption.

2.5 OPTIMISATION

Optimisation has been used to improve many systems and processes. The goal of optimisation is to maximise or minimise some objective function given constraints in the system.

A general optimisation problem is defined by its objective function. This is the aim of the optimisation, the quantity that must be optimised. Generally, an objective function has the following form:

$$\text{maximise } z = f(x,y,z). \tag{2.9}$$

Equation 2.9 describes a function of x, y and z and must be maximised. Simply, the aim is to find the highest value that the function can achieve.

However, there may be certain restrictions on the quantities x, y and z. These are the constraints of the optimisation problem and can be written as:

$$\text{subject to } g_i(x,y) \leq b_i \; i = 1,...,m \tag{2.10}$$

$$x_k \leq 1 \; k = 1,2,.....,r \tag{2.11}$$

$$z_k \geq 0 \; k = 1,2,.....,r. \tag{2.12}$$

Here, the functions g_i constrain the quantities x and y to some limit, b_i. This may be due to the physical limits of x and y or to any other valid reason. Furthermore, the variables x, y and z can be individually constrained to some limits or certain values. The objective function will have to be maximised with these constraints in place.

Every system has a desirable quality or a minimum performance requirement. These requirements can be throughput, minimum power use or network time [79]. In a network where multiple SUs are competing for access and they have certain requirements pertaining to them, the network needs to provide reasonable distribution of available resources to allow each SU to meet its respective requirement. This can be solved by finding an optimal point; that is, some network parameters can be adjusted to determine the best environment for all involved. This is usually achieved by determining an objective function to suit all clients. Objective functions, however, may not be straightforward and must be carefully determined to ensure that the solution found is indeed optimal. Determining a fair optimisation point remains a challenge, especially if the requirements are different. However, objective functions can be defined to meet multiple requirements, given different constraints in the network.

2.5.1 Classification of optimisation problems

Optimisation problems can be classified into different classes [80]. These classes depend largely on the objective and the constraint function. The classification is as follows:

1. Based on constraints

   An optimisation problem can be classified as either constrained or unconstrained. If the goal is to optimize an objective function and no constraints are imposed on the inputs of that function, this is deemed to be an unconstrained problem. Similarly, if the inputs are constrained, this is a constrained optimisation problem.

2. Based on the goal of the optimisation

   Optimisation sometimes has the sole aim of maximising and minimising the objective function. These problems can be classified according to the goal, i.e. the problem is either a maximisation or a minimisation problem.

3. Based on the type of the objective and constraint functions

   The objective function will depend heavily on what kind of optimisation is required and how it should be done. It is therefore important to classify problems depending on the nature of the problem. The nature is determined by the nature of the equations involved in the objective and constraint function.

   If both the objective and constraint functions are made up of only linear terms, then the problem can be classified as a linear programming problem. If any of the functions in the objective or constraint function is non-linear, then the problem is classified as a non-linear programming problem. Furthermore, a problem in which the terms are polynomials can be classified as geometric programming. A quadratic programming problem is one in which the terms in the objective function are quadratic and the ones in the constraint function are linear. Geometric and quadratic problems are special cases of non-linear programming.

4. Based on decision variable

   Should some or all the decision variables require only integers, the problem is called a mixed integer programming or integer programming problem. Similarly a real valued programming problem requires all the decision variables to take in real values. This, combined with linear

and non-linear programming, will result in the common classifications, mixed integer linear programming (MILP) and integer linear programming (ILP).

5. Number of objective functions

   If the problem has more than one objective function the problem is deemed a multi-objective programming problem, else it is a single-objective programming problem.

2.5.2 Optimisation approaches

There are various techniques that may be employed to solve an optimisation problem. This will depend on the class of problem at hand. An example of a classification is a constrained linear single-objective maximisation problem. Hence the technique employed must consider the classification and be applied appropriately. Therefore multiple techniques may be used to deal with one problem [81].

1. Branch and bound

   The first step is to relax the constraints and solve the problem. Should a feasible solution be obtained, then that is the optimal solution. However, if no solution is found the approach adds a new constraint and splits the problem into two. The two branches' can be split further into four and so on. Each branch will be bound by a different or tighter constraint. This goes on until the optimal solution is found. It may be easy to tell when and if a particular branch is too far out of the feasible region and discontinue that branch. The approach can be computationally taxing, especially if the first branches are far from the solution.

2. Heuristics

   Heuristics have been used to solve optimisation problems. They are a logically based approach that carefully look at the problem and develop a solution to that problem. Therefore heuristics work particularly well to solve a given problem and are difficult to transfer.

3. Lagrangian duality

   This approach dualises all the major constraints and solves the optimisation problem using the classical Karush-Kuhn-Tucker conditions.

2.5.2.1 Linear programming

Linear programming methods can be applied not only to linear problems but to non-linear problems as well. By using approximation and/or relaxation, some problems can be solved as linear programming problems. Linear programming describes a problem where the objective and the constraint functions are all linear. A number of techniques and methods have been developed to solve both classes such as Dantzig's simplex method [82].

A further generalisation of linear optimisation occurs when the objective function itself is a convex function. This is known as convex optimisation. Should the objective function be convex any local minimum of the function is also a global function and hence the minimal (optimal) point. Convex optimisation has been used to find unknown signals in wireless sensor networks [83]. It is found that the detection scheme performs very much like an ideal detector. In order to determine the objective function in a multiple SU scenario the network layout and behaviour must be defined first.

With convex optimisation the objective function does not need to be closed form as long as the function itself is convex. However, convexity must be clearly established. At first glance some functions can appear to be non-convex but by approximation or relaxation convexity can be observed. Quasi- and pseudo-convexity are terms that describe different forms of convexity. Pseudo-convexity describes a function that is sufficiently convex in terms of finding its minimum point.

2.6 CONCLUSION

It is observed that a gap exists in terms of allowing multiple SUs access the same channel under a channel share scheme determined by power allocation. A hierarchical spectrum access scheme in the form of a CRN network is implemented in a centralised approach. The interference power is chosen as the criteria for SU requirements and control. An M/M/1 based processor sharing queue together with heuristics was deployed to model and analyse the system behaviour. Optimisation is done based on the goal of SU performance with linear programming, specifically convex programming, employed to find the optimal point.

CHAPTER 3 METHODS

3.1 CHAPTER OVERVIEW

This chapter describes the methodology of the research. The proposed solution and assumptions are thoroughly developed and the corresponding equations and algorithms are given. The queuing models are presented and the queue dynamics and behaviour are determined. The optimisation technique is given. Only two SUs are considered for the proposed models and the PU is assumed to be always present in the network.

3.2 RESEARCH PROCEDURE

The following approach to the research was taken:

- First, the question of how two or more SUs can share a single channel is discussed under network layout.
- Assumptions based on this initial layout are given and used to develop the specific detailed network models.
- Two network models (pre-emptive and non-pre-emptive) are then developed to realise the network layout suggested after considering the assumptions.
- Queueing theory is then used to manage and analyse the two models that have been developed.
- The results are obtained and compared to existing models (Chapter 4).
- Optimisation is then carried out in the form of convex programming to improve the results.

3.3 NETWORK LAYOUT

Given that at least two SUs already exist, an improvement may be achieved in two parts.

- First, by allowing both SU's to transmit simultaneously via allocating a fixed share of bandwidth to each SU. This will ensure that more users or applications have access to the network although the cost is a reduction in bandwidth for the incumbent SU.
- Then, improve that network scenario by allowing the SUs to share capacity. That is, in the absence of another, an SU is allowed to increase its transmission power up to the limit. This means that depending on the idle length periods, SU will be able to achieve better transmission rates than at fixed allocation rates.

Simultaneous transmission by two or more SUs on the same channel in an underlay CRN is proposed. Firstly, the network layout must be defined to model a realistic scenario. Interference is of concern here and harmful interference occurs when a receiver is unable to decipher the intended signal from any other signals in the vicinity. Receivers typically work by detecting the strongest signal and then filtering out all other signals that are at least 3 dB below the strongest one. This allows a threshold to be determined and enforced to prevent any interference on a receiver. The threshold will be determined based on the desired signal strength and the 3 dB point.

Transmissions are affected by multiple factors, such as multipath and shadowing. This will degrade a transmitted signal and hence it is necessary to understand the possible path that a signal may take. What is ultimately important, is the signal power that arrives at the transmitter. A signal will also degrade with distance depending on the channel characteristics. There are some ways that can be used to mitigate interference issues. Directional antennas can also be used to help reduce interference. Thus by keeping all non-intended receivers away from this direction interference can be minimised. Positioning and distance are thus some of the factors that can be utilised to combat interference.

Interference signals are also constructive, i.e. interference from multiple sources will have a cumulative effect on the receiver. Hence it is crucial that any secondary transmitters in the network be aware of each other. For this research a system consisting of a single primary transmitter receiver pair and only two secondary transmitter receiver pairs was considered.

Fig. 3.1 shows the suggested network layout. In the figure two SU transmitter-receivers pairs are shown. The network also contains a PU receiver and the links for the pairs are shown. SU_1 will transmit to its receiver with channel coefficient h_{s1}. Similarly, SU_2 will transmit to its receiver with channel coefficient h_{s2}. However, because of the presence of the PU receiver and depending on its proximity, there will interference from both SU transmissions to the PU. These are given in the figure as h_{p1} and h_{p2}. These coefficients will play a role in determining the allowable transmission power on both SUs because ultimately they play a role in how much interference reaches the PU receiver.

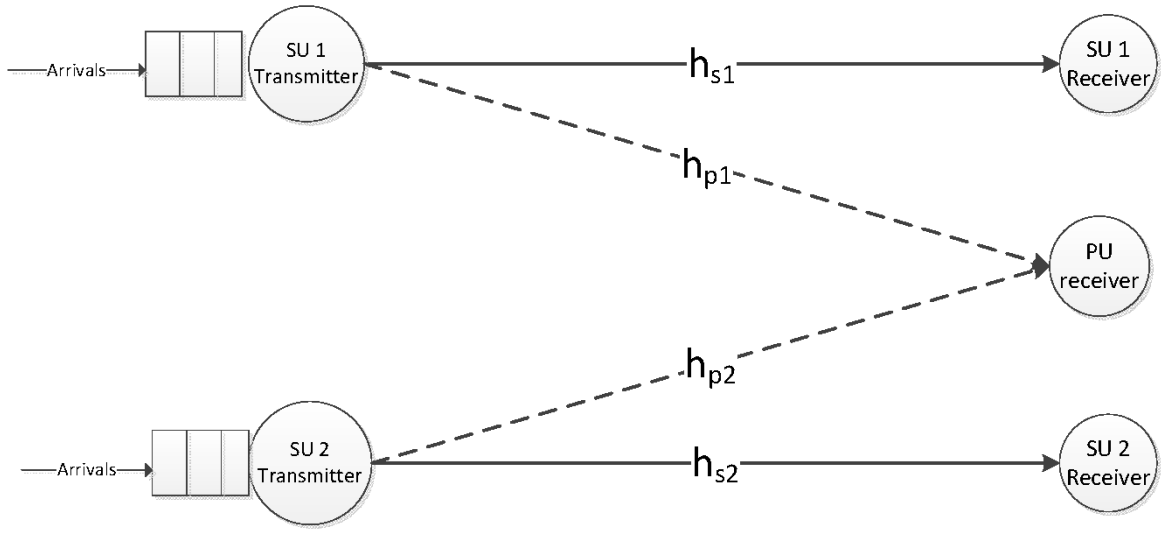

Figure 3.1. Network layout.

3.4 ASSUMPTIONS

The following assumptions were made during the system design:

- The PU will always be present in the system. The system is therefore analysed with a the interference constraint present. The motivation is that, if the system can function level in such a situation, it should follow that the system will work well when the PU is absent.
- The CRN approach is central. That is, all decisions will be taken by a main node. This node will be aware of the ITL and inform all SUs in the network.
- There is negligible SU-to-SU interference. This allows the transmission time to be derived mainly in terms of the ITL on the PU.

- The service type is exhaustive. SU queues will be served to completion.
- The ITL limit is adequate to serve at least two SUs reliably. The motivation is that the ITL cannot be so low that any SU communication will cause interference in the primary network. The ITL must be such that the SU has enough power to send and receive packets consistently.

3.5 TRANSMISSION TIME ANALYSIS

Paramount to the system model is the transmission time distribution. This is defined as the time that a packet will take to leave the transmitter and be received fully at the receiver. The transmission will represent the service time in the queueing analysis. The lower the transmission time the better the network quality.

Transmission takes place over wireless radio channels that differ according to their environment. This will have a great impact on the transmission time. Channel coefficients are used to define the nature of the channel, i.e. whether the channel is good or bad. The coefficients will vary from 0 to 1 with 1 indicating an ideal channel and 0 a very poor channel where transmission is impossible. While there are many factors that contribute to channel conditions such as multipath effects and Doppler effects the focus is on primarily distance in this work. Hence if a receiver is close to a transmitter, the channel coefficient will be higher than when it is further away. This power attenuation is described more accurately by the path loss phenomena [26].

Therefore let:

- h_{pi} be the channel coefficients for the channel SU_i to the PU receiver.
- h_{si} be the channel coefficients for the channels SU_i transmitter to SU_i receiver.
- Q be the interference power constraint imposed by the PU receiver.

Q is determined by the PU and all SUs in the network are aware of it. It is the limit for interference power reaching the PU receiver hence the need for the SUs to know the channel conditions for the links from them to the PU receiver.

Given this information the total received interference signal is calculated as:

$$\gamma_r = \sum_{i=1}^{\text{Number of SUs}} h_{pi} x_{si} + n_o \tag{3.1}$$

where x_{si} is the signal from SU_i with transmission power P_{si}. n_o is the additive white Gaussian noise (AWGN).

The transmission signal powers are to be constrained such that the total signal power reaching the PU receiver is less than Q. This can be done by ensuring than all SUs are allocated a portion of the allowable transmission power up to Q. Therefore, let r_i be the given allocation for SU_i such that:

$$\sum_{i=1}^{\text{Number of SUs}} r_i = 1. \tag{3.2}$$

r_i can take on any value between 0 and 1 inclusive. If an SU queue is not transmitting then its allocation will be 0 thus making the remaining SU allocation equal to 1. This means for a given allocation, an SU can has two states of transmission power, the allocated power level and the maximum power level when the other SU is not transmitting.

Taking into account the channel conditions and the allocation per SU, the transmission power for each SU must then be capped at:

$$P_{si} \leq \frac{r_i Q}{|h_{pi}|^2},\ i \in \{1,2...\text{Number of SUs}\}. \tag{3.3}$$

However, there are physical limitations on the transmission power of the SU transmitter. This is the highest amount that the transmitter can physically output, dependent on its hardware is denoted by P_{max_i}. It is possible than P_{max_i} may be less than P_{si}, hence the instantaneous SNR of an SU is given by:

$$\gamma_{si} = \min\{\frac{P_{max_i}|h_{si}|^2}{n_o}, \frac{r_i Q|h_{si}|^2}{n_o|h_{pi}|^2}\},\ i \in \{1,2...\text{Number of SUs}\}. \tag{3.4}$$

3.5.1 Shannon's theorem

Shannon's theorem relates the maximum capacity over a channel to the bandwidth, noise or interference and to transmission power over that channel [84].

Using the theorem, the transmission time for an SU is given by:

$$T_i = \frac{N}{B\log_2(1+\gamma_{si})},\ i \in \{1,2...\text{Number of SUs}\} \tag{3.5}$$

where B is the bandwidth of the transmission channel and N is the number of bits per packet. These will usually be dictated by an IEEE standard.

3.6 QUEUEING ANALYSIS

The goal of this research is to improve the QoS of the scenario, either by serving more SUs or by improving performance measures such as queue length or queue delay. Two models are proposed for the implementation. In both models the SUs will be given priority weighting by some network-defined criteria. Examples of criteria are cost, fairness, type of job and importance. Higher weighted SUs are given higher priority over lower weighted ones [85].

Head of line processor sharing is combined with priority discipline to develop the specific models. Transmission power is used as the basis for allocating the priority or weights to the SUs. A higher weighted SU will be given a larger share of the transmission power, hence a bigger capacity resulting in less transmission time. This means that higher weighted queues will have better QoS. When considering transmission time and that the SUs are initially transmitting at equal power and there are only two SUs (SU_1 and SU_2) in the system, SUs will be allowed to increase their transmission power to the interference limit when the other SU is absent. The implication is that at certain times an SU is transmitting at a power higher than normal. This therefore means that there would be improvement in service time.

Fig. 3.2 shows how the power allocations work through time. Here the transmission power of SU_1 varies with the presence of SU_2. If SU_2 is present the power is shared, else all the power is allocated to SU_1.

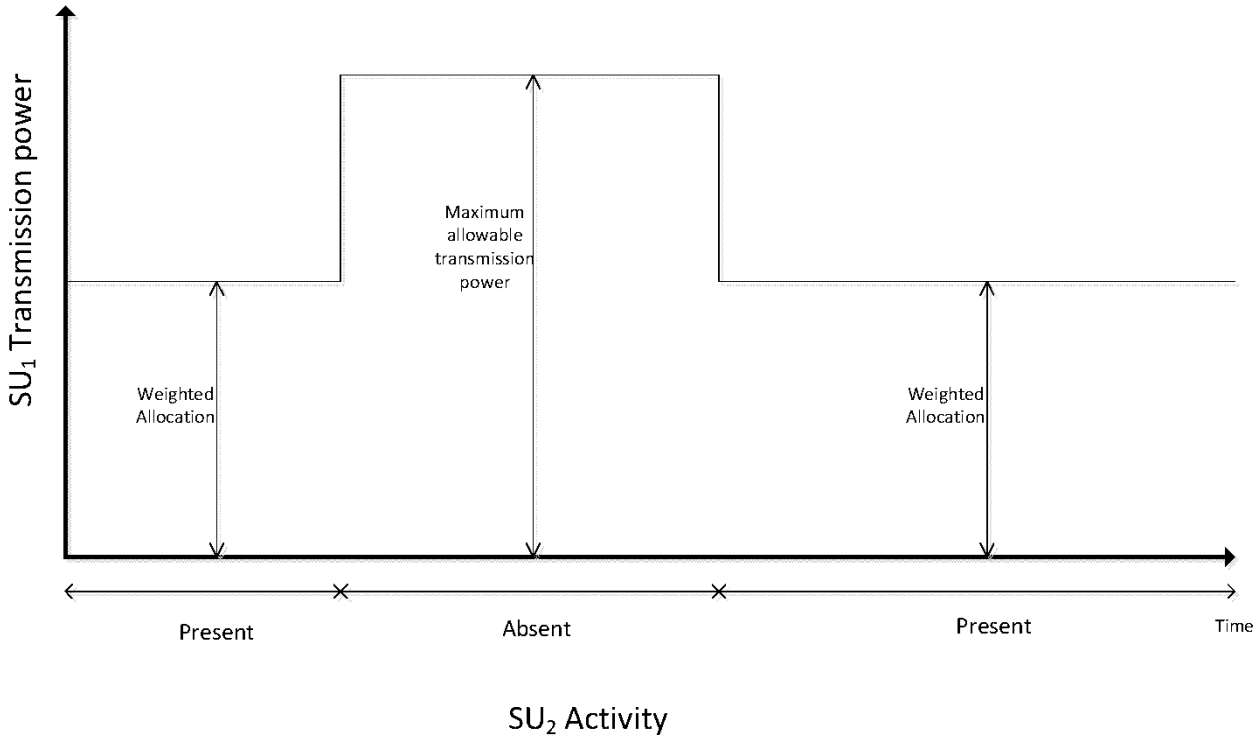

Figure 3.2. Power allocation as a result of queue activity.

The number of packets allowed in the system is infinite, however a buffer can be implemented, especially on the lower weighted SU of any size K. Therefore there can only be a maximum of K+1 lower weighted SU packets at any given time.

The model parameters are defined as indicated below. Let:

- λ_1 is the mean arrival rate of the SU_1 queue;
- λ_2 is the mean arrival rate of the SU_2 queue;
- μ_1 is the mean service rate of the SU_1 queue when there are no SU_2 packets;
- μ_3 is the mean service rate of the SU_2 queue when there are no SU_1 packets;
- μ_2 is the mean service rate of the SU_1 queue when there are SU_2 packets in the system; and

- μ_4 is the mean service rate of the SU_2 queue when there are SU_1 packets in the system.

Therefore:

$\lambda = \lambda_1 + \lambda_2$ is the mean arrival rate of the entire system. Let the higher weighted SU be HW and the lower weighted SU be LW. For this text Queue 1 will always be the HW and Queue 2 will be the LW.

3.6.1 Pre-emptive model

SUs in the network share the transmission power and will transmit at maximum allowable power in the absence of others. When SU_1 is not transmitting, SU_2 will transmit at full allowed power below the interference threshold. If SU_1 restarts transmission, the transmit power of SU_2 will immediately adjusted to the pre-allocated power levels and SU_1 will also begin transmitting. Both SUs behave the same hence similarly, SU_1 will transmit at maximum allowable power in absence of SU_2 and will also immediately adjust transmit power when SU_2 restarts transmission.

3.6.1.1 State space

Fig. 3.3 shows the state space for the pre-emptive queueing model. Let a state description be (h,j)

where

- h = number of HW packets in the system, h $\in 0, 1, 2, 3, ...$
- j = number of LW packets in the system, j $\in 0, 1, 2..., K+1$.

The squares represent the state in which the system currently is in. The arrows show valid and possible transitions from one state to the other. For example, to move from state (2,2) to state (2,1) there must be a service completion on the LW queue before any other event. Note that because of the continuous nature of the queue there cannot be simultaneous events. That is, there cannot be an arrival and a departure at the same time.

Consider an initially empty system and then an arrival at the LW queue, that is, λ_2. The system will now move to state (0,1). A packet may arrive at the HW queue and will begin service immediately. The system will move to state (1,1). From state (1,1) there can be a service completion to either the HW or the LW queue. The system will move to state (0,1) or (1,0) respectively. Only the transitions shown by the arrows are allowed.

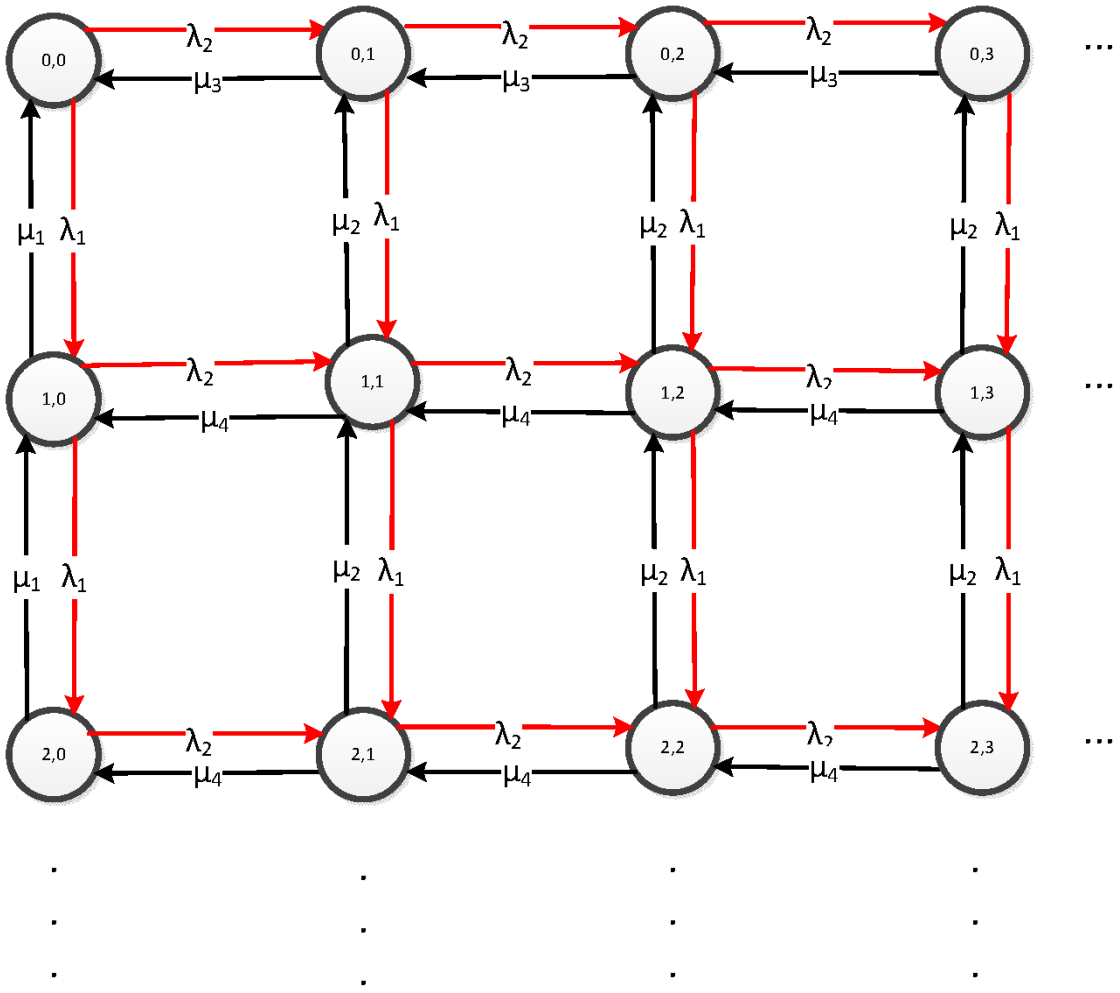

Figure 3.3. Pre-emptive model state space.

3.6.1.2 Transition matrix

The transition matrix is a quasi-birth-death (QBD) process. Define Q_{ij} as the rate of transition from state i to state j for the HW queue. Since there are two queues in the system the elements of Q will be

matrices to capture the rates of the LW queue. A simple way of looking at it is first to identify the state of the HW queue and then to check the transition rates of the LW queue in that state.

$$Q = \begin{bmatrix} B & C & & & & \\ E & A_1 & A_0 & & & \\ & A_2 & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \end{bmatrix} \tag{3.6}$$

where

$$B = \begin{array}{c} \\ 0,0 \\ 0,1 \\ 0,2 \\ \vdots \\ 0,K+1 \end{array} \begin{array}{c} \begin{array}{ccccc} 0,0 & 0,1 & 0,2 & \dots & 0,K+1 \end{array} \\ \begin{pmatrix} -\lambda & \lambda_2 & & & \\ \mu_3 & -(\lambda+\mu_3) & \lambda_2 & & \\ & \mu_3 & -(\lambda+\mu_3) & \lambda_2 & \\ & & \ddots & \ddots & \ddots \\ & & & \mu_3 & -(\lambda_1+\mu_3) \end{pmatrix} \end{array}$$

$$C = \begin{array}{c} \\ 0,0 \\ 0,1 \\ 0,2 \\ \vdots \\ 0,K+1 \end{array} \begin{array}{c} \begin{array}{ccccc} 0,0 & 0,1 & 0,2 & \dots & 0,K+1 \end{array} \\ \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_1 & & & \\ & & \lambda_1 & & \\ & & & \ddots & \\ & & & & \lambda_1 \end{pmatrix} \end{array}$$

$$A_0 = \begin{array}{c} \\ n,0 \\ n,1 \\ n,2 \\ \vdots \\ n,K+1 \end{array} \begin{array}{c} \begin{array}{ccccc} n,0 & n,1 & n,2 & \dots & n,K+1 \end{array} \\ \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_1 & & & \\ & & \lambda_1 & & \\ & & & \ddots & \\ & & & & \lambda_1 \end{pmatrix} \end{array} n \in 1,2,3...$$

$$E = \begin{array}{c} \\ 1,0 \\ 1,1 \\ 1,2 \\ \vdots \\ 1,K+1 \end{array} \begin{array}{c} \begin{array}{ccccc} 1,0 & 1,1 & 1,2 & \dots & 1,K+1 \end{array} \\ \begin{pmatrix} \mu_1 & & & & \\ & \mu_2 & & & \\ & & \mu_2 & & \\ & & & \ddots & \\ & & & & \mu_2 \end{pmatrix} \end{array}$$

$$A_2 = \begin{array}{c} \\ n,0 \\ n,1 \\ n,2 \\ \vdots \\ n,K \end{array} \begin{array}{c} \begin{array}{ccccc} n,0 & n,1 & n,2 & \dots & n,K+1 \end{array} \\ \begin{pmatrix} \mu_1 & & & & \\ & \mu_2 & & & \\ & & \mu_2 & & \\ & & & \ddots & \\ & & & & \mu_2 \end{pmatrix} \end{array} n \in 2,3,4...$$

$$A_1 = \begin{array}{c} \\ n,0 \\ n,1 \\ n,2 \\ \vdots \\ n,K+1 \end{array} \begin{array}{c} \begin{array}{ccccc} n,0 & n,1 & n,2 & \dots & n,K+1 \end{array} \\ \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_2 & & & \\ \mu_4 & -(\lambda+\mu_b) & \lambda_2 & & \\ & \mu_4 & -(\lambda+\mu_b) & \lambda_2 & \\ & & \ddots & \ddots & \ddots \\ & & & \mu_4 & -(\lambda_1+\mu_b) \end{pmatrix} \end{array}$$
$$n \in 1,2,3...$$

3.6.1.3 Rate matrix

The Rate matrix is the minimal non-negative solution to:

$$0 = A_0 + RA_1 + R^2A_2. \tag{3.7}$$

The stationary distribution, $\boldsymbol{\pi}$ of the Markov chain can now be obtained through [86]:

$$\begin{aligned} \pi_0(B+RA_2) &= \mathbf{0} \\ \pi_0(I-R)^{-1}\mathbf{1} &= 1 \end{aligned} \tag{3.8}$$

and

$$\begin{aligned} \pi_{k+1} &= \pi_k R \\ \sum \boldsymbol{\pi} &= 1. \end{aligned} \tag{3.9}$$

The R-matrix has a somewhat special structure similar to the one developed by Miller given in equation 3.10 [87]. The rows are repeating because the LW arrivals are independent of the number of HW packets in the system [7]. The special structure can be advantageous in developing an efficient method for calculating the state probabilities.

$$R = \begin{bmatrix} r_0 & r_1 & r_2 & r_3 & r_4 & \dots \\ 0 & r_0 & r_1 & r_2 & r_3 & \dots \\ 0 & 0 & r_0 & r_1 & r_2 & \dots \\ 0 & 0 & 0 & r_0 & r_1 & \dots \\ 0 & 0 & 0 & 0 & r_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix} \tag{3.10}$$

In the given model LW packets can reduce whilst there are HW packets in the system because they are both being served at the same time. The matrix will no longer be upper triangular. However, a repetitious structure can still be observed and utilised to develop an efficient method for obtaining the steady state probabilities. The structure is given in equation 3.11.

$$R = \begin{bmatrix} k_0 & k_1 & k_2 & k_3 & k_4 & k_5 & k_6 & \dots \\ h_0 & r_0 & r_1 & r_2 & r_3 & r_4 & r_5 & \dots \\ h_1 & w_0 & r_0 & r_1 & r_2 & r_3 & r_4 & \dots \\ h_2 & w_1 & w_0 & r_0 & r_1 & r_2 & r_3 & \dots \\ h_3 & w_2 & w_1 & w_0 & r_0 & r_1 & r_2 & \dots \\ h_4 & w_3 & w_2 & w_1 & w_0 & r_0 & r_1 & \dots \\ h_5 & w_4 & w_3 & w_2 & w_1 & w_0 & r_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix} \tag{3.11}$$

3.6.2 Non-pre-emptive model

An arriving class will not pre-empt the service of the class already in service. When SU_1 is transmitting and SU_2 packets arrive, the transmission powers will not be immediately adjusted to accommodate SU_2. Hence the service of the SU packet already in the system will be completed at full allowable power or the power with which it started service. The transmission power will only be adjusted after the packet in service has been completed. The two classes of SUs will continue to be served at the shared rate until one of the SUs has no packets left to be served. The packets of the remaining SU will then be served at full power until the other SU restarts transmission. This applies to both HW and LW users.

3.6.2.1 State diagram

The state space of the proposed non-pre-emptive queue model is rather dense because of the need to keep track of the SU currently in service. Let state (h,j,i) {h = number of HW packets in the system, j = number of LW packets in the system, i = the type (HW or LW) of a packet currently in service}. x $\in 0,1,2,3,....,$ j $\in 0,1,2,....,K+1$.

$$i = \begin{cases} 0, & \textit{when both classes are in service.} \\ 1, & \textit{when an HW user is in service.} \\ 2, & \textit{when an LW user is in service.} \end{cases}$$

The circles in Fig. 3.4 represent the state in which the system is currently and the squares are dummy states used to determine which SU packet is currently being served. It is necessary to keep track of the current packet as this packet will need to be served to completion before any other packet is served. There can also still be more arrivals at both queues. In order to provide a clearer representation two scenarios are given below to show the progression of the queue between states.

Scenario one

Consider an initially an empty system (no SU in service or in the queues) that begins service with an LW queue arrival (λ_2). The system moves to state (0,1,2). Before service has been completed on

packet one, a second packet arrives but on the HW queue (λ_1). The system must continue serving packet one at maximum allowable power and must note the arrival of a packet at the HW queue. The system moves to state (1,1,2). The first packet service is now completed at full power (μ_3), leaving the packet on the HW alone in the system to begin service at full power. The system moves to state (1,0,1). A third packet arrives in the system (λ_2) on the LW queue before the HW packet has finished service. The system moves to state (1,1,1) indicating that an HW packet is currently in service at maximum allowable power. The HW packet completes service at maximum allowable power (μ_1) leaving the newly arrived LW packet alone in the system. The system moves to state (0,1,2). Finally the last LW packet is served at maximum allowable power (μ_3), leaving the system empty again, state (0,0,0). Fig. 3.4 shows this scenario.

The events described above are summarized as follows: $\lambda_2 \rightarrow \lambda_1 \rightarrow \mu_3 \rightarrow \lambda_2 \rightarrow \mu_1 \rightarrow \mu_3$.

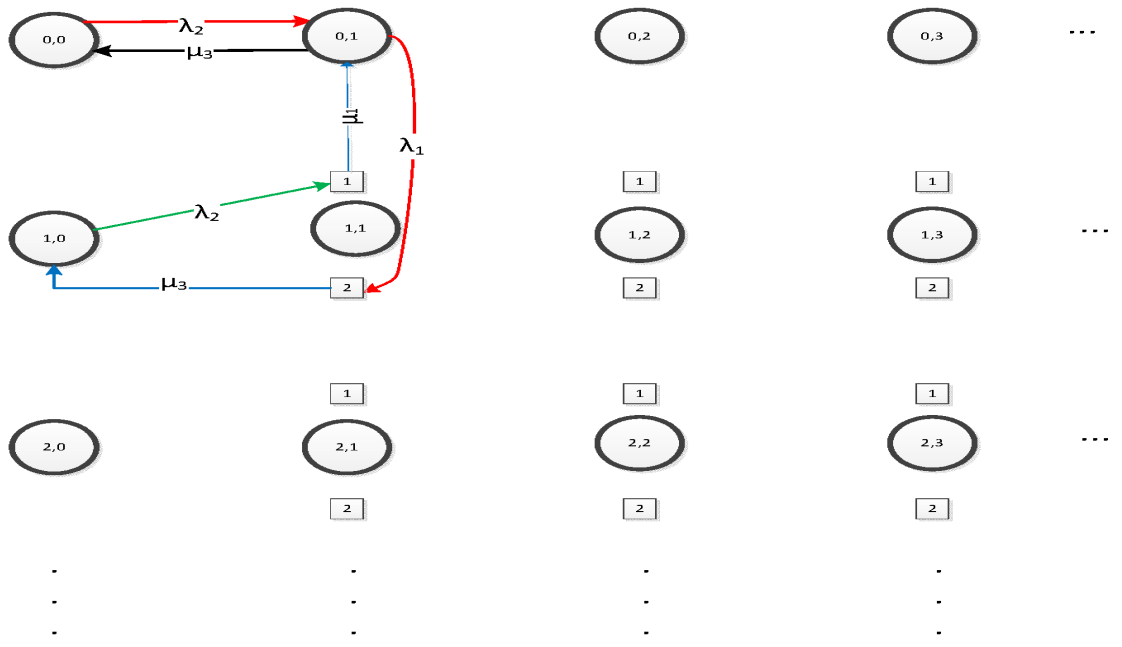

Figure 3.4. Non-pre-emptive model state space showing scenario one.

Scenario two

Consider initially an empty system that begins service with an LW queue arrival (λ_2). The system moves to state (0,1,2). Before service has been completed on packet one, a second packet arrives but

on the HW queue (λ_1). The system must continue serving packet one at maximum allowable power and must note the arrival of a packet at the HW queue. The system moves to state (1,1,2). The first packet service is now completed at full power (μ_3), leaving the packet on the HW alone in the system to begin service at full power. The system moves to state (1,0,1). A third packet arrives in the system (λ_2) on the LW queue before the HW packet has finished service. The system moves to state (1,1,1), indicating that an HW packet is currently in service at maximum allowable power. Another packet arrives at the LW queue (λ_2) before service is completed. The system moves to state (1,2,1), indicating that there is one HW and two LW packets in the system and that the HW packet is enjoying maximum allowable service. An HW packet arrives (λ_1) and the system moves to state (2,2,1). The service of the first HW packet is finally completed (μ_3) and the system moves to state (1,2,0), indicating one HW and two LW packets and that both queues are being served. An LW packet completes service at its allocated rate (μ_4). Now one HW packet and one LW packet are being served. The system moves to state (1,1,0). The second LW packet completes service (μ_4), leaving only the HW queue. The system moves to state (1,0,1). Again before service is completed, an LW packet arrives (λ_2) and the system once again goes to state (1,1,1). The HW packet completes service at maximum allowable power (μ_1). Finally the last LW packet is served at maximum allowable power (μ_3), returning the system to state (0,0,0). Fig. 3.5 shows this scenario.

The events described above are summarised as follows: $\lambda_2 \rightarrow \lambda_1 \rightarrow \mu_3 \rightarrow \lambda_2 \rightarrow \lambda_2 \rightarrow \lambda_1 \rightarrow \mu_1 \rightarrow \mu_4 \rightarrow \mu_4 \rightarrow \lambda_2 \rightarrow \mu_1 \rightarrow \mu_3$.

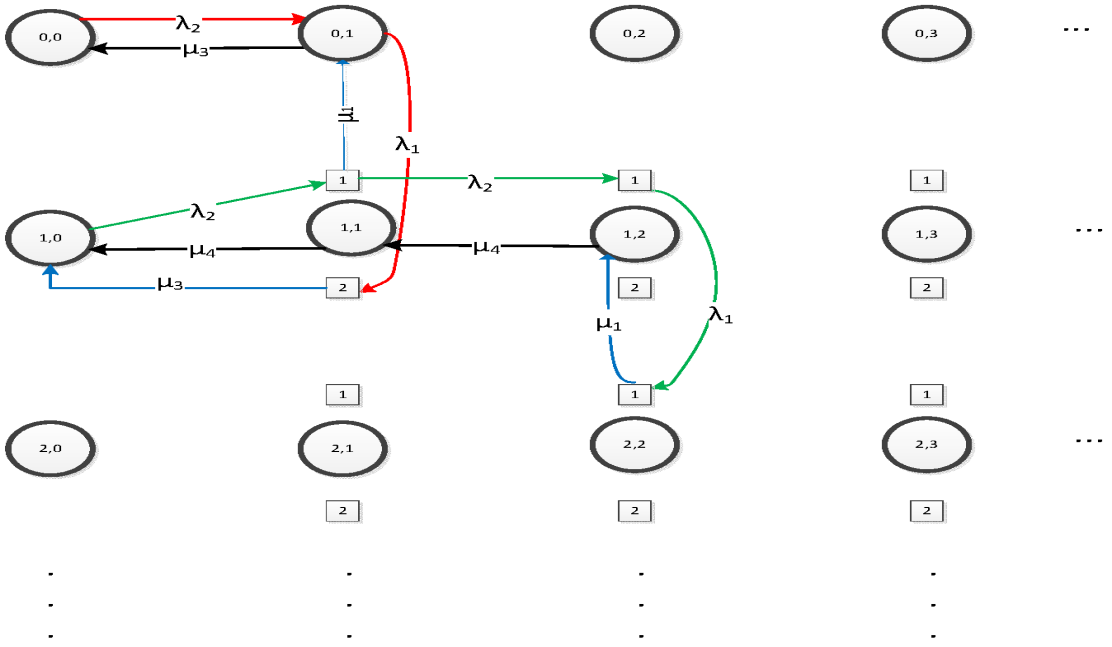

Figure 3.5. Non-pre-emptive model state space showing scenario two.

The state space for the entire system complete with all possible transitions is shown in Fig. 3.6

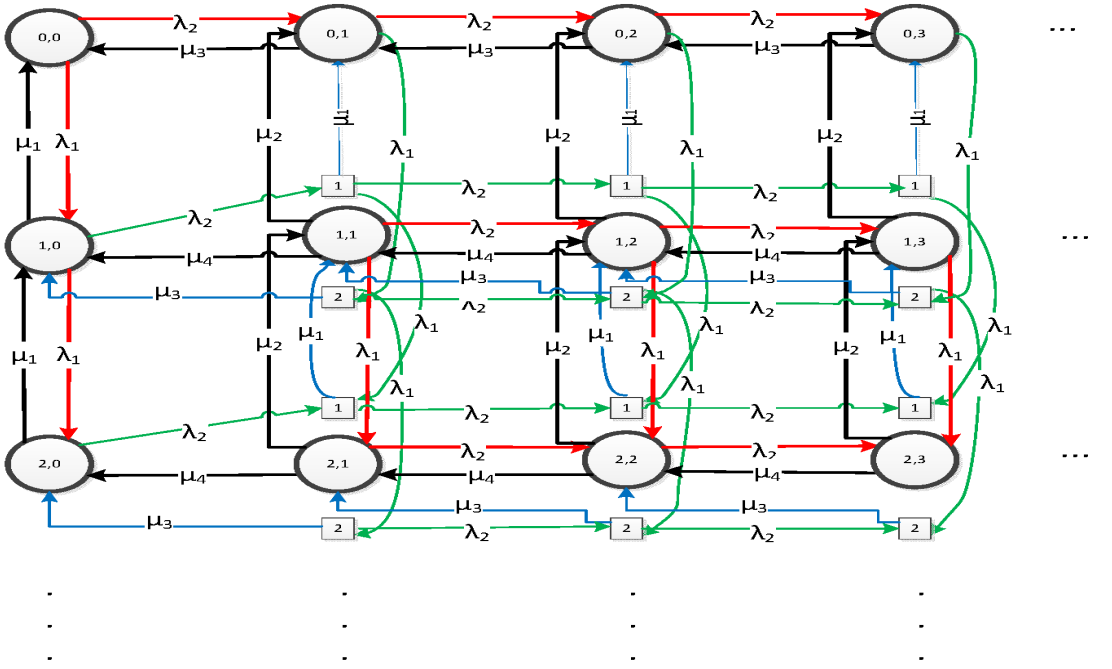

Figure 3.6. Non-pre-emptive model state space.

3.6.2.2 Transition matrices

The transition matrix is a QBD process. Define Q_{ij} as the rate of transition from state i to state j for the HW queue. Because there are two queues in the system, the elements of Q will be matrices as well. in addition, the dummy states have to included in the transition matrix. This will cause the elements of the inner matrices to be matrices as well. Another way of looking at it is that the outer matrix represents the states of the HW queue, the first inner matrix represent the states of LW and the innermost matrix represents the SU currently being served.

$$Q = \begin{bmatrix} B & C & & & \\ E & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \tag{3.12}$$

where

$$B = \begin{array}{c} \\ 0,0 \\ 0,1 \\ 0,2 \\ \vdots \\ 0,K+1 \end{array} \begin{array}{c} \begin{array}{ccccc} 0,0 & 0,1 & 0,2 & \dots & 0,K+1 \end{array} \\ \begin{pmatrix} B_1 & B_2 & & & \\ B_3 & B_4 & B_2 & & \\ & B_6 & B_4 & B_2 & \\ & & \ddots & \ddots & \ddots \\ & & & B_6 & B_5 \end{pmatrix} \end{array}$$

$$B_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} -\lambda & & \\ & & \\ & & \end{pmatrix} \end{array} \quad B_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & \lambda_2 & \\ & & \\ & & \end{pmatrix} \end{array}$$

$$B_3 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & & \\ & & \\ \mu_3 & & \end{pmatrix} \end{array} \quad B_4 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & & \\ & & \\ & & -(\mu_3+\lambda) \end{pmatrix} \end{array}$$

$$B_5 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & & \\ & & \\ & & -(\mu_3+\lambda_1) \end{pmatrix} \end{array} \quad B_6 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & & \\ & & \\ & & \mu_3 \end{pmatrix} \end{array}$$

$$C = \begin{array}{c} \\ 0,0 \\ 0,1 \\ 0,2 \\ \vdots \\ 0,K+1 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0,0 & 0,1 & 0,2 & \dots & 0,K+1 \end{array} \\ \begin{pmatrix} C_1 & & & & \\ & C_2 & & & \\ & & C_2 & & \\ & & & \ddots & \\ & & & & C_2 \end{pmatrix} \end{array}$$

$$C_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & \lambda_1 & \\ & & \\ & & \end{pmatrix} \end{array} \quad C_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & & \\ & & \\ & & \lambda_1 \end{pmatrix} \end{array}$$

$$E = \begin{array}{c} \\ 1,0 \\ 1,1 \\ 1,2 \\ \vdots \\ 1,K+1 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1,0 & 1,1 & 1,2 & \dots & 1,K+1 \end{array} \\ \begin{pmatrix} E_1 & & & & \\ & E_2 & & & \\ & & E_2 & & \\ & & & \ddots & \\ & & & & E_2 \end{pmatrix} \end{array}$$

$$E_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & & \\ \mu_1 & & \\ & & \end{pmatrix} \end{array} \quad E_2 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & & \\ & \mu_2 & \\ & \mu_1 & \end{pmatrix} \end{array}$$

$$A_2 = \begin{array}{c} \\ n,0 \\ n,1 \\ n,2 \\ \vdots \\ n,K+1 \end{array} \begin{array}{c} \begin{array}{ccccccc} n,0 & n,1 & n,2 & \dots & n,K+1 \end{array} \\ \begin{pmatrix} A_{21} & & & & \\ & A_{22} & & & \\ & & A_{22} & & \\ & & & \ddots & \\ & & & & A_{22} \end{pmatrix} \end{array}$$

$$n \in 2,3,4,...$$

$$A_{21} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} \mu_1 & & \\ & \mu_1 & \\ & & \mu_1 \end{pmatrix} \end{array} A_{22} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} \mu_2 & & \\ \mu_1 & & \\ & & \end{pmatrix} \end{array}$$

$$A_1 = \begin{array}{c} \\ n,0 \\ n,1 \\ n,2 \\ \vdots \\ n,K+1 \end{array} \begin{array}{c} \begin{array}{ccccccc} n,0 & n,1 & n,2 & \dots & n,K+1 \end{array} \\ \begin{pmatrix} A_{11} & A_{15} & & & \\ A_{13} & A_{12} & A_{15} & & \\ & A_{13} & A_{12} & A_{15} & \\ & & \ddots & \ddots & \ddots \\ & & & A_{13} & A_{14} \end{pmatrix} \end{array}$$

$$n \in 1,2,3,...$$

$$A_{11} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & & \\ & -(\mu_1+\lambda) & \\ & & \end{pmatrix} \end{array}$$

$$A_{13} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} & \mu_4 & \\ & & \\ & \mu_3 & \end{pmatrix} \end{array} A_{15} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} \lambda_2 & & \\ & \lambda_2 & \\ & & \lambda_2 \end{pmatrix} \end{array}$$

$$A_{12} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} -(\lambda+\mu_b) & & \\ & -(\lambda+\mu_1) & \\ & & -(\lambda+\mu_3) \end{pmatrix} \end{array}$$

$$A_{14} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} -(\lambda_1+\mu_b) & & \\ & -(\lambda_1+\mu_1) & \\ & & -(\lambda_1+\mu_3) \end{pmatrix} \end{array}$$

$$A_0 = \begin{array}{c} \\ n,0 \\ n,1 \\ n,2 \\ \vdots \\ n,K+1 \end{array} \begin{array}{c} \begin{array}{ccccccc} n,0 & n,1 & n,2 & \dots & n,K+1 \end{array} \\ \begin{pmatrix} A_{01} & & & & \\ & A_{01} & & & \\ & & A_{01} & & \\ & & & \ddots & \\ & & & & A_{01} \end{pmatrix} \end{array}$$

$$n \in 1,2,3,...$$

$$A_{01} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} \lambda_1 & & \\ & \lambda_1 & \\ & & \lambda_1 \end{pmatrix} \end{array}$$

3.6.2.3 Rate matrix

The R-matrix is the minimal non-negative solution to equation 3.7.

The difference in this case is that the R-matrix now consists of matrices. The stationary distribution of the Markov chain can now be obtained through equations 3.8 and 3.9.

Normalisation can be carried out to find π_0, the stationary distribution.

$$R = \begin{bmatrix} K_0 & K_1 & K_2 & K_3 & K_4 & K_5 & K_6 & \dots \\ H_0 & R_0 & R_1 & R_2 & R_3 & R_4 & R_5 & \dots \\ H_1 & W_0 & R_0 & R_1 & R_2 & R_3 & R_4 & \dots \\ H_2 & W_1 & W_0 & R_0 & R_1 & R_2 & R_3 & \dots \\ H_3 & W_2 & W_1 & W_0 & R_0 & R_1 & R_2 & \dots \\ H_4 & W_3 & W_2 & W_1 & W_0 & R_0 & R_1 & \dots \\ H_5 & W_4 & W_3 & W_2 & W_1 & W_0 & R_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix} \tag{3.13}$$

where

$$K_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} k_{i1} & p_{i1} & 0 \\ p_{i2} & k_{i2} & 0 \\ p_{i3} & p_{i4} & k_{i3} \end{pmatrix} \end{array}$$

$$H_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} 0 & h_{i1} & 0 \\ 0 & h_{i2} & 0 \\ 0 & h_{i4} & 0 \end{pmatrix} \end{array}$$

$$R_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} r_{i1} & s_{i1} & 0 \\ s_{i2} & r_{i2} & 0 \\ s_{i3} & s_{i4} & r_{i3} \end{pmatrix} \end{array}$$

$$W_i = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} w_{i1} & t_{i1} & 0 \\ t_{i2} & w_{i2} & 0 \\ t_{i3} & s_{i4} & 0 \end{pmatrix} \end{array}$$

3.7 OPTIMISATION

Given the model, there are two approaches that may be utilised for optimisation of the model. The first is focusing on the optimisation of a network parameter such as transmission time for both SUs and the second is focusing on a performance measure such as the time spent in the system. The objective function will consequently be based on either of these two. The system constraints will have to be adjustable network parameters such that the network can vary them as necessary to achieve optimisation. From the model above only the power allocation ratio, r_i, can be adjusted. The other parameters, such as the interference limit and the channel coefficients, are environment-dependent and hence cannot be adjusted by the network. For the case in this work, the optimisation problem is a minimization problem. By definition, the optimal solution to a minimisation problem is the lowest objective function value within the feasible region [88].

Here an objective function is defined and general constraint formulation given. This is a previously seen problem in CRNs. A general objective function for a resource allocation problem is given below [81]. The objective function z is:

$$\text{minimise } z = f(x,y). \tag{3.14}$$

General constraints are:

$$\text{subject to } g_i(x,y) \leq b_i i = 1,....m \tag{3.15}$$

$$x_k \geq 0, k = 1,2,....,r \tag{3.16}$$

$$y_j \geq 0, j = 1,2,....,r. \tag{3.17}$$

However, for the given model above the objective function z is not straightforward to derive. The two SUs are dependent on each other regardless of the constraint; this results in difficulty in developing a closed form objective function. In addition, given that the optimisation of a performance measure is more desirable queueing theory is used to develop and determine the queue behaviour. Queuing theory will allow performance measures

such as waiting time and number of packets in the network to be obtained. Therefore, the already developed queueing models are used to determine $x, y, f(x,y), g_i, b_i$ etc. Optimisation will then be done according to the results of the queueing models.

3.7.1 Convex optimisation

Convex programming was found to be adequate and efficient for the given problem. Firstly, the objective function and the constraints need to be mapped to the queueing models. Therefore the objective is to minimise the mean number of packets in the system subject to the power allocation ratio. Equation 3.14 now defines a performance measure. That is, z is now the total average number in the queue for the whole system.

$$\min z = f(x,y) = x + y \tag{3.18}$$

x and y are now the average number of packets in the HW and LW queues respectively. $g_i(x,y)$ is now the power allocation ratio r_i, which is the only constraint in the system. Equation 3.15 now becomes:

$$\text{s.t. } r_1 + r_2 = 1. \tag{3.19}$$

Equations 3.16 and 3.17 remain unchanged because there cannot be fewer than zero packets in the queues at any time. For convex optimisation to be valid, the function $f(x,y)$ has to be convex and satisfy:

$$f(\alpha w_1 + (1-\alpha)w_2) \leq \alpha f(w_1) + (1-\alpha) f(w_2) \tag{3.20}$$

where $w = (x,y)$ and $0 \leq \alpha \leq 1$.

Equation 3.20 describes a situation where x and y, the average number of packets in the HW and LW queues respectively, are mapped onto a single function by w. The equation describes that the value of a point between w_1 and w_2 must be less than the corresponding value of that point on a straight line connecting w_1 and w_2.

Also let $f(x)$ and $f(y)$ be equal to x and y respectively. Given that $f(x)$ and $f(y)$ are convex their sum, given weights, will also be convex. Let the weight associated with the HW queue be ϕ and β for the LW queue.

The power allocation ratio, r, has a significant effect on the system performance. The ratio affects the service rates, hence the traffic intensity. This leads to the average number and waiting times in the queues being affected and this in turn means the optimisation is affected. The purpose, though, is to determine a power allocation ratio point that after taking all this into account, will allow an optimal solution to be found.

The power allocation ratio, r, will be varied by decreasing the allocation to the HW queue and therefore increasing the allocation to the LW queue by the same amount. This is to find how best the power allocation can be distributed in order to find the best performing region. Let θ be a variable such that $r_1 = 1 - \theta$ and $r_2 = \theta$. Fig. 3.7 shows the functions $f(x)$ and $f(y)$ varied across θ. Therefore, $f(x)$ and $f(y)$ can be considered convex. As θ increases, more power is allocated to Queue 2 and therefore Queue 2 performs better all things being equal. This can be seen in the figure, as the average length of Queue 2 is now less than that of Queue 1 when θ is greater than 0.5.

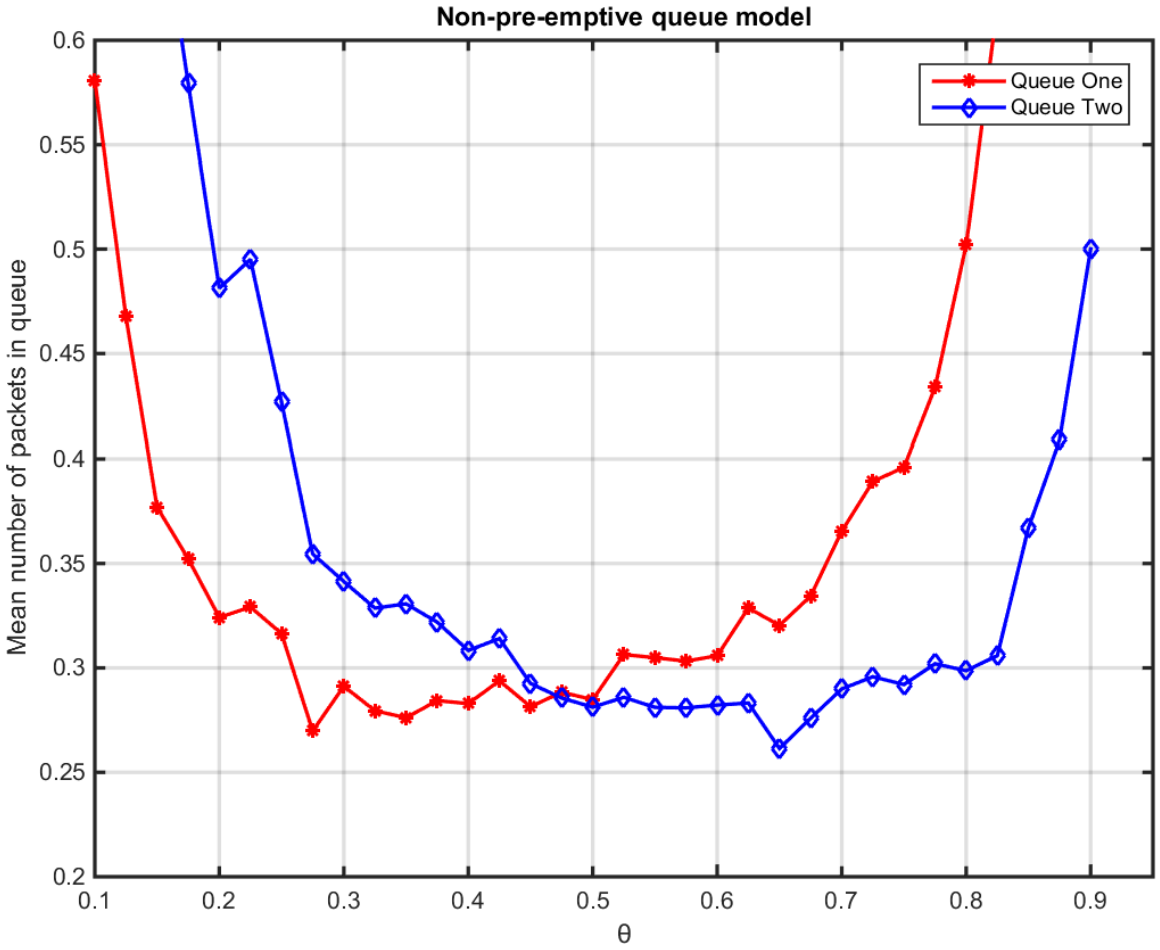

Figure 3.7. Mean number of packets for Queue 1 and 2 showing the convex property.

The system presents a feedback issue whereby the power allocation ratio affects the system performance, which in turn affects the optimisation results. That implies that adjusting the allocation ratio due to previous optimisation results will lead to change in service times and hence the system performance.

3.8 CONVEXITY OF RESULTS

Fig. 3.7 shows the result on system performance of varying the power allocation ratio through the range of possible values of the allocation ratio. The results fit a pseudo-convex function. Pseudo-convex functions are sufficient to be deemed convex with regard to determining the minimum point. A function can generally be considered pseudo-convex if it is increasing in a direction with positive directional derivative [89]. However, to further illustrate the point that the results are indeed sufficient for convexity to be established, a technique developed by Whitt is used [90]. The technique employs observation points in order to determine the pattern or shape of given scaling and the number of plot points can be changed. The scaling can be achieved adjusting the vertical scales so as to limit the observable region to only the relevant data. That means that considering only the first queue f_x the scaling becomes:

$$plot(\{f_x : 0 \leq k \leq n\}) \equiv plot(\{(f_x - min)/range : 0 \leq k \leq n\}), \tag{3.21}$$

where

$$min \equiv \min\ f_x : 0 \leq k \leq n$$

and

$$range \equiv \max\ f_x - \min f_x : 0 \leq k \leq n.$$

n is the number of plot points that also determines the horizontal scaling, since the plot points are spaced by $\frac{1}{n}$.

Simulations were then made to conclude that the function f_x (and similarly f_y) is adequate to be determined convex. Fig. 3.8 shows the function f_x with $n = 40$. It is still not clear if convexity can be established, the 40 points are not enough. Fig. 3.9 shows the same function, but with $n = 1600$.

From Fig. 3.9 it can be seen that convexity has now been achieved by applying the technique of equation 3.21. The results are conclusive enough for convexity to be established and hence convex optimisation to be carried out.

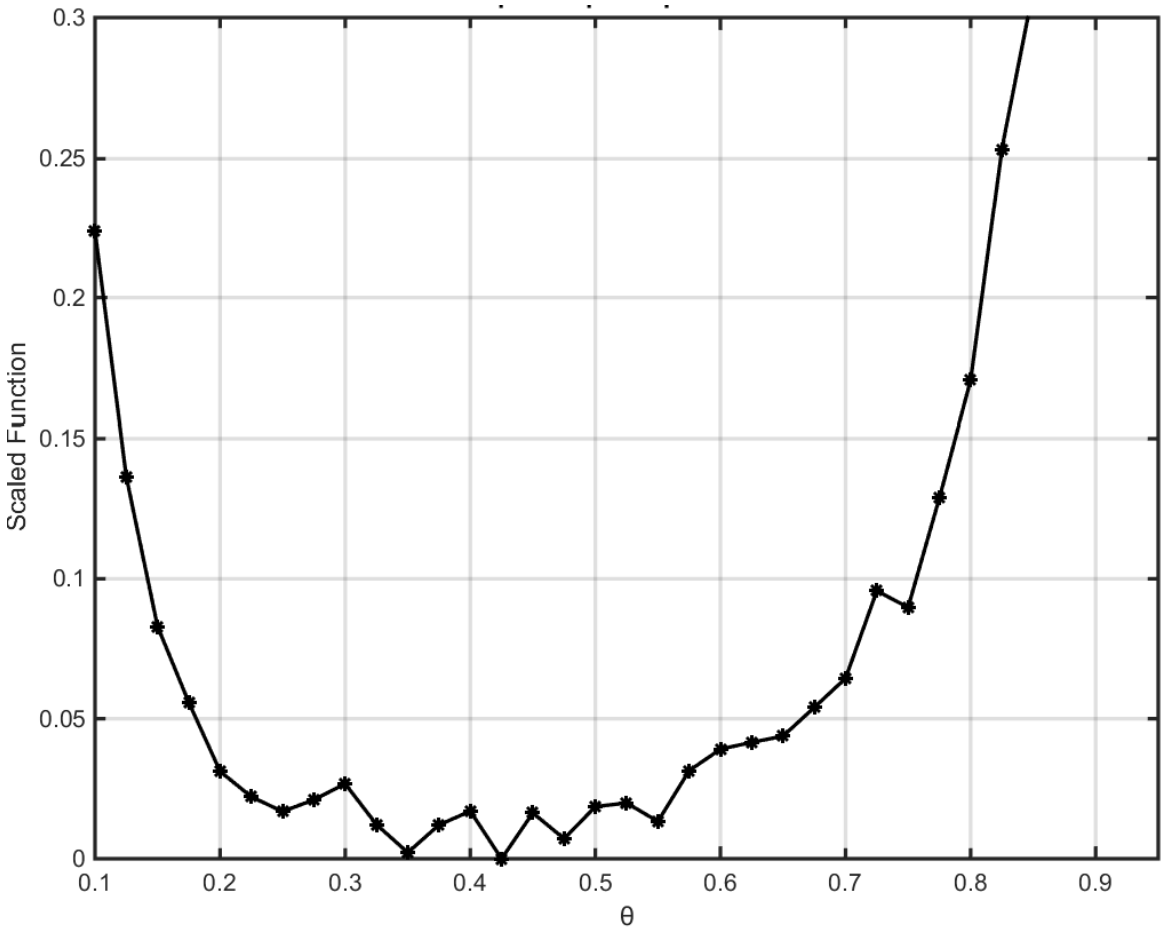

Figure 3.8. Queue 1 length vs. allocation power for $n = 40$.

3.9 CONCLUSION

Interference power is the major constraint of the SU and transmission power is allocated according to weights assigned to SUs. Two models are proposed that behave differently in terms of service to the SUs. Transmission times which are then also the service times are determined from equation 3.5 according to the respective channel coefficients and the allocated power. In the absence of packets in one queue, the other queue's allocation becomes unity.

The proposed schemes can be used for urban areas where there are licensed spectrum users with the necessary infrastructure already in place. The ISM band in urban areas is beginning to become congested with the increase of home based transmitters. The proposed models provide a way for these transmitters, to (1) co-exist with licensed user and (2) extend their range.

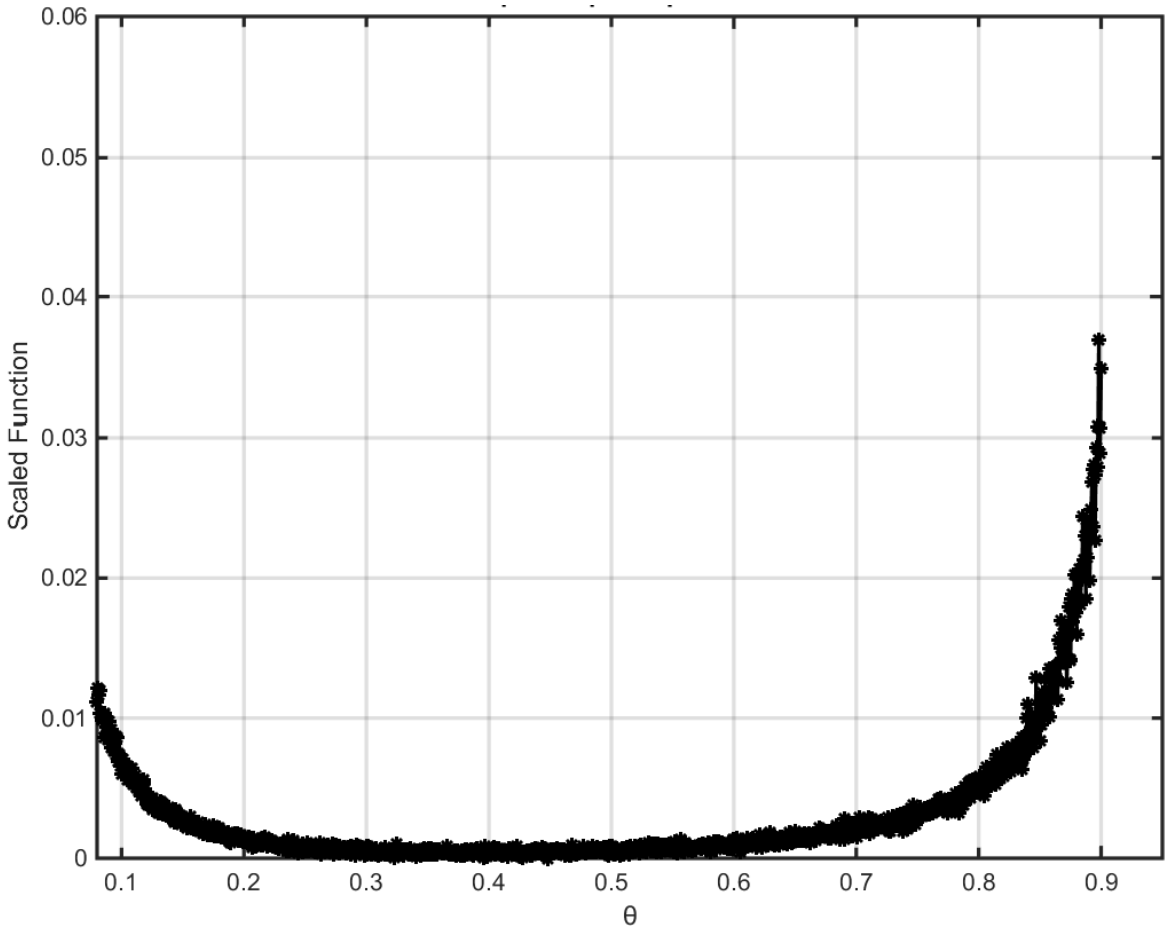

Figure 3.9. Queue 1 length vs. allocation power for $n = 1600$.

CHAPTER 4 RESULTS

4.1 CHAPTER OVERVIEW

The pre-emptive and non-pre-emptive queue models are analysed in depth. Firstly, simulation conditions are set up and explained. Simulations are then run in these conditions for each model. Variations are applied to observe the queue behaviour in different conditions. The results are primarily based on performance measures namely queue length and waiting times.

4.2 SIMULATION SET UP

From the state space of the models in Chapter 3 the simulation can now be set up. An algorithm that simulates the arrival and departure process of the system given the rules of the model, was developed and is given in this chapter. The simulation was allowed to run until steady state was reached so as to capture the non-transient behaviour of the system.

The results aim to show the effect that the added SU will have on the incumbent SU and also how the added SU will cope with the incumbent SU. Therefore, most of the results are of an SU performance being compared to the varying activity of the other SU under different scenarios. Fig. 4.1 shows a flowchart of the simulation procedure. The simulation is repeated for firstly, different traffic intensities on SU_1 and then secondly for different power allocation ratios. The flowchart shows that after setting the power allocation and SU_1 traffic intensity the traffic intensity, for SU_2 is then varied from just above zero to just below one. This way the results can be compared for many different scenarios while holding the allocation ratio and SU_1 traffic intensity constant.

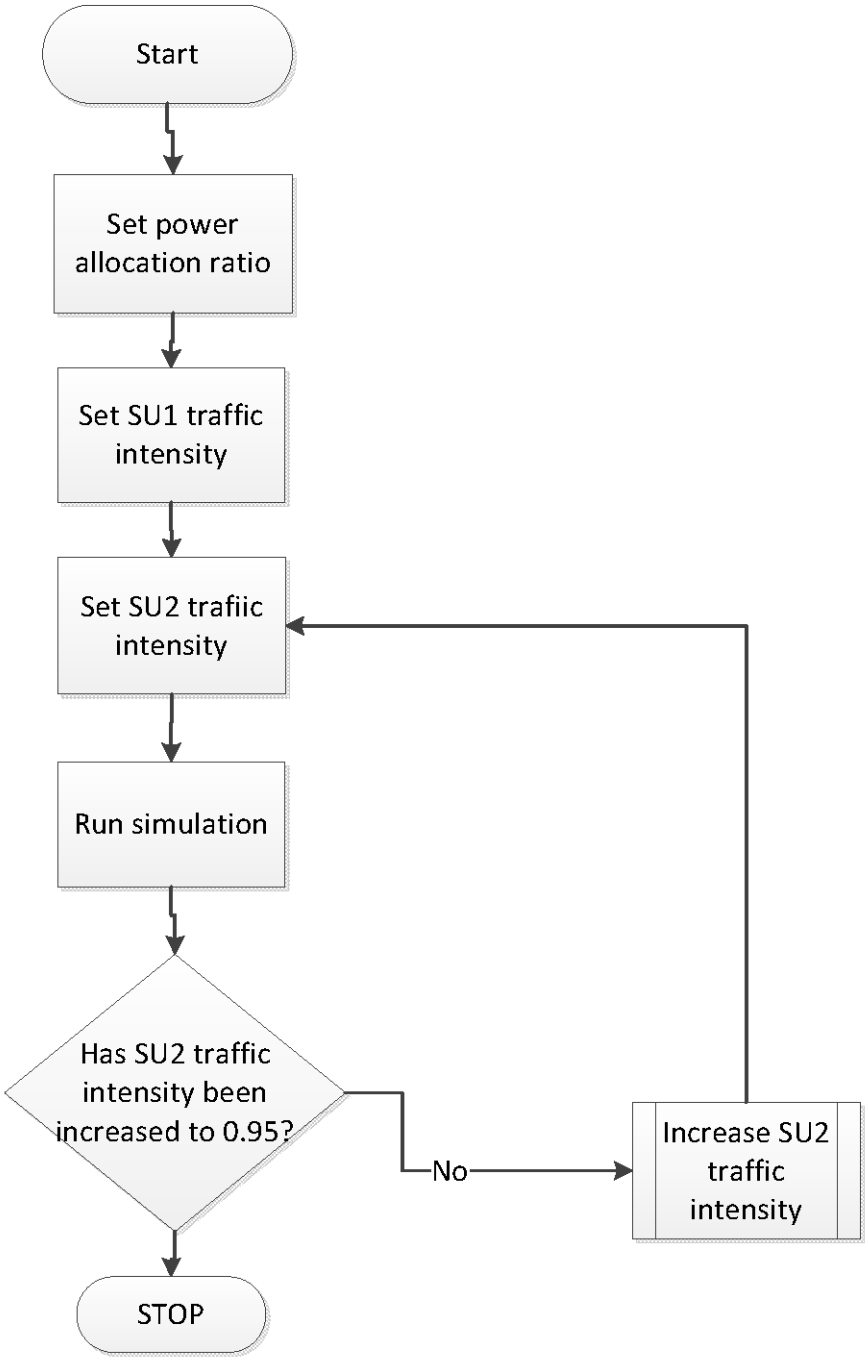

Figure 4.1. Flowchart for the system simulation

4.3 NETWORK CONDITIONS

The system behaviour or efficiency largely depends on the network conditions. The transmission time (or service time) is the key variable that will affect the network. Transmission time is affected by a number of variables. These variables as well as the equation for transmission time are explained in Chapter 3.

For the simulation, certain values and variables will have to explicitly defined. An IEEE standard was used to define some parameters such as bandwidth and number of bits per channel. The interference limit, Q, the maximum device power, P_{max}, and channel coefficients, h, are selected according to typical real world values. Channel coefficients are determined via the path loss propagation model and will be an indication of the distance between the transmitter and the receiver in question, i.e. h_{si} is indicative of the distance from transmitter SU i to receiver SU i.

4.3.1 Chosen parameter and variable Values

The standard chosen is the IEEE 802.22 [91]. Based on this the following parameters were set:

- Bandwidth = 6 MHz
- Number of bits per packet = 8184
- $P_{max} = 25$ dB
- $Q = 15$ dB
- $h_{s1} = 0.7$
- $h_{s2} = 0.6$
- $h_{p1} = 0.8$
- $h_{p2} = 0.9$.

Since channel conditions can vary more readily than all the other parameters, Fig. 4.2 shows the variations in transmission time given various channel coefficients. The service time values are then obtained equation 3.5 with the ratio r_i given after the respective plots.

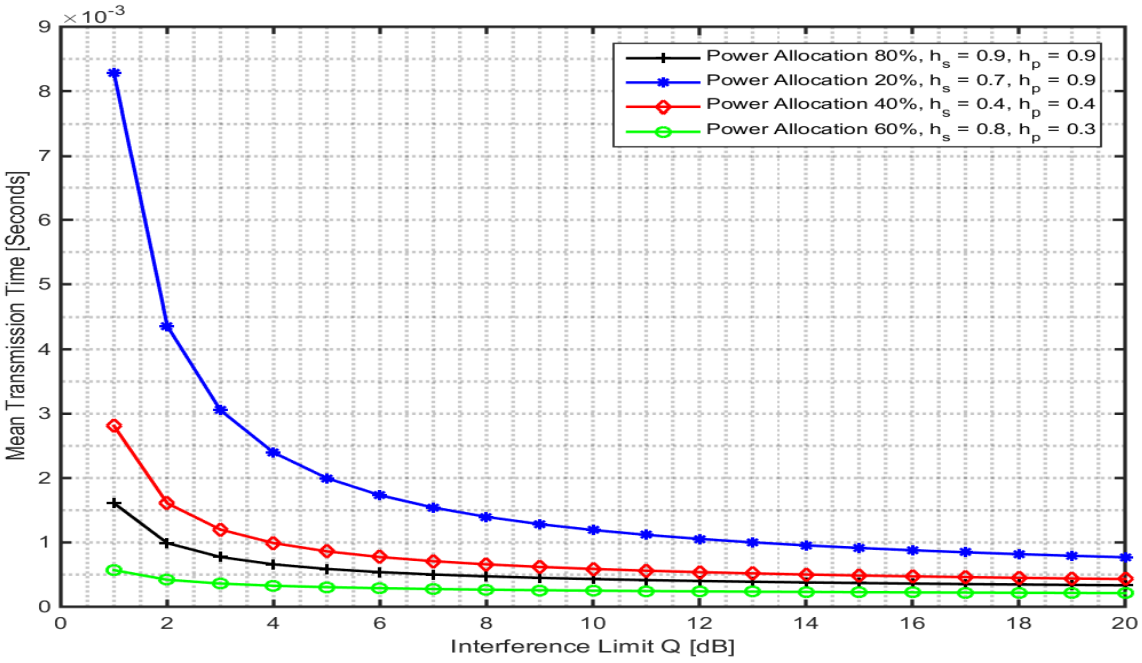

Figure 4.2. Mean transmission time vs. interference limit Q/N_o (SNR) under various scenarios.

Firstly, however, the simulation must adhere to theoretical values under the same conditions. That means that if Queue 2 arrival rate, λ_2, is set to zero, the system is effectively turned into an ordinary M/M/1 queue. The results must then match the well-documented theoretical results. Fig. 4.3, showing the simulation results, proves that the average length simulation results do indeed match the expected values from theory. Similarly, Fig. 4.4 shows the delay simulation.

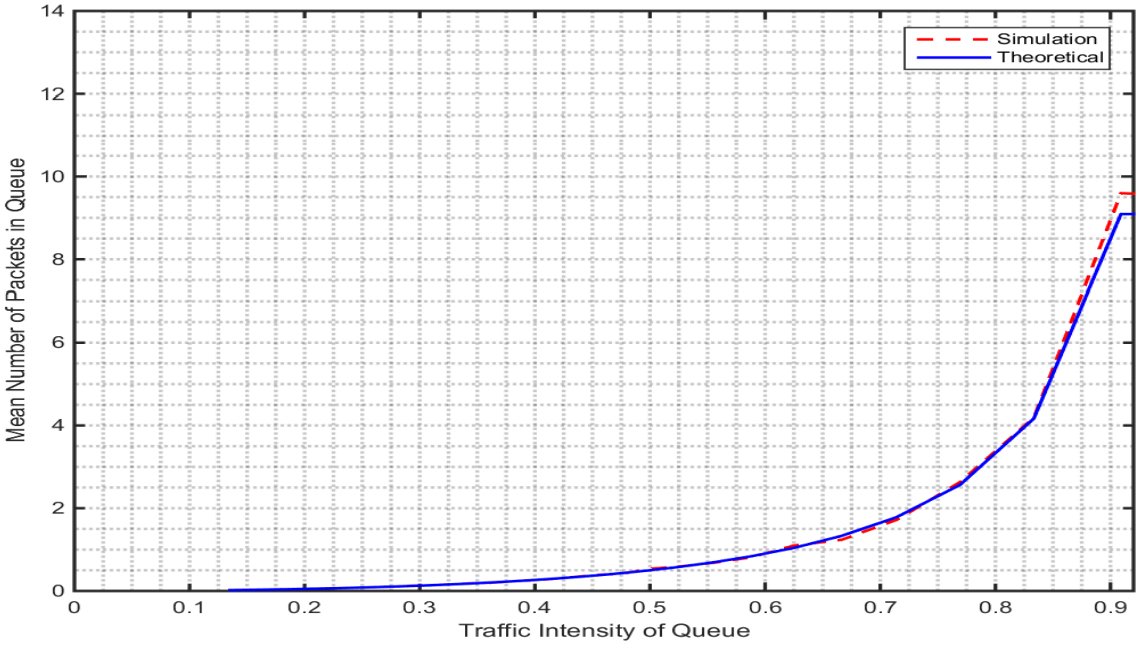

Figure 4.3. Simulated vs. theoretical mean number of packets for an M/M/1 queue

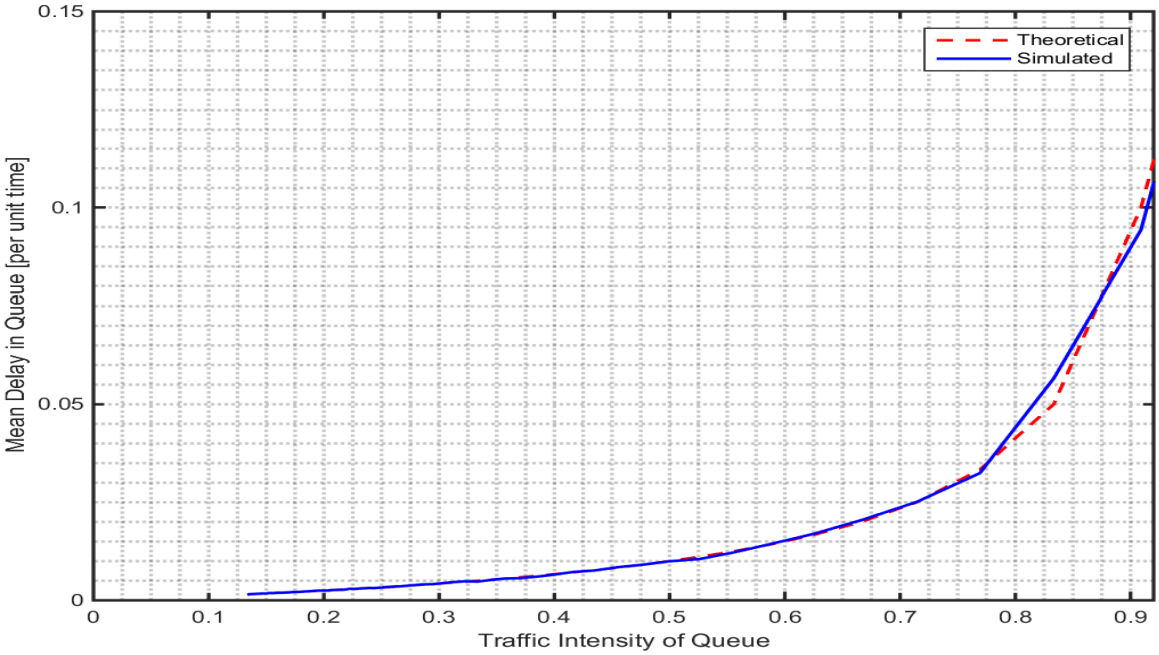

Figure 4.4. Simulated vs. theoretical mean delay for an M/M/1 queue

4.4 PRE-EMPTIVE QUEUE MODEL

After setting the parameters to the initial values given above, the queue length for the system can be determined. This is done according to traffic intensity, hence the arrival rates λ_1 and λ_2, will be varied. Using equation 2.3, it is possible to find the theoretical value of an ordinary M/M/1 if the traffic intensity is known. These values are given in the figures to show some comparison of the performance difference. The following is the algorithm used to generate the results:

Require: r_i $i \in \{1,2\}$

1: Set $h_{p1}, h_{p2}, h_{s1}, h_{s2}, P_{max}, Q, B, N, totalPackets, buffSize$
2: Use the values set is step one and r_i to calculate the service times for Queue one and two. μ_2 and μ_4 will be determined using the ratio r_i. μ_1 and μ_3 will be determined using a power allocation equal to unity for both. Equation 3.5 is used here.
3: Generate arrival times using a random exponential distribution and a mean inter-arrival time such that the traffic intensity of Queue one is 0.2 with μ_2 service rate.
4: **while** Queue traffic intensity is less than one **do**
5: Generate arrival times using a random exponential distribution and a mean inter-arrival time such that the traffic intensity of Queue two is 0.05 with μ_4 service rate.
6: **while** Number of packets in system is less than total number of packets **do**
7: Using the simulation time, arrival times and service times determine which event occurs first and adjust the queue length accordingly. *% There are four types of events in the simulations, namely Queue one arrival, Queue one departure, Queue two arrival and Queue two departure.*
8: Any packet, from any queue, that arrives first will enter into service immediately and depending on the queue, will be allocated either μ_1 or μ_2 service rate.
9: **if** A Queue one packet arrives while a Queue one packet is in service **then**
10: Increase Queue one size by one and generate next Queue one arrival time.
11: **end if**
12: **if** A Queue two packet arrives while a Queue two packet is in service **then**
13: **if** Queue two size is less than buffer size **then**
14: Increase Queue two size by one.
15: **end if**
16: Generate next Queue two arrival time.
17: **end if**
18: **if** A Queue one/two packet arrives while a Queue two/one packet is in service **then**
19: Adjust service rates to μ_3 or μ_4 immediately to accommodate arriving packet.
20: **end if**
21: **if** A Queue one/two packet arrives while both Queue one and two packets are in service **then**
22: Increase Queue one size by one.

23: **if** Queue two size is less than buffer size **then**
24: Increase Queue two size by one.
25: **end if**
26: Generate next Queue one and two arrival times.
27: **end if**
28: Use time spent in the queue and the simulation time to determine the average number of packets in the queue and the average delays in the queues.
29: Increase mean inter-arrival time such that traffic intensity of Queue two increases by 0.05.
30: **end while**
31: Increase Queue one traffic intensity by 0.2.
32: **end while**
33: Plot simulation graph

4.4.1 Queue length

Figs. 4.5 to 4.10 show the results of the mean number of packets in the queue simulation. As expected, the average number in the queue increases with traffic intensity. The mean service rates can be seen to change with change in the power allocation. This is to show the performance of the system under different power allocation or channel conditions. Overall the simulations were carried out over four different traffic intensities for the incumbent SU in order to simulate for a high range of scenarios.

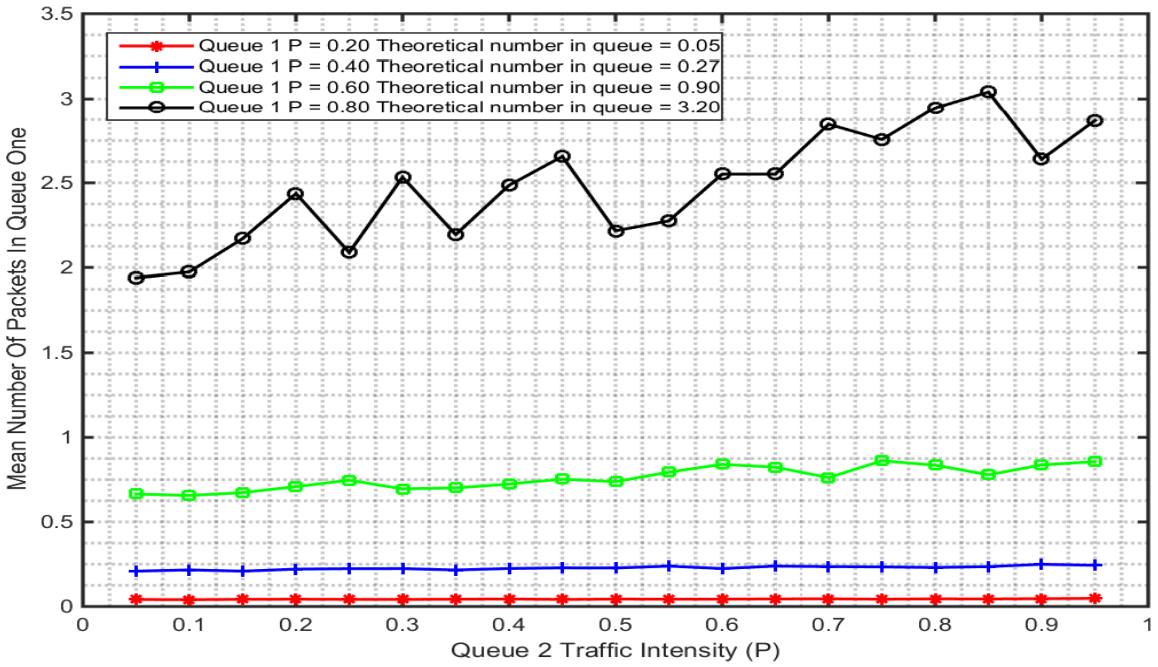

Figure 4.5. Queue 1 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

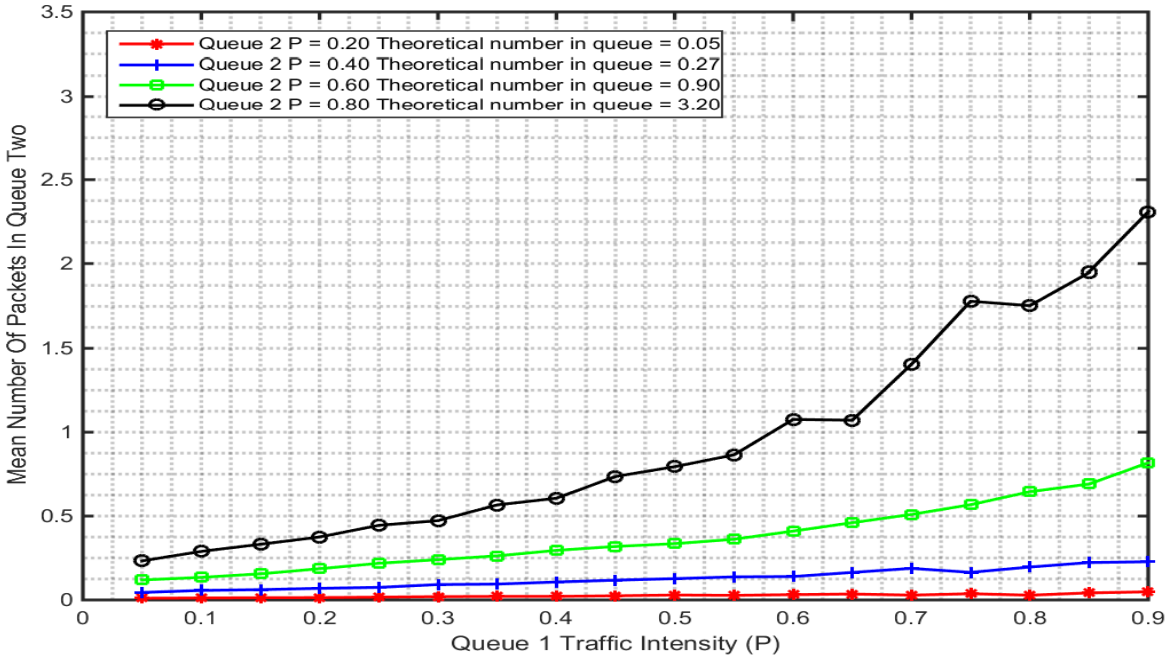

Figure 4.6. Queue 2 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

In Figs. 4.5 and 4.6 the power allocation ratio is 80% Queue 1 to 20% Queue 2. Hence μ_2 is much greater than μ_4. It takes Queue 2 more time to transmit a packet when both SUs are transmitting.

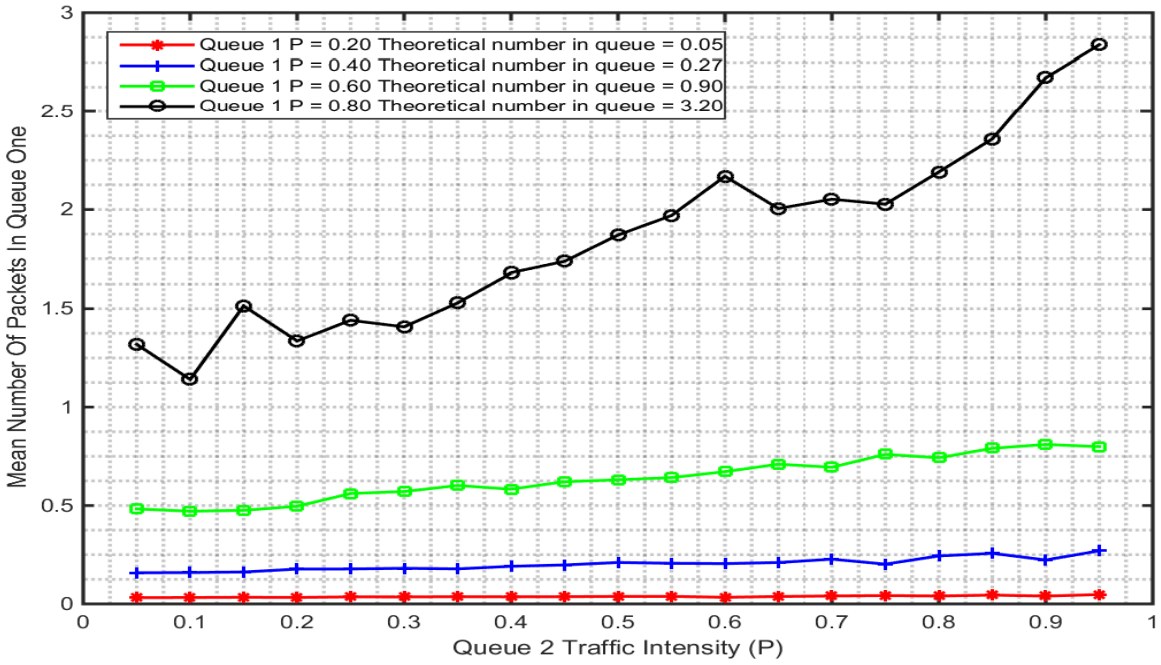

Figure 4.7. Queue 1 average number of packets. Here, $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00051$ sec/packet and $\mu_4 = 0.00064$ sec/packet

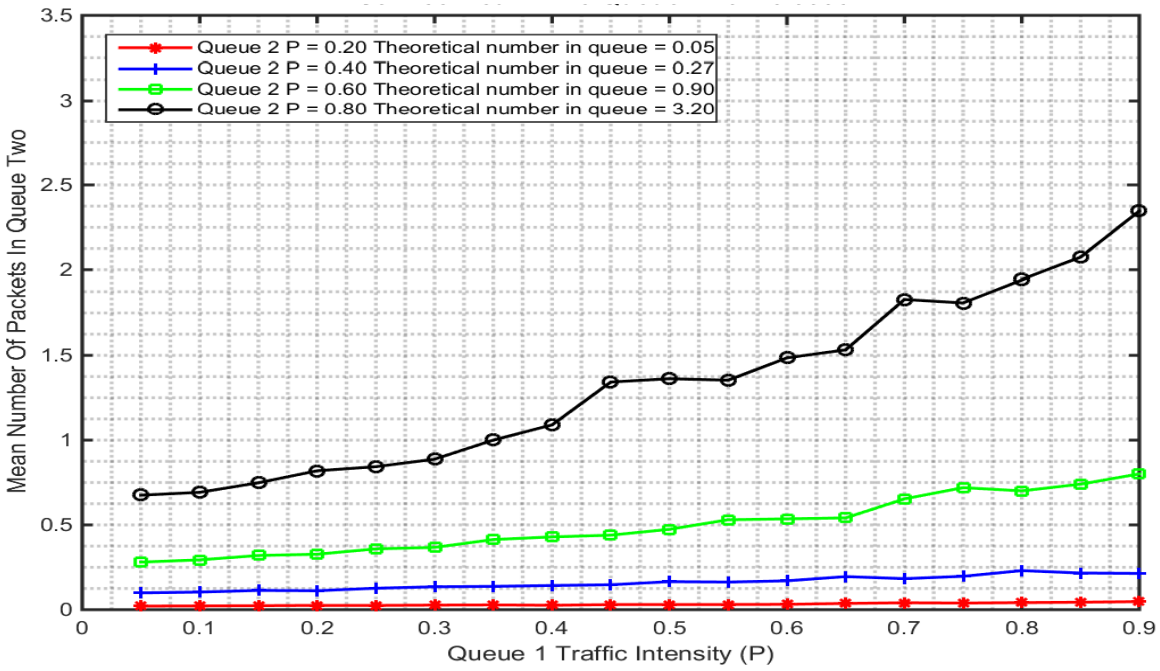

Figure 4.8. Queue 2 average number of packets. Here, $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00051$ sec/packet and $\mu_4 = 0.00064$ sec/packet

In Figs. 4.7 and 4.8 the power allocation ratio is 60% Queue 1 to 40% Queue 2. Queue 1 takes a bit longer to transmit a packet under shared conditions.

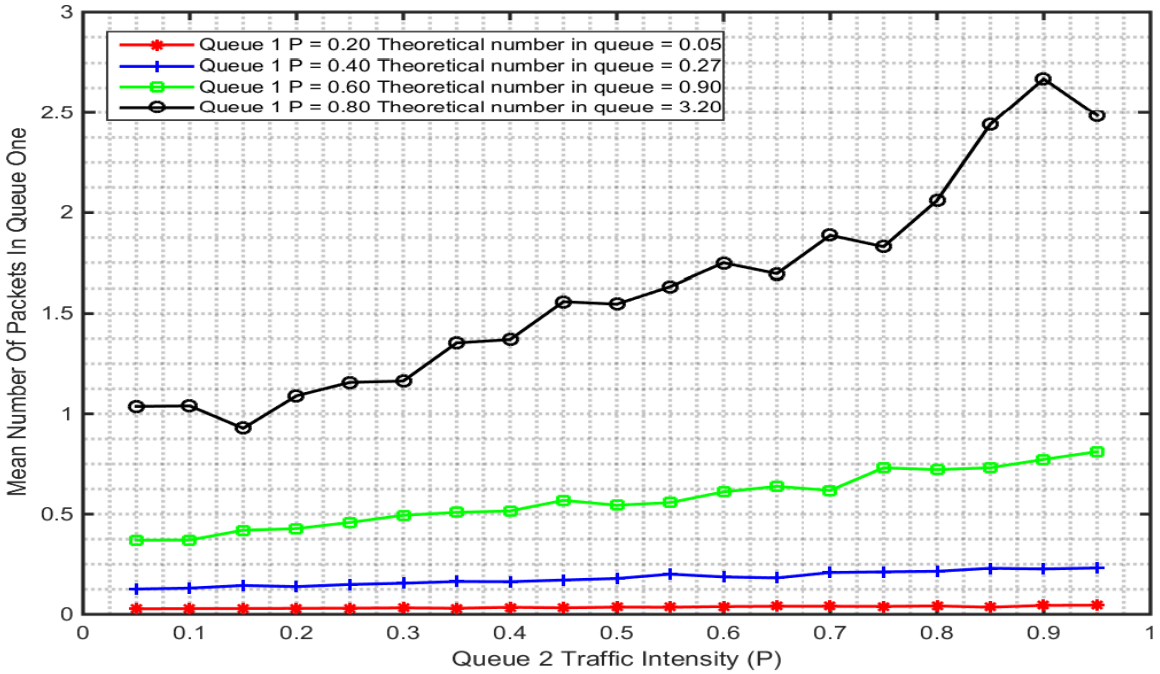

Figure 4.9. Queue 1 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00055$ sec/packet and $\mu_4 = 0.00057$ sec/packet

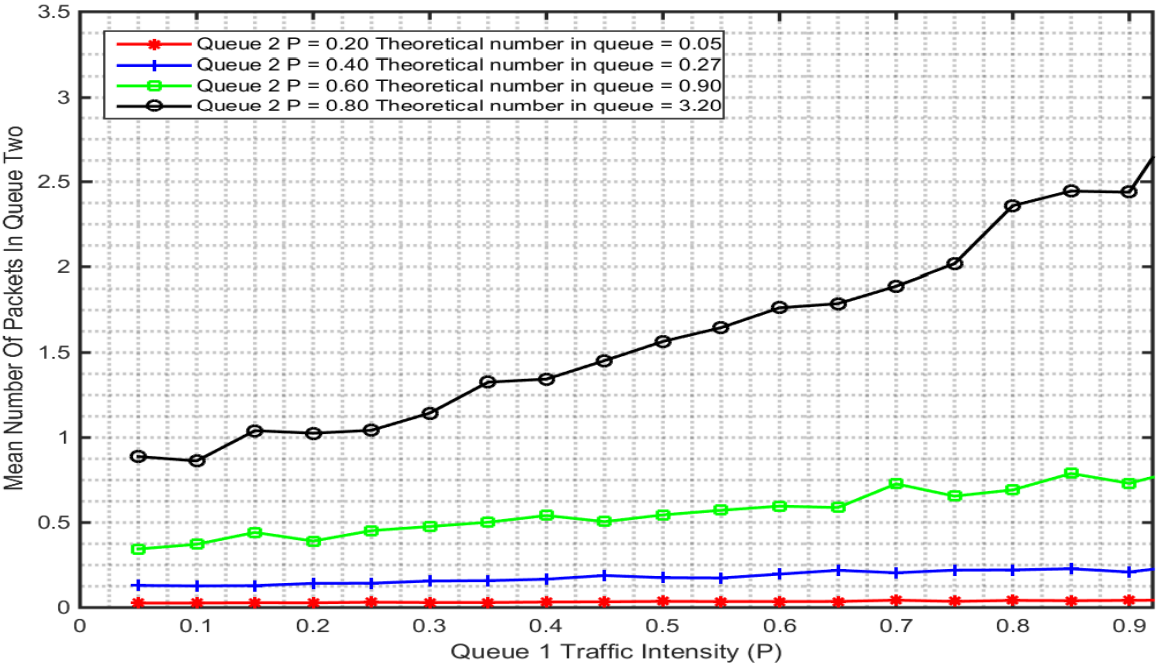

Figure 4.10. Queue 2 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00055$ sec/packet and $\mu_4 = 0.00057$ sec/packet

In Figs. 4.9 and 4.10 the power allocation ratio is 50% Queue 1 to 50% Queue 2. μ_2 and μ_4 are almost the same, only differing because of the different channel conditions.

4.4.2 Waiting time

Figs. 4.11 to 4.16 show the expected delay in the queues against the traffic intensity of the other queue under various traffic intensities. Again the delay increases with traffic intensity because of the queues being served at the reduced rate. Should the service rate of Queue 2 be high enough, then Queue 1 will be negatively affected owing to the higher waiting times before commencing service.

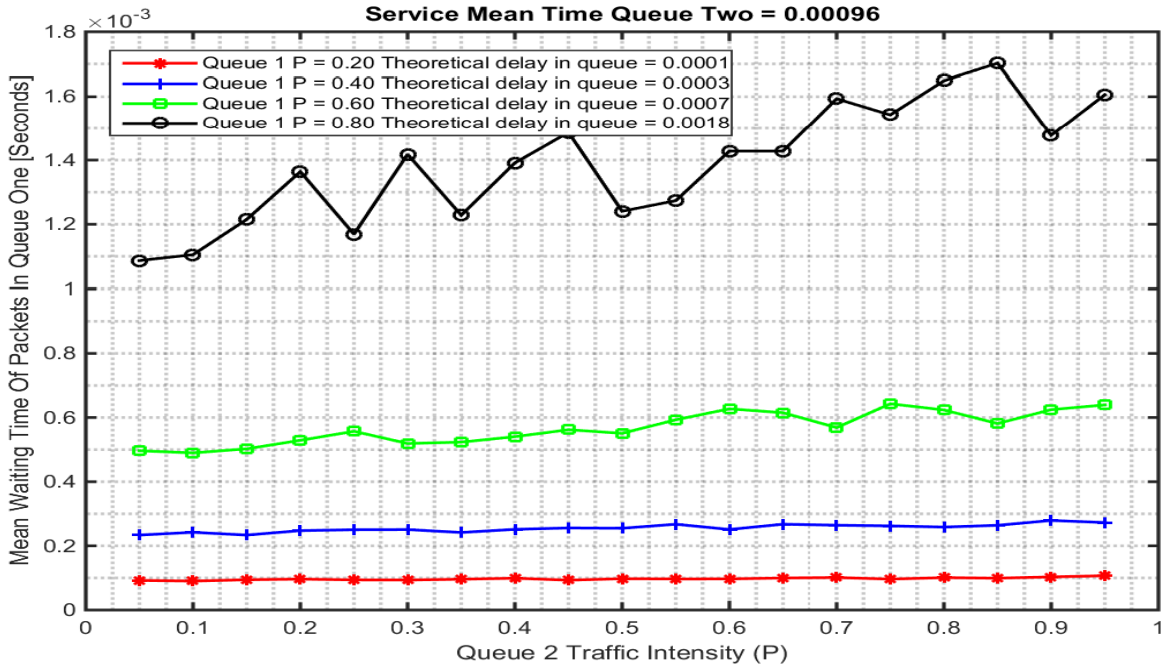

Figure 4.11. Queue 1 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

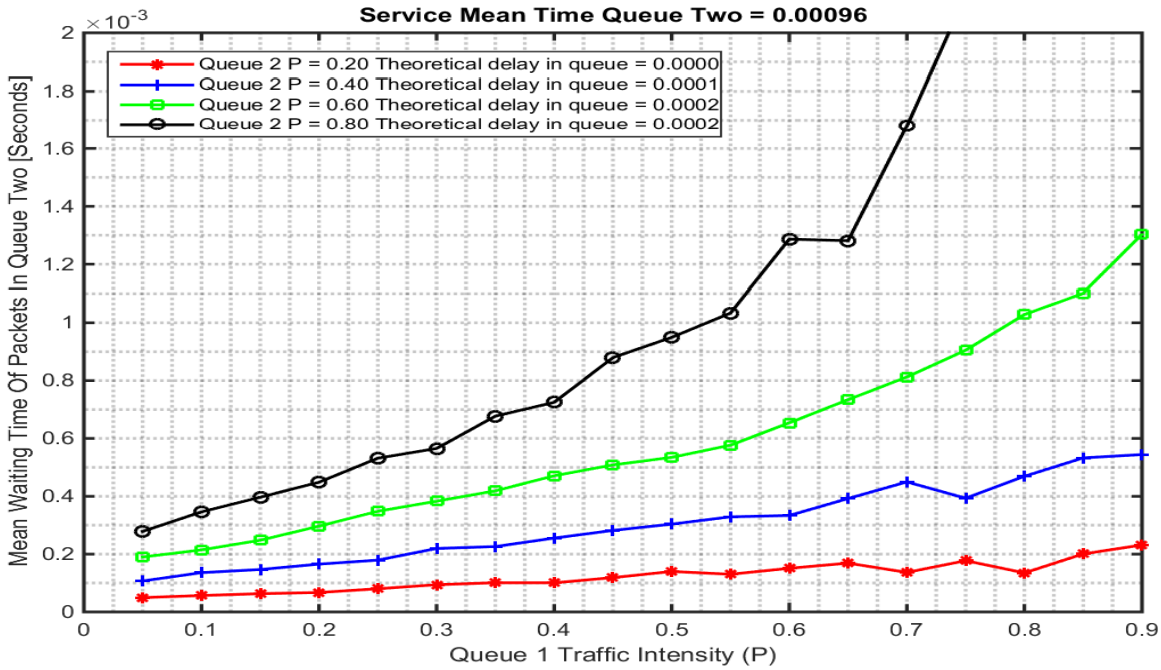

Figure 4.12. Queue 2 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

In Figs. 4.11 and 4.12 the power allocation ratio is 80% HW to 20% LW. Hence μ_2 is much greater than μ_4. It takes the LW more time to transmit a packet when both SUs are transmitting. Hence there will be more delay. Figs 4.5 and 4.11 are similar. This is because they are proportional, as confirmed by Little's formula in equation 2.7.

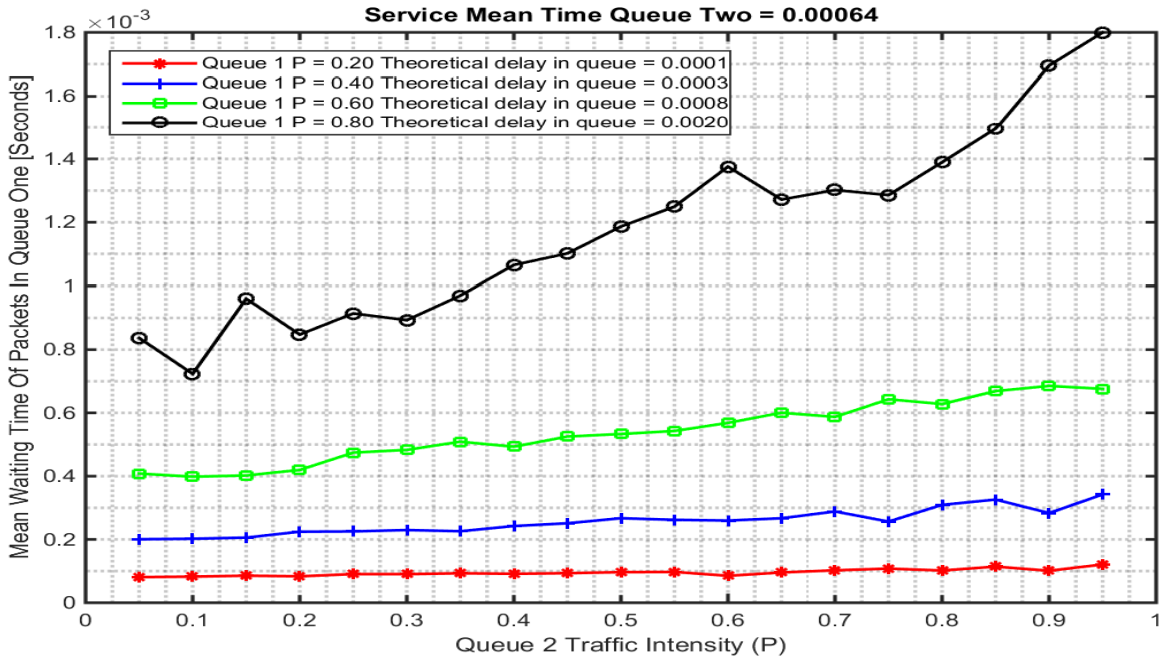

Figure 4.13. Queue 1 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00051$ sec/packet and $\mu_4 = 0.00064$ sec/packet

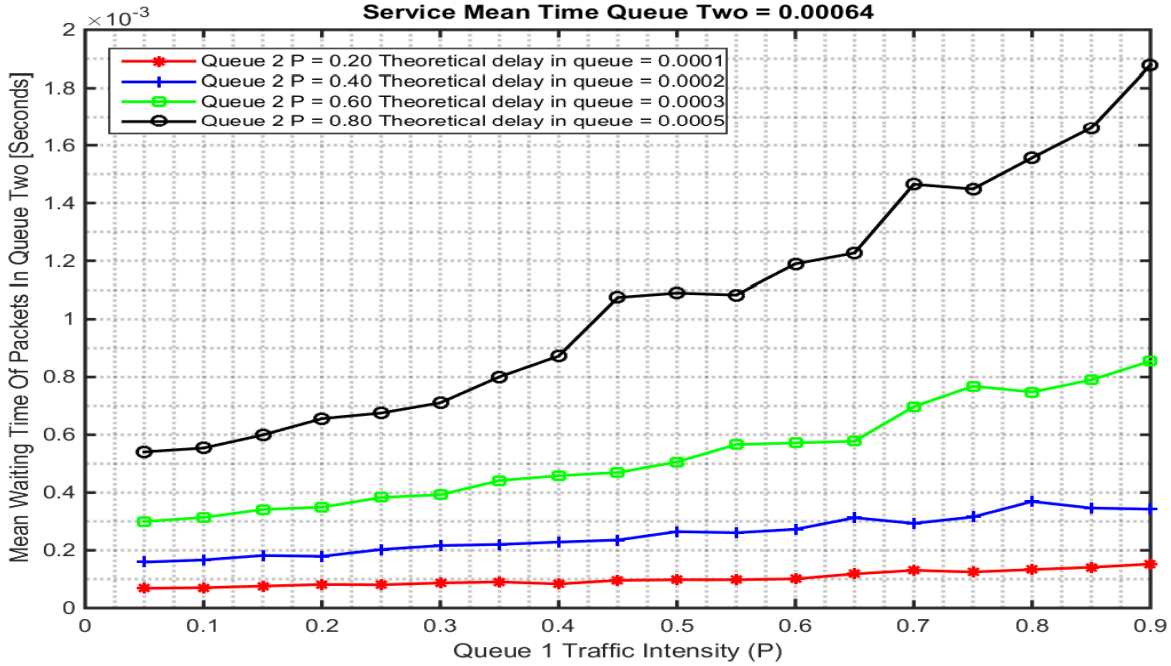

Figure 4.14. Queue 2 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00051$ sec/packet and $\mu_4 = 0.00064$ sec/packet

In Figs. 4.13 and 4.14 the power allocation ration is 60% Queue 1 to 40% Queue 2.

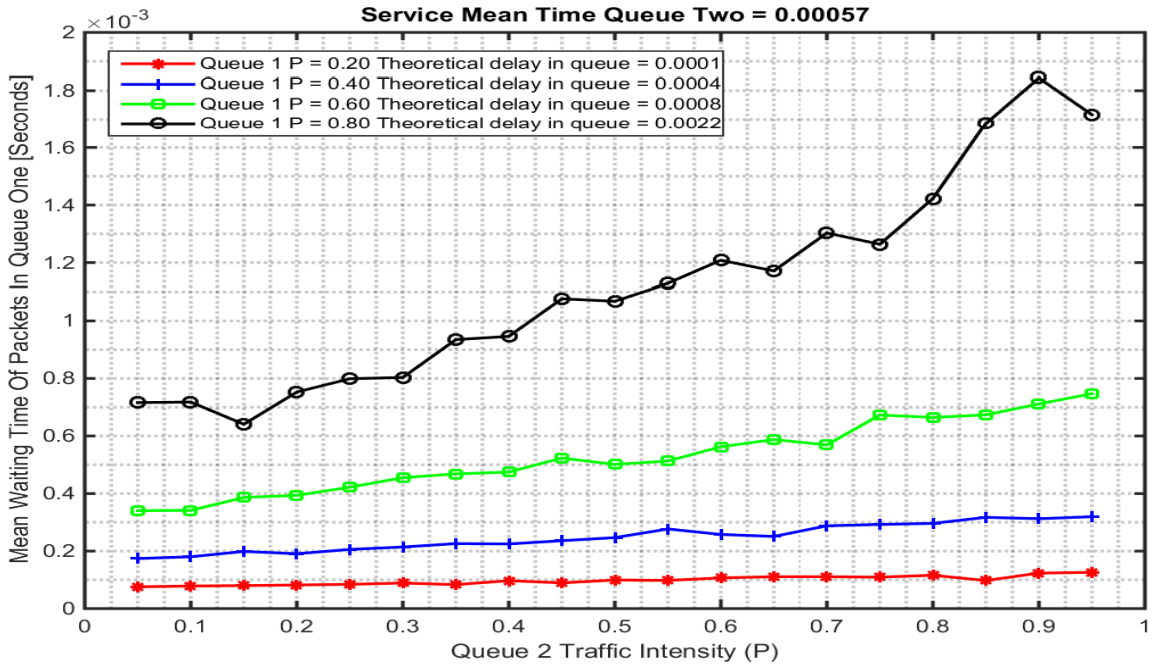

Figure 4.15. Queue 1 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00055$ sec/packet and $\mu_4 = 0.00057$ sec/packet

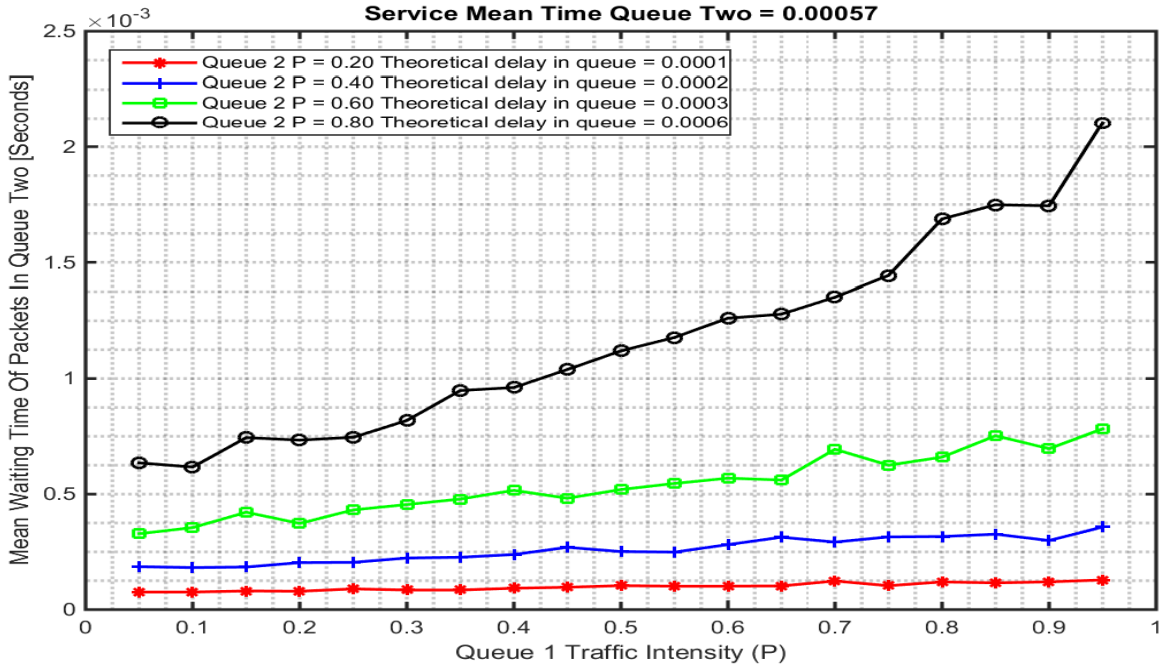

Figure 4.16. Queue 2 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00055$ sec/packet and $\mu_4 = 0.00057$ sec/packet

In Figs. 4.15 and 4.16 the power allocation ratio is 50% Queue 1 to 50% Queue 2.

4.4.3 Buffer considerations

The effect of the buffer size for the pre-emptive queue model is shown in Fig. 4.17. The buffer is set to K = 2. The mean length and waiting times decrease with a decrease in buffer length.

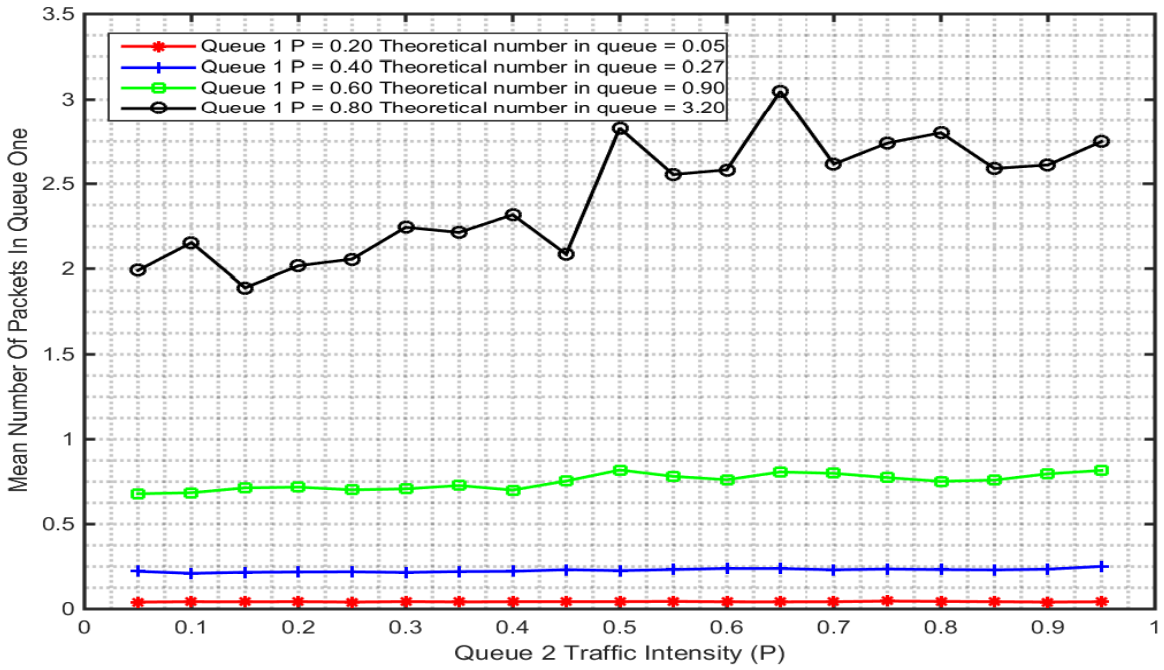

Figure 4.17. Queue 1 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

Fig. 4.18 shows the mean queue length of the LW queue under the same conditions as in Fig. 4.17.

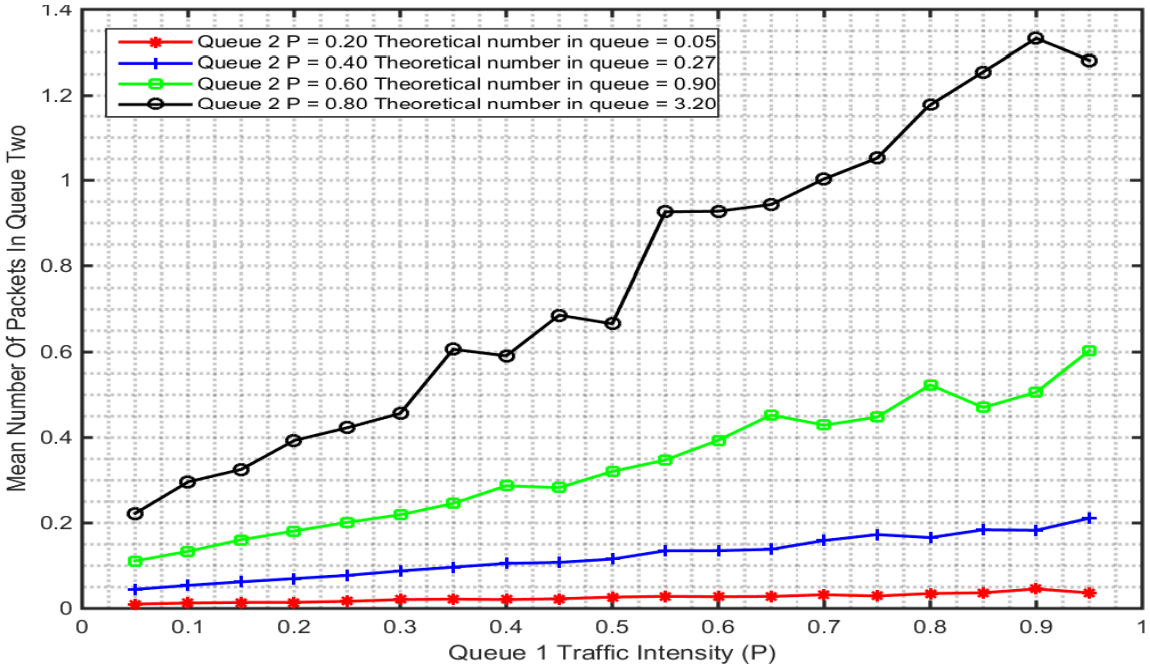

Figure 4.18. Queue 2 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

4.5 NON-PRE-EMPTIVE QUEUE MODEL

Figs. 4.19 to 4.24 show the expected number of packets in the queues against the traffic intensity of the other queue under various traffic intensities. The mean number in the queue increases with the traffic intensity because more Queue 1 packets will result in more Queue 1 packets having to wait in the queue because of the non-pre-emption rule. Initially the buffer is set to infinite. The algorithm used to generate the results is given as follows:

Require: r_i $i \in \{1,2\}$

1: Set $h_{p1}, h_{p2}, h_{s1}, h_{s2}, P_{max}, Q, B, N, totalPackets$

2: Use the values set is step one and r_i to calculate the service times for Queue one and two. μ_2 and μ_4 will be determined using the ratio r_i. μ_1 and μ_3 will be determined using a power allocation equal to unity for both. Equation 3.5 is used here.

3: Generate arrival times using a random exponential distribution and a mean inter-arrival time such that the traffic intensity of Queue one is 0.2 for μ_2 service rate.

4: **while** Queue traffic intensity is less than one **do**

5: Generate arrival times using a random exponential distribution and a mean inter-arrival time such that the traffic intensity of Queue two is 0.05 for μ_4 service rate.

6: **while** Number of packets in system is less than total number of packets **do**

7: Using the simulation time, arrival times and service times determine which event occurs first and adjust the queue length accordingly. *% There are four types of events in the simulations, namely Queue one arrival, Queue one departure, Queue two arrival and Queue two departure.*

8: The first packet, from any queue, that arrives first will enter into service immediately and depending on the queue, will be allocated either μ_1 or μ_2 service rate.

9: **if** A Queue one packet arrives while a Queue one packet is in service **then**

10: Increase Queue one size by one and generate next Queue one arrival time.

11: **end if**

12: **if** A Queue two packet arrives while a Queue two packet is in service **then**

13: **if** Queue two size is less than buffer size **then**

14: Increase Queue two size by one.

15: **end if**

16: Generate next Queue two arrival time.

17: **end if**

18: **if** A Queue one/two packet arrives while a Queue two/one packet is in service **then**

19: **if** Queue one **then**

20: Increase Queue one size by one.

21: **else if** Queue two size is less than buffer size **then**

22: Increase Queue two size.

23: **end if**

24: Set service rates of next packets in Queue one and two to μ_3 and μ_4 respectively.

25: **end if**

26: **if** A Queue one/two packet arrives during a busy period and there packets in Queue two/one **then**

27: **if** Queue one **then**

28: Increase Queue one size by one.

29: **else if** Queue two size is less than buffer size **then**

30: Increase Queue two size

31: **end if**

32: **end if**

33: **if** A Queue one/two packet arrives while both Queue one and two packets are in service **then**

34: Increase Queue one size by one.

35: **if** Queue two size is less than buffer size **then**

36: Increase Queue two size by one.

37: **end if**

38: Generate next Queue one and two arrival times.

39: **end if**

40: Use time spent in the queue and the simulation time to determine the average number of packets in the queue and the average delays in the queues.

41: Increase mean inter-arrival time such that traffic intensity of Queue two increases by 0.05.

42: **end while**

43: Increase Queue one traffic intensity by 0.2.

44: **end while**

45: Plot simulation graph

4.5.1 Queue length

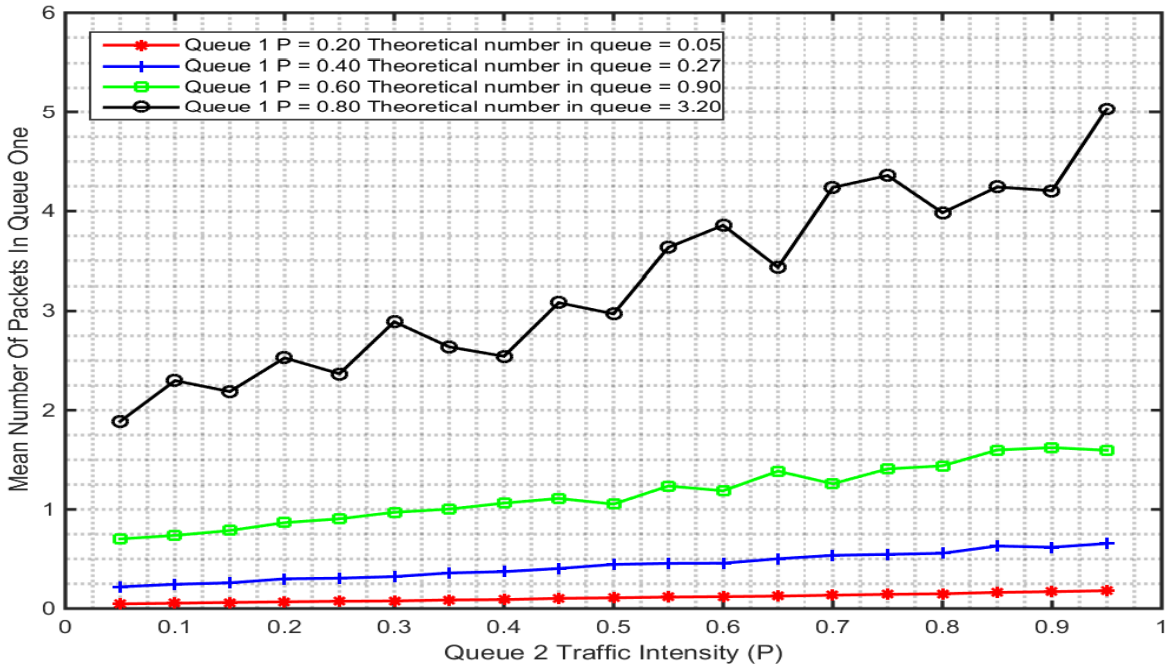

Figure 4.19. Queue 1 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

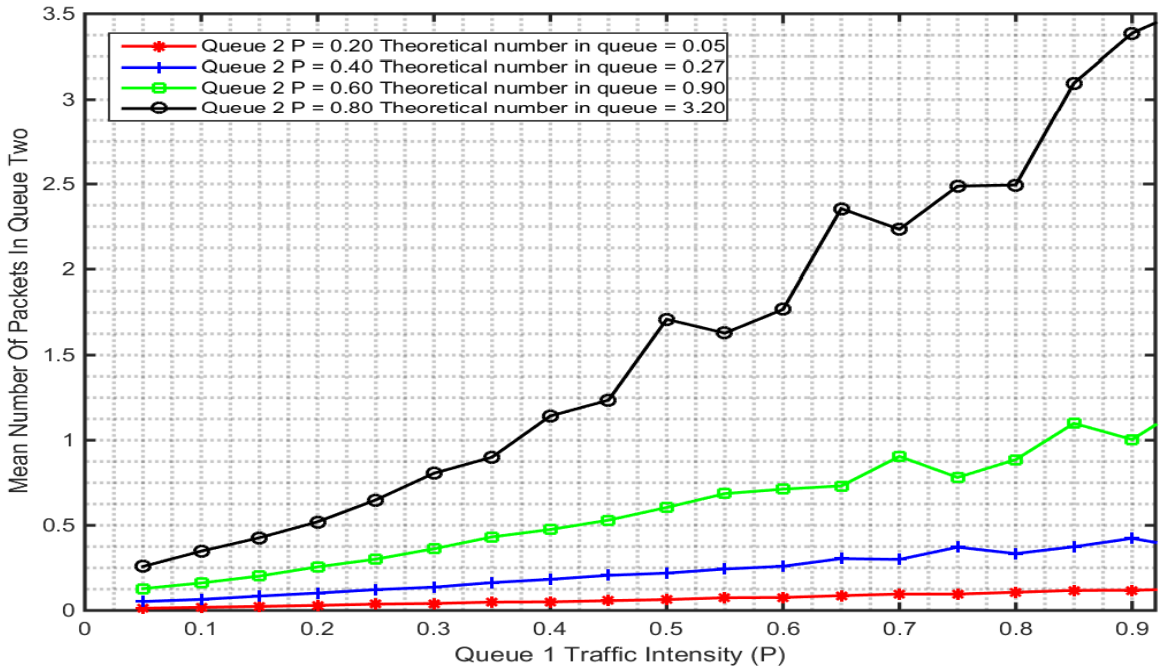

Figure 4.20. Queue 2 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

In Fig. 4.19 and 4.20 the power allocation ratio is 80% Queue 1 to 20% Queue 2. Hence μ_2 is much greater than μ_4. It takes Queue 2 more time to transmit a packet when both SUs are transmitting.

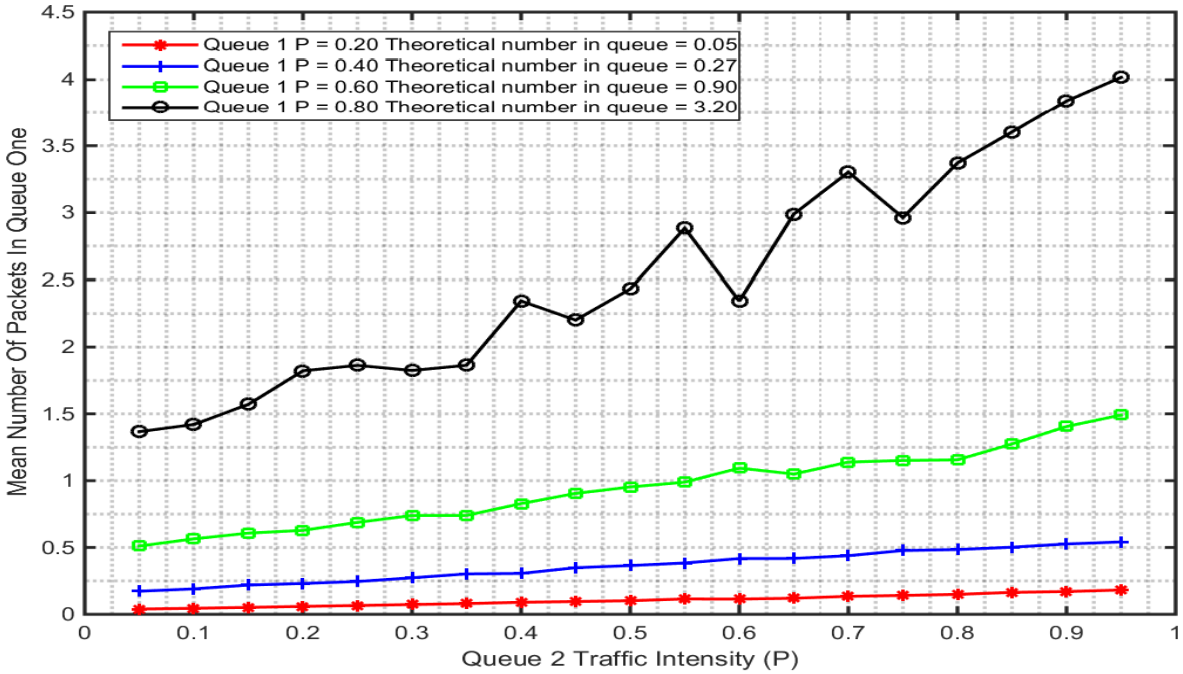

Figure 4.21. Queue 1 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00051$ sec/packet and $\mu_4 = 0.00064$ sec/packet

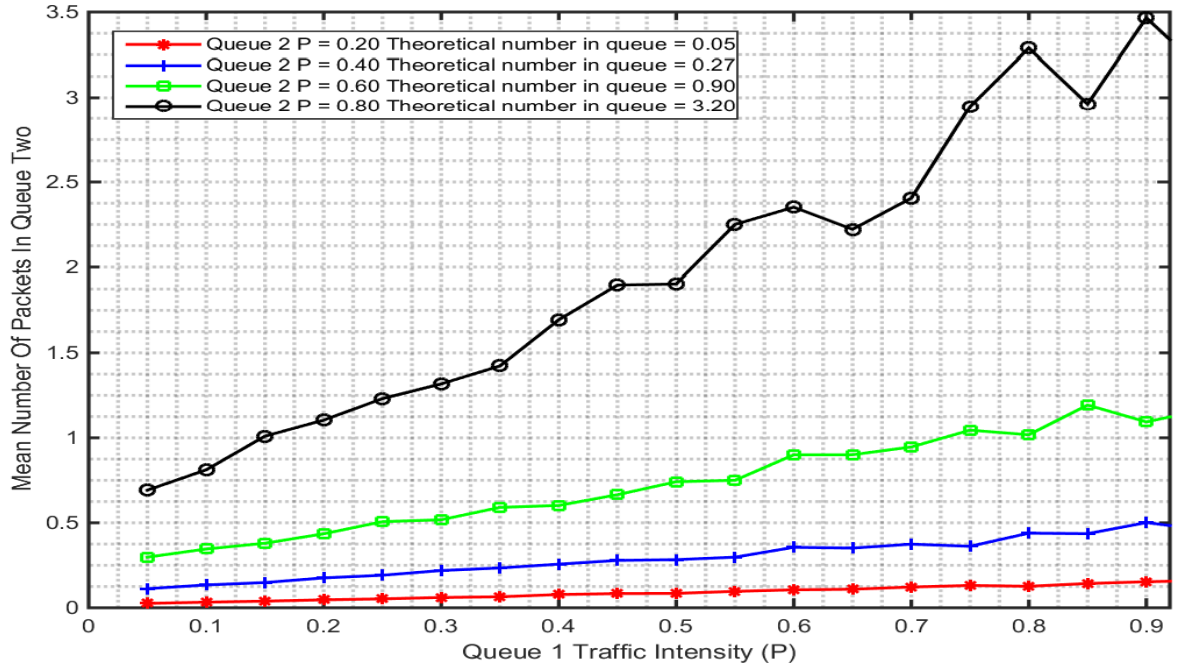

Figure 4.22. Queue 2 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00051$ sec/packet and $\mu_4 = 0.00064$ sec/packet

In Figs. 4.21 and 4.22 the power allocation ratio is 60% Queue 1 to 40% Queue 2. Queue 1 now takes a bit longer to transmit a packet under shared conditions.

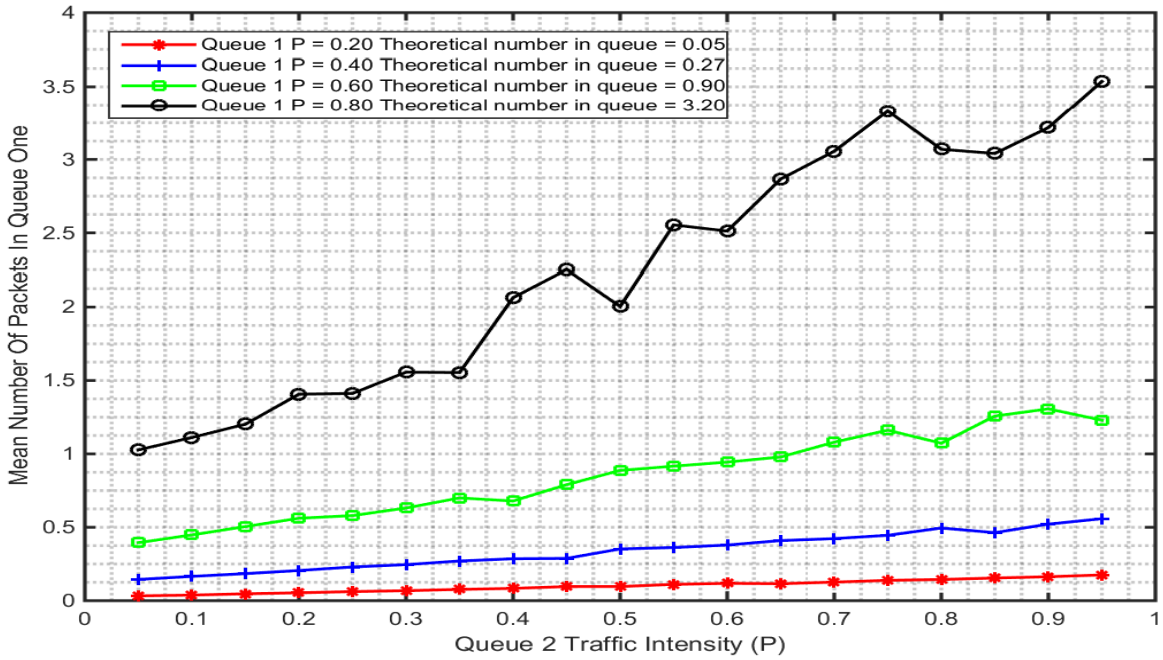

Figure 4.23. Queue 1 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00055$ sec/packet and $\mu_4 = 0.00057$ sec/packet

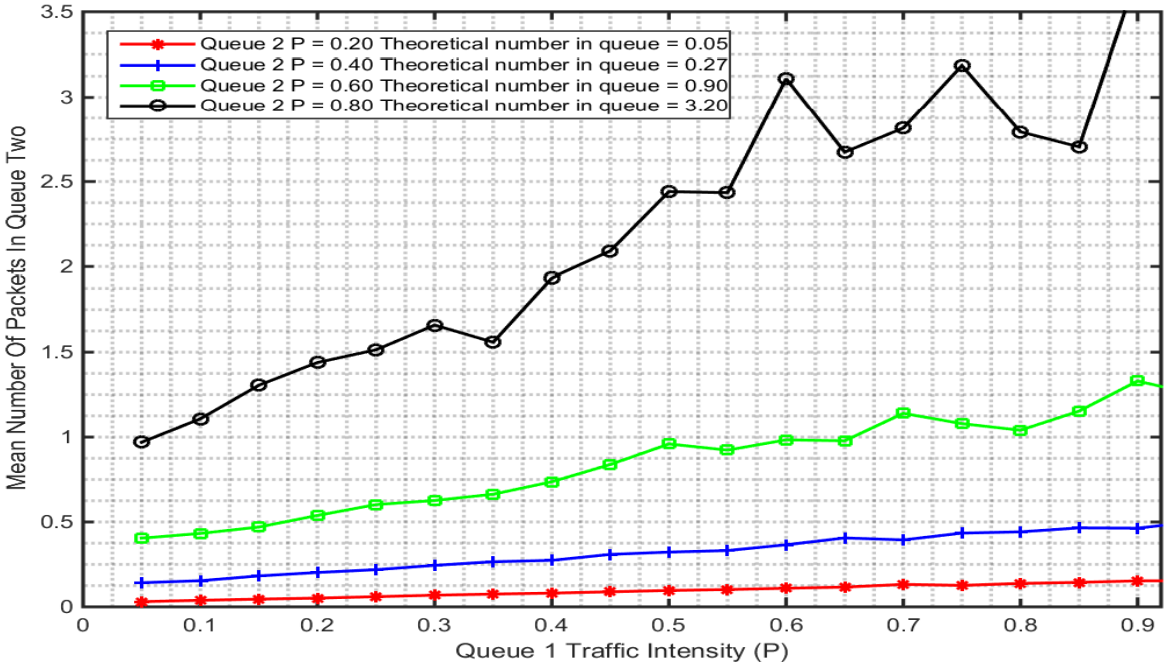

Figure 4.24. Queue 2 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00055$ sec/packet and $\mu_4 = 0.00057$ sec/packet

In Figs. 4.23 and 4.24 the power allocation ratio is 50% Queue 1 to 50% Queue 2. μ_2 and μ_4 are almost the same, only differing because of the different channel conditions.

4.5.2 Waiting time

Figs. 4.25 to 4.30 show the expected delay in Queue 1 against the traffic intensity of Queue 2 under various traffic intensities.

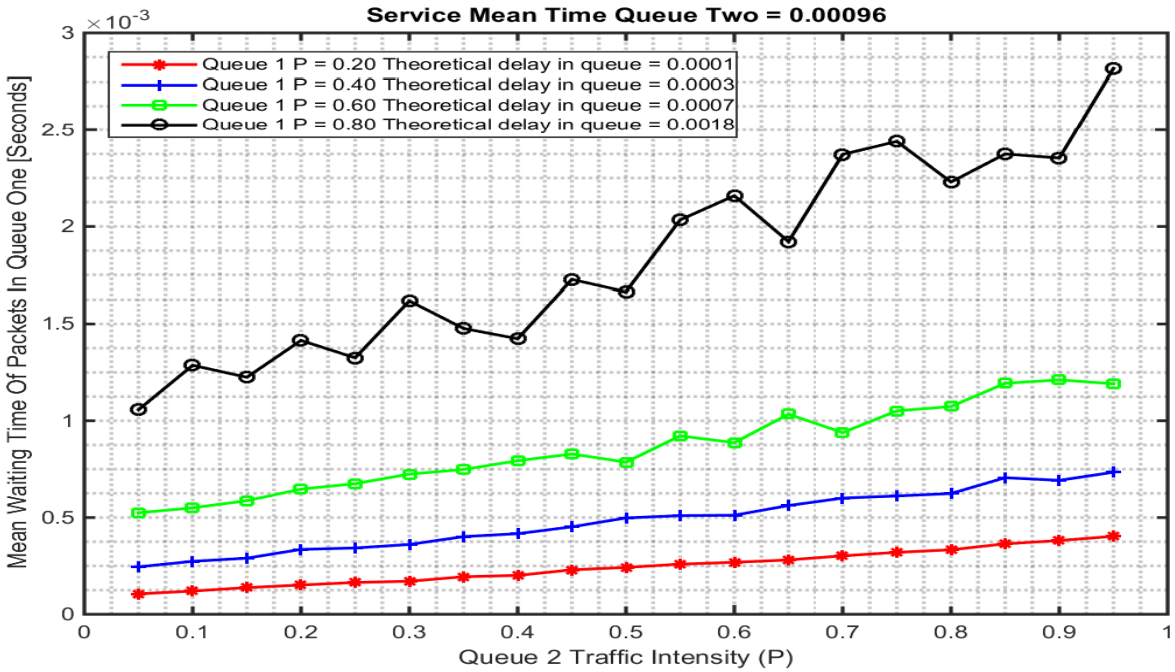

Figure 4.25. Queue 1 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

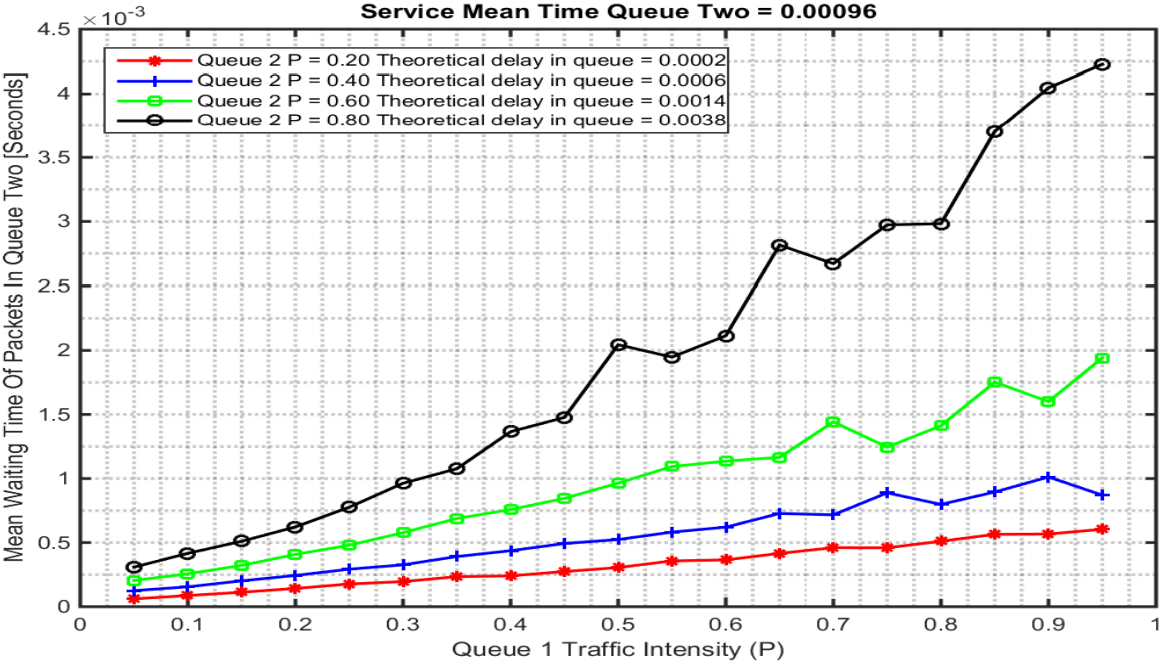

Figure 4.26. Queue 2 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

In Figs. 4.25 and 4.26 the power allocation ration is 80% HW to 20% LW. Hence μ_2 is much greater than μ_4. The LW queue takes more time to transmit a packet when both SUs are transmitting. A higher waiting time can be observed in the queues.

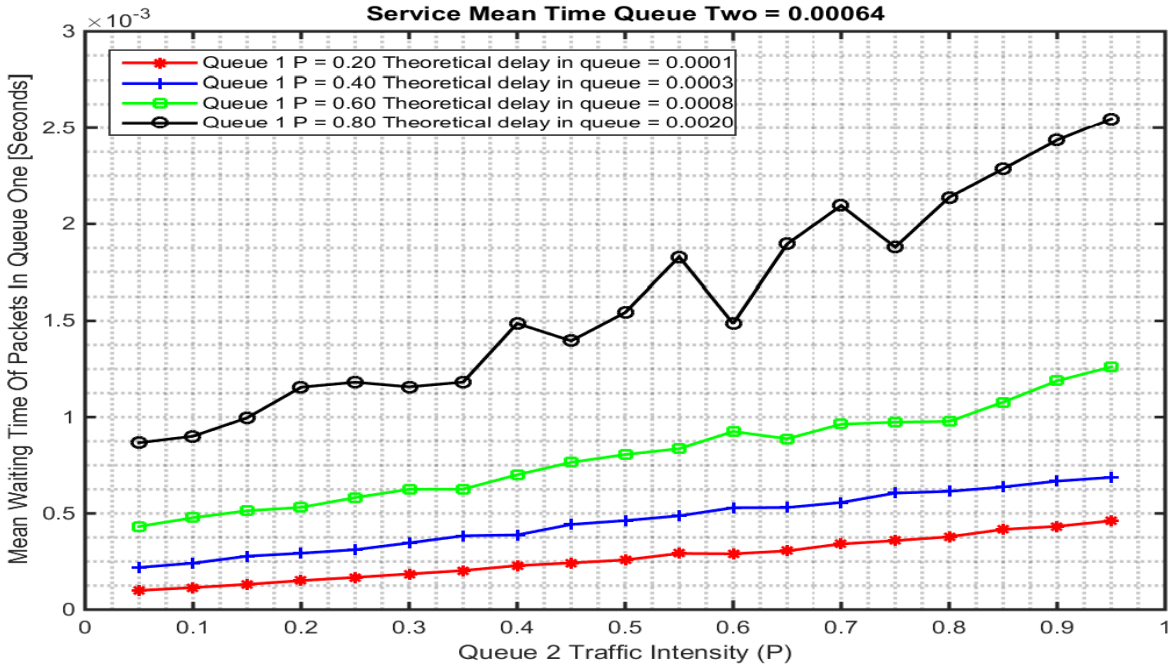

Figure 4.27. Queue 1 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00051$ sec/packet and $\mu_4 = 0.00064$ sec/packet

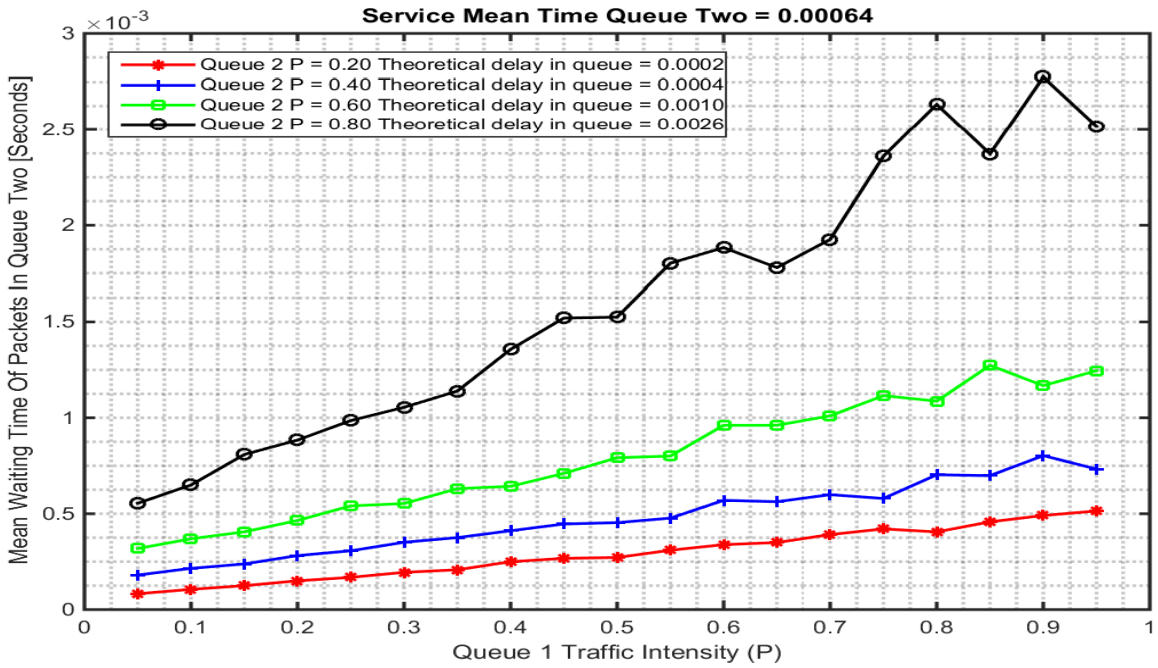

Figure 4.28. Queue 2 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00051$ sec/packet and $\mu_4 = 0.00064$ sec/packet

In Figs. 4.27 and 4.28 the power allocation ratio is 60% HW to 40% LW. Hence μ_2 is much greater than μ_4. The waiting times are longer than for the corresponding simulation for the pre-emptive model.

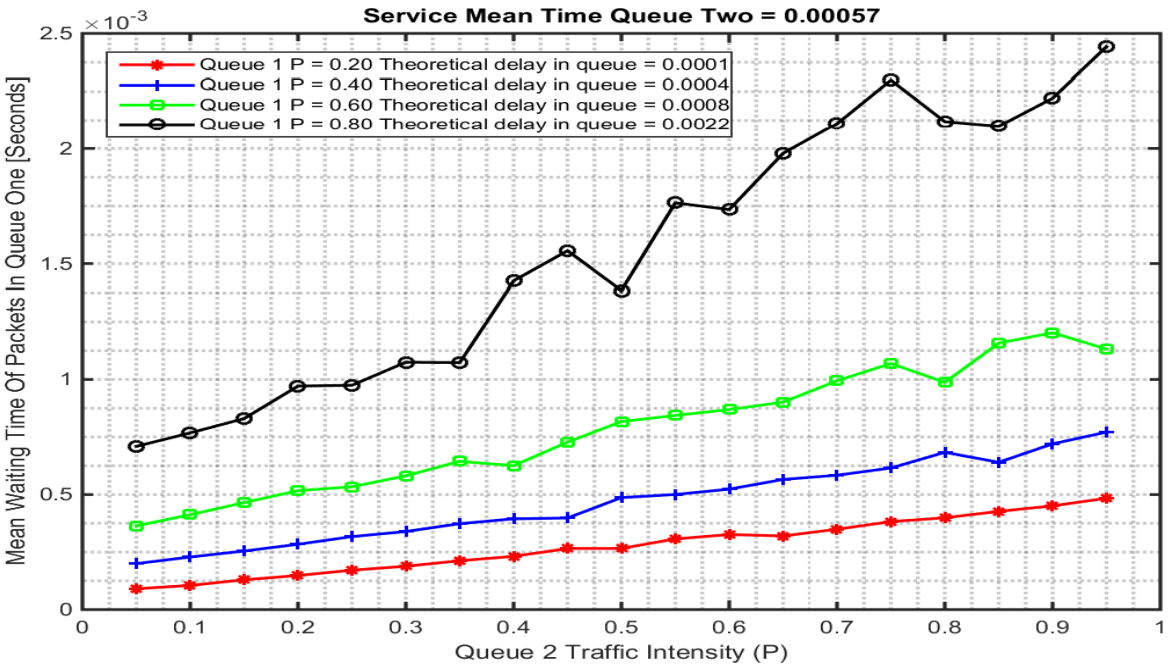

Figure 4.29. Queue 1 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00055$ sec/packet and $\mu_4 = 0.00057$ sec/packet

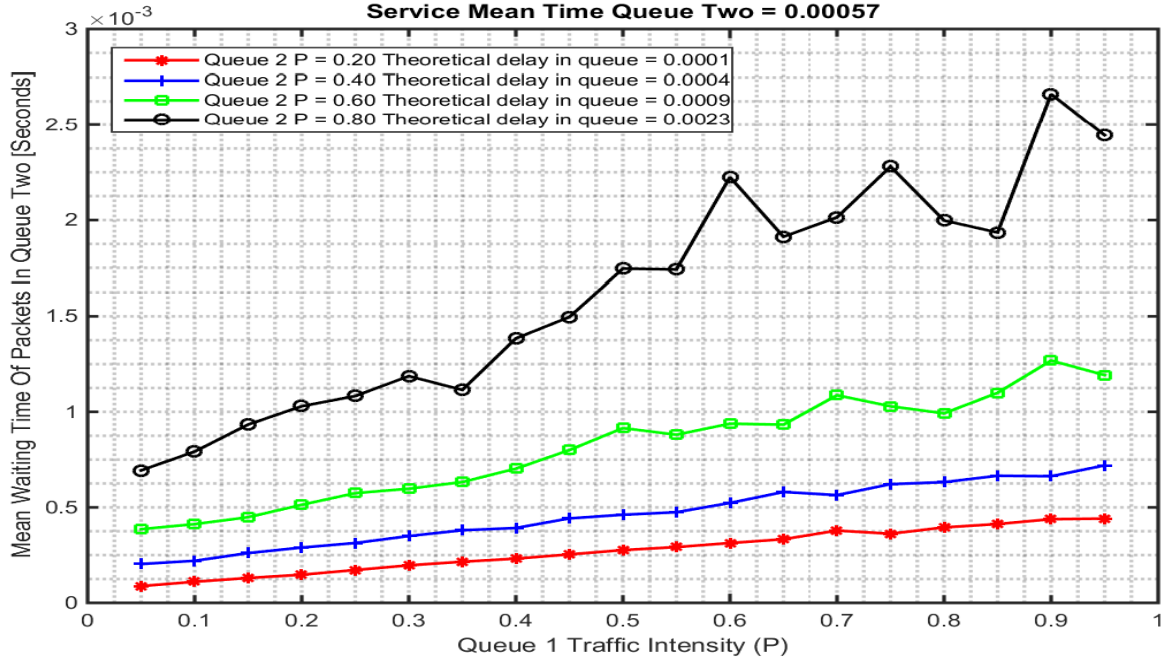

Figure 4.30. Queue 2 average waiting time of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00055$ sec/packet and $\mu_4 = 0.00057$ sec/packet

In Figs. 4.27 and 4.30 the power allocation ratio is 50% HW to 50% LW. Hence μ_2 is much greater than μ_4. It can be observed that the queue performance is not identical despite the equal power allocation. Channel conditions play a role here. A poor channel will result in decreased performance.

The non-pre-emption rule will result in the packets higher waiting times as the transmission time increases.

4.5.3 Buffer considerations

The addition of a buffer will decrease the waiting times. The buffer size effect for the non-pre-emptive queue model can be seen in Fig 4.31. The buffer is set to K = 2. The mean length and waiting times decrease with a decrease in buffer length.

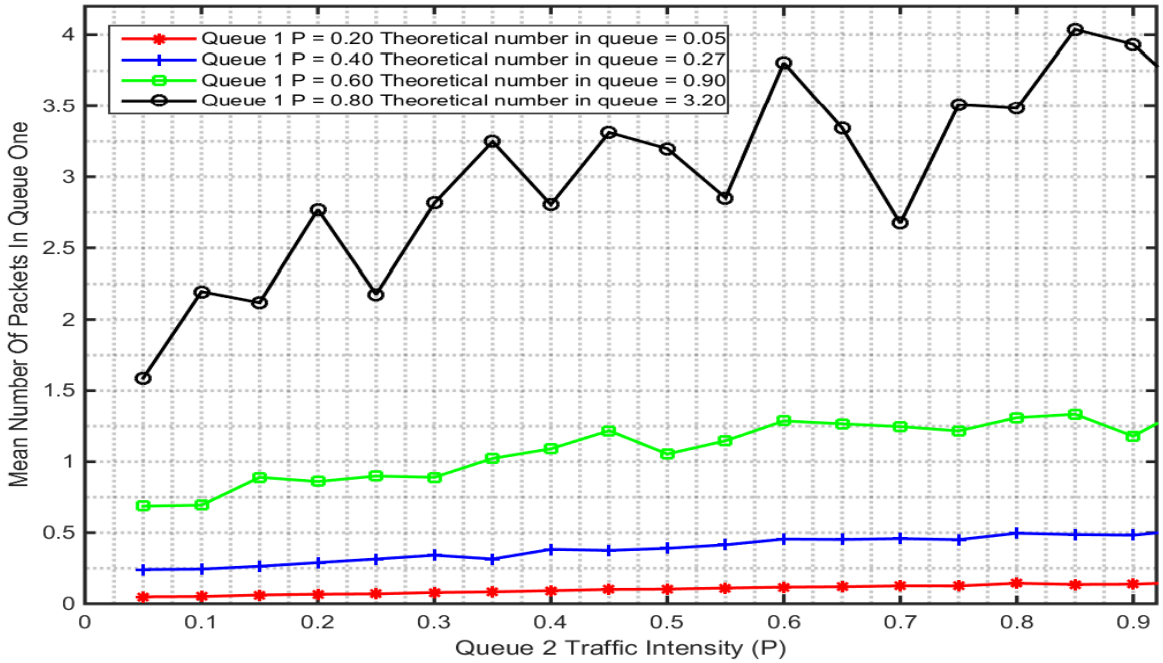

Figure 4.31. Queue 1 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

Fig. 4.32 shows the mean queue length of the LW queue for the non-pre-emption model under the same conditions as in Fig. 4.17. The LW queue has much a much lower expected queue length than that of the HW queue.

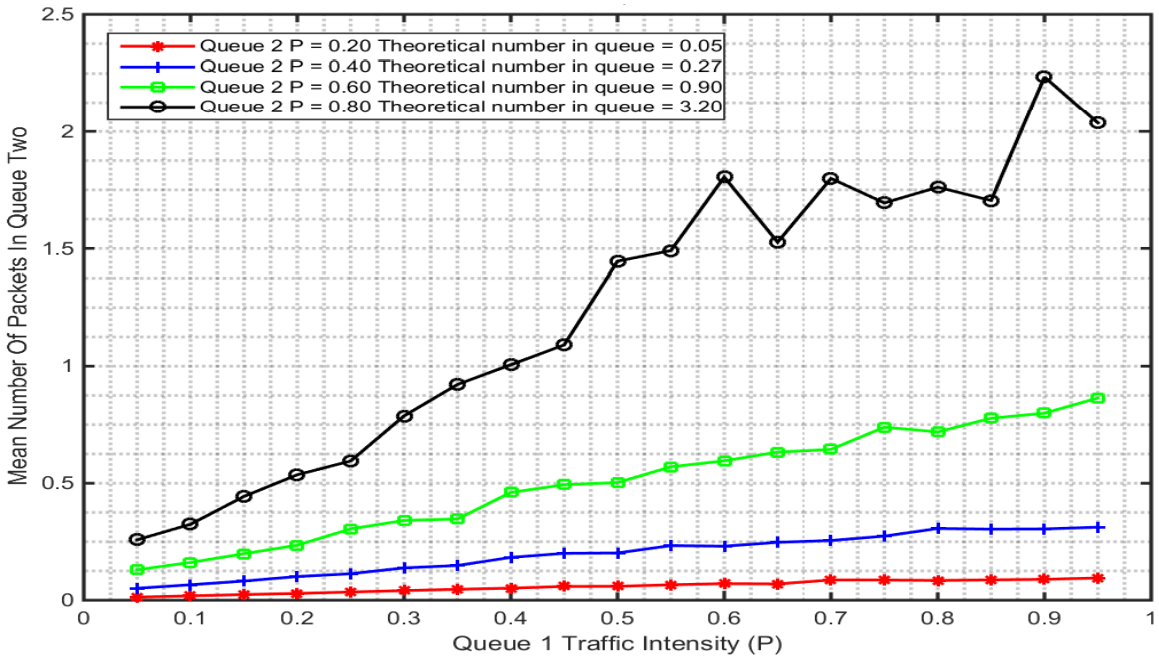

Figure 4.32. Queue 2 average number of packets. Here $\mu_1 = 0.00041$ sec/packet, $\mu_3 = 0.00042$ sec/packet, $\mu_2 = 0.00045$ sec/packet and $\mu_4 = 0.00096$ sec/packet

4.6 OPTIMISATION

The power allocation ratio r has a significant effect on the system performance. The ratio affects the service rates, hence the traffic intensity. This leads to the average number and waiting times in the queues being affected and this in turn means the optimisation is affected. The goal is to determine a power allocation ratio point that after taking all this into account, will allow an optimal solution to be found.

The power allocation ratio, r, will varied by decreasing the allocation to the HW while simultaneously increasing the allocation to LW by the same amount. Let θ be a variable such that $r_1 = 1 - \theta$ and $r_2 = \theta$. Fig. 4.33 shows the functions $f(x)$ and $f(y)$ varied across θ. Convexity was established in Chapter 3. The algorithm applied is given as follows:

Require: r_i $i \in \{1,2\}$

1: Determine μ_1 and μ_3. Use initial r_i to define the weights allocated to each SU.
2: Make weights equal to initial r_i
3: $r_1 = 1, r_2 = 0$
4: **while** $r_1 > 0.01$ **do**
5: Decrease r_1 by 0.01

6: Increase r_2 by 0.01

7: Determine μ_2 and μ_4

8: Run queue simulation

9: **end while**

10: Apply weights and sum the two queues

11: Plot graph of simulation

12: Obtain r_i for minimum point

4.6.1 Non-pre-emptive model results

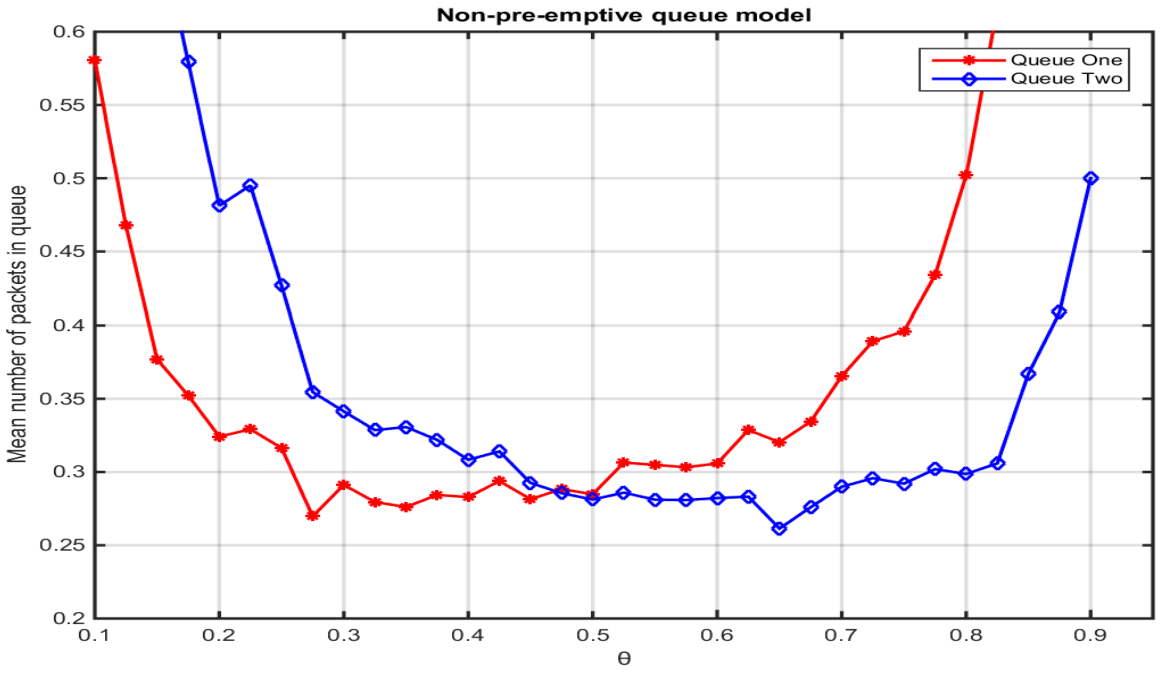

Figure 4.33. Mean number of packets for Queues 1 and 1 showing the convex property.

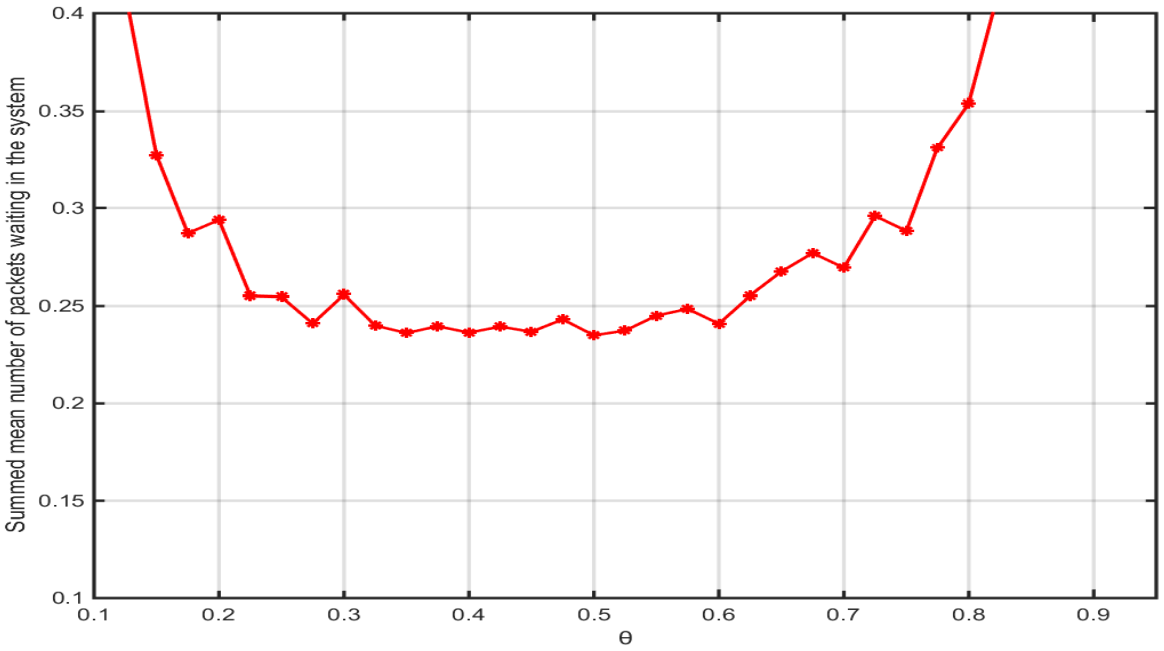

Figure 4.34. Mean number of packets for the system. $\phi = 0.8$, $\beta = 0.2$

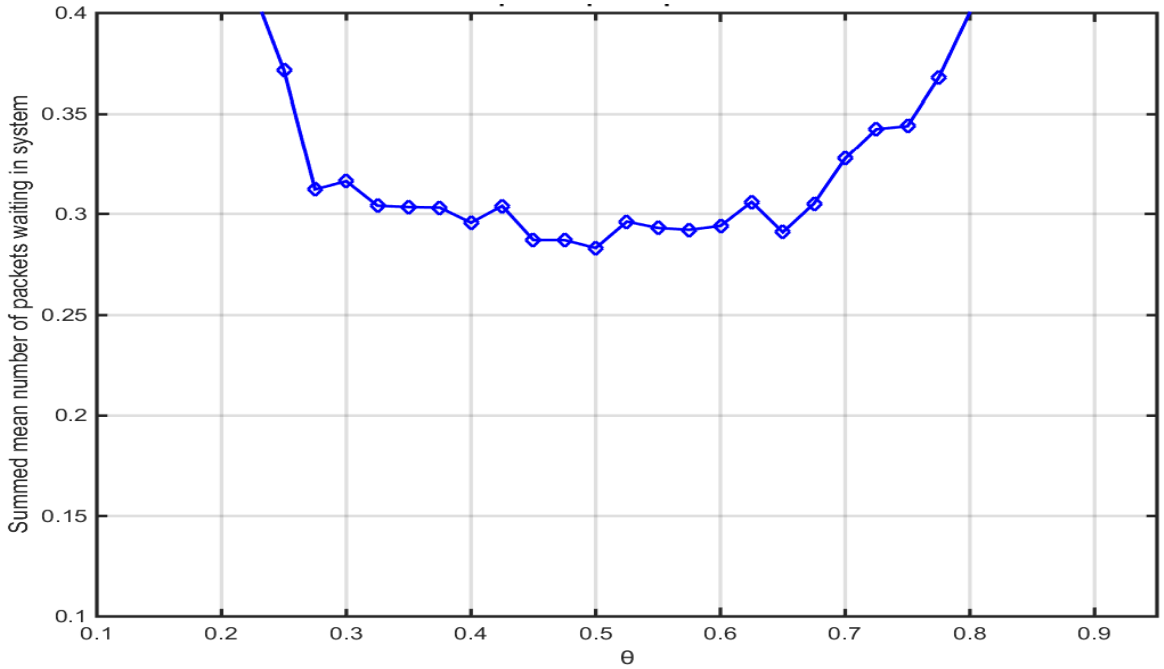

Figure 4.35. Mean number of packets for the system. $\phi = 0.6$, $\beta = 0.4$

Fig. 4.34 shows the minimum point for the graph to be around $\theta = 0.4$. That implies that the power allocation ratio is 60% of the power to the HW queue for optimum results despite it being weighted at 80% initially.

Fig. 4.35 shows the minimum point for the graph to be around $\theta = 0.35$. This would dedicate 65% of the power to the HW queue for optimum results.

4.6.2 Pre-emptive model results

The optimisation of a scenario for the pre-emptive queue model is given in Fig.4.36 for completeness. The behaviour is similar to the non-pre-emptive queue model.

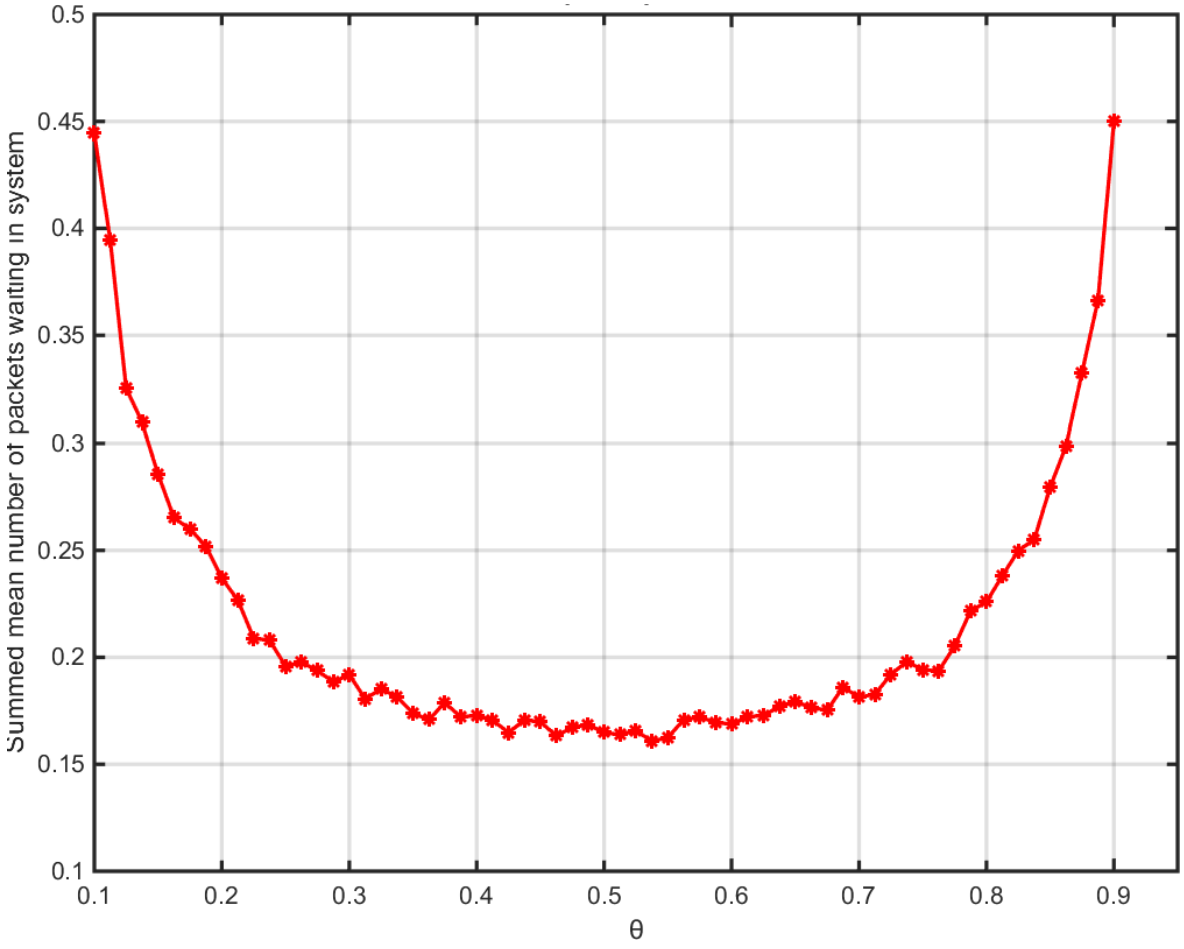

Figure 4.36. Mean number of packets for the pre-emptive model with equal weighting

4.7 COMPARISON

The proposed model is compared to a standard M/M/1 two-class priority system in Tables 4.1 to 4.4. A standard priority system completely halts service for the LW queue if there are HW packets to be transmitted. Here the

results are measured when queues 1 and 2 are at 80% and 20% traffic intensity respectively. The power ratio is the power allocated to Queue 1/power allocated to Queue 2.

Table 4.1. Comparison Between Processor Sharing and Standard Priority Average Queue Lengths: Pre-emptive Model

Model	Queue 1	Queue 2	Total
Processor Sharing Power Ratio 80/20	2.5	0.13	2.6
Standard Priority	0.9	20	20.9
Optimised Processor Sharing Power Ratio 80/20	1.5	0.7	2.2

Table 4.2. Comparison Between Processor Sharing and Standard Priority Average Queue Lengths: Non-pre-emptive Model

Model	Queue 1	Queue 2	Total
Processor Sharing Power Ratio 80/20	2.5	0.7	3.2
Standard Priority	1.4	>20	>20
Optimised Processor Sharing Power Ratio 80/20	2	1	3

Table 4.3. Comparison Between Processor Sharing and Standard Priority Average Queue Waiting Times: Pre-emptive Model

Model	Queue 1	Queue 2	Total
Processor Sharing Power Ratio 80/20	1.4 ms	0.1 ms	1.5 ms
Standard Priority	0.5 ms	11 ms	11.5 ms
Optimised Processor Sharing Power Ratio 80/20	0.88 ms	0.4 ms	1.2 ms

Table 4.4. Comparison Between Processor Sharing and Standard Priority Average Queue Waiting Times: Non-pre-emptive Model

Model	Queue 1	Queue 2	Total
Processor Sharing Power Ratio 80/20	1.4 ms	0.5 ms	1.9 ms
Standard Priority	0.5 ms	13 ms	13.5 ms
Optimised Processor Sharing Power Ratio 80/20	1.1 ms	0.6 ms	1.7 ms

The tables show that the proposed sharing scheme is beneficial in terms of the mean number of packets in the system overall. This means the second SU can be reliably served simultaneously with the first SU, thus increasing the number of users making use of the spectrum.

CHAPTER 5 DISCUSSION

5.1 CHAPTER OVERVIEW

The results presented in Chapter 4 are discussed extensively. Explanations of the results are discussed and comparisons are made. Firstly, the queueing considerations that were done before setting up the models are examined. Then the queuing results themselves are discussed separately for each model and explanations are given for the difference in behaviour. The optimisation results are explained with reference to the figures given in Chapter 4 and how optimisation affects the system. Lastly a discussion on which model is better for which conditions follows.

5.2 PROPOSED COGNITIVE RADIO NETWORK

The results indicate that the proposed network is indeed feasible. Two SUs can reliably transmit packets simultaneously on the same channel. The channel conditions have considerable effects on the network performance. From Fig. 4.2 it can be seen that varying the channel coefficients will affect the transmission time and hence the service time. The service time will in turn affect the entire system.

The respective channel links also play a significant role. For example if the SU-to-SU is link very good but the SU-to-PU link is poor the SU can transmit at slightly increased power and will experience lower transmission times thus better performance. However, if the link from the SU to the PU is very good, there is a higher chance of interference and the SU has to transmit at lower power regardless of the condition of the channel to its own SU receiver.

5.3 QUEUEING CONSIDERATIONS

The domain and queue type were the factors considered during the queue design.

5.3.1 Queueing domain

Continuous time domain modelling was selected over discrete time modelling to allow for ease of analysis. In discrete time, packets can arrive in and leave the system simultaneously; however, in continuous time this is not possible because there will be a difference in time between events, however small. The channel considered can be modelled as exponential using the path loss propagation model and hence an exponential service queueing model is sufficient. Considering a general service distribution creates the problem of where to embed the Markov chain. Since there are at least two queues that are dependent on each other the immediate problem becomes where to embed this chain, on departure of the first queue or on departure of the second queue. Either choice will mean that the other queue cannot be studied sufficiently.

For these reasons, continuous time queue models are sufficient to provide strong proof of concept.

5.3.2 Phase type

Upgrading the system to phase type queuing is possible but a rather Herculean task. The queues are interdependent on each other and would have to be assigned at least four phases per queue for the non-pre-emptive queue model: the first phase to indicate the currently in service, the second phase to indicate when the SU is the only one present, the third phase to indicate when both SUs are being served and the fourth phase to indicate when the SU is waiting to be served.

Incorporating all these phases into the phase type matrix is very difficult to achieve. Extending the model to phase type is therefore possible but does not add much value to the model at this stage.

5.4 QUEUEING RESULTS

The system must only be observed after a long time of activity so that the arrival rate can be assumed and measured and the service time is known for steady state analysis.

In addition, an important condition is that the queue must be stable, i.e. the traffic intensity, ρ, must be less than one. The service rate cannot be less than the arrival rate otherwise the queue will continue to grow uncontrollably. Since power allocation affects the service rate any power allocated to the SUs must take stability into account.

Fluctuations in the simulations are due to the randomness of the exponential function. A generally increasing trend in all the graphs can still be observed.

5.4.1 Pre-emptive model

The results show that, as expected, there is a trade-off between the two queues. Allocating more transmission power to one queue will result in increased performance from that queue but decreased performance on the other queue. Figs. 4.5 and 4.6 show the behaviour of the HW and LW queues under the same conditions. The HW queue performs much better because of the higher power allocation. Figs. 4.7 and 4.8 show when the power allocation has been adjusted. The performance of the HW has decreased and that of the LW has increased. The HW queue outperforms the LW queue because the allocation still favours the HW queue. However, from the point of view the LW queue alone it experiences a major performance benefit.

Fig. 4.7 shows the behaviour of the system when the HW queue power allocation has been reduced. Comparing this to Fig. 4.5 when the allocation is 20% higher, it can be seen that the average number of packets in the system has increased. This pattern is observed up to when the allocation is 50%, as shown in Fig. 4.9. This proves that the power allocation does indeed affect the service rate, which in turn affects the system performance.

In Fig. 4.31 a buffer size of two has been imposed on the LW queue. This means that a maximum of only three packets on the LW system can be present at any time, two in the queue and one being served. Whereas there is much improvement in terms of the performance of the HW queue, the overall network performance does not improve as much. The reason for this is that the LW queue is now serving fewer packets. Packets that arrive when the system is full are blocked and are lost.

The figures for the waiting are similar to the ones for the average length because of the relationship according to Little's Law. The less power allocated to the queue, the higher the delay. Hence there is more waiting time going from Figs. 4.11 to 4.15. Similarly, the addition of a buffer in Fig. 4.17 will see much improvement in both the HW and the LW queues. However packets are still lost and the overall number of packets served in the LW queue is reduced as well.

The theoretical values are derived from a standard M/M/1 queue. That means that, they show the expected average length and waiting times in the queue, provided that the queues are independent and working at the allocated ratio. The improved performance of the proposed model over the theoretical model can be attributed to allowing either queue to tap into the extra power left by the other queue's vacuum. Eventually, with an increase in the LW queue traffic intensity, the proposed model's performance will deteriorate to just under the theoretical value. This is because since there is now much more activity in the LW queue, the HW queue is mostly operating

at the allocated power and only rarely reaches the maximum allowable power. Hence the conditions are almost exactly similar to a standard M/M/1 queue. However, the LW queue will be empty at some point and hence the performance will never be same as the theoretical M/M/1 queue value.

5.4.2 Non-pre-emptive model

The results show that similar to the pre-emptive queuing model and as expected, there is a trade-off between the two queues. Allocating more transmission power to one queue will result in increased performance from that queue but decreased performance on the other queue. Figs. 4.19 and 4.20 show the behaviour of the HW and LW queues under the same conditions. The HW queue is performs much better because of the higher power allocation. Figs. 4.21 and 4.20 illustrate the performance when the power allocation has been adjusted. The performance of the HW has decreased and that of the LW has increased. The HW queue outperforms the LW queue because the allocation still favours the HW queue.

In Fig. 4.31 a buffer size of two has been imposed on the LW queue. Whereas there is much improvement in terms of the performance of the HW queue, the overall network performance does not improve as much. The reason for this is that the LW queue is now serving significantly fewer packets. Similar to the pre-emptive model, packets that arrive when the system is full are blocked and are lost.

Little's Law also holds here. The less power allocated to the queue, the longer the delay. Hence there is more waiting time going from Figs. 4.25 to 4.29. Similarly, the addition of a buffer in Fig. 4.31 will see much improvement in both the HW and the LW queues. However, packets are still lost and the overall number of packets served in the LW is reduced as well.

Unlike the pre-emptive model, the non-pre-emptive model performance can become worse than the theoretical value. This is due to the fact that the non-pre-emptive rule will institute an extra delay in the system. In Fig. 4.19 the theoretical value for when the HW queue is at 80% traffic intensity and the LW queue is at 60%, is higher than the expected value for the M/M/1 queue in those conditions. This is explained by the fact that since the LW queue is much busier, the HW packets will have to wait longer in the queue, thus increasing the queue length, since the arrival rate is not affected. The explanation is that because of the higher traffic intensity of the LW queue, there is a higher chance that a packet arriving at the HW queue will find an LW packet already in service. The sojourn time for the HW packet will then include the time it spent waiting for service.

5.5 OPTIMISATION RESULTS

Given the difficulty in finding the objective function, convex optimisation is very useful in this situation. The solution to the minimisation problem becomes the minimum point of the graph. The only constraint to the minimisation problem was the power allocation ratio, which was varied through the range of possible values. The objective function shows pseudo-convexity and is sufficient to determine the minimum point. In addition, after applying scaling, the objective function was determined to be convex.

There is improvement in the queue performance after optimisation. However, the improvements are marginal, even for large variations of the power allocation ratio. The reason for this is that since the system has already been allocated weights, optimisation will not change the performance much outside the allocated weights.

The optimisation system is a feedback system in which any change in the power ratio will change the service rate and as a result will change the system performance.

5.6 PRE-EMPTIVE VS. NON-PRE-EMPTIVE MODEL

The choice of the best model comes down to individual network requirements. The overall performance of each model is almost similar. With this in mind, however, the non-pre-emptive model may suit networks that have reliability and consistency as key requirements. This is supported by the fact that the model experiences fewer interruptions even when the traffic intensity is high. This can be seen from the extended waiting times when comparing the two models in Fig. 4.13 and Fig. 4.14. The pre-emptive model will suit networks that are time- or delay-sensitive better. A packet will enter service as soon as it arrives ensuring that service is always provided. This is supported by the results in Table 4.3. The pre-emptive model has shorter waiting times than the non-pre-emptive model in Table 4.4.

CHAPTER 6 CONCLUSION

6.1 CONCLUSION

In this dissertation a solution to the resource allocation problem is developed. A technique adapting weighted processor sharing queues to include at least two SUs per channel in an underlay CRN is proposed.

- In Chapter 2, extensive literature on the resource allocation problem is discussed. Insight into why it is a problem, why it exists and why it should be solved is given. Background on CRNs is given, specifically the challenges and criteria that make up a CRN and rules that must be followed. The different types of CRNs are also examined. Queuing theory, which is an integral part of this research, is introduced. Specific queueing aspects, namely processor sharing and M/M/1 queues, are discussed. Finally optimisation is presented. Background on what has been done is provided and insight into convex optimisation is given.
- In Chapter 3, the methodology of the research is presented, the assumptions made are given and the CRN network models and behaviour are defined graphically. The transition matrices are given to support the defined models and validate the simulation approach. The optimisation technique employed is presented and the work done in order to prove that convexity can be established is also presented.
- In Chapter 4, the simulation results obtained from multiple varying scenarios are given. The results are extensive so as to capture the approach undertaken and observe any patterns. Results are presented in the form of performance measures. Finally, optimisation of the said performance measures is done in order to find the best-performing network scenario.
- In Chapter 5, an overall discussion is given on the research. The results from Chapter 4 are extensively discussed and reasons are given to explain their nature. Comparisons are done to determine if the proposed solution is indeed an improvement on existing solutions or not.

6.2 RESULTS ACHIEVED

A technique to allow at least two SUs to operate simultaneously with a PUon the same channel in CRN has been developed. The results indicate that the proposed solution has great potential to improve on already existing solutions. Optimisation is also applied to find the best possible scenario to maximise network performance.

Thus in brief, a centralized underlay CRN is presented to allow at least two SUs to exist by giving them the ability to determine the amount of transmission power available to use and make maximum use of it, all while being controlled by some network-defined criteria.

The applications can be used in areas where spectrum is already strained, such as densely populated urban areas where the ISM band is beginning to suffer from overcrowding. Another application is in situations where infrastructure is limited and transmission bands cannot be changed or increased.

6.3 FUTURE WORK

The research presented herein could be used to further research on aspects such as the analysis domain and more SUs.

6.3.1 Analysis domain

As more and more telecommunications systems are being analysed in discrete time rather than continuous time, it will be worthwhile to extend the proposed queueing models to discrete time. Although the results show that the continuous time domain is indeed sufficient for length and delay analysis, discrete time models will extend the analysis into other aspects, such as busy period distribution of the two queues. The challenge, however, is to develop an accurate model based in discrete time because given the nature of the domain, two events can occur in a given discrete block, i.e. a departure and an arrival are allowed to occur at the same time.

6.3.2 Secondary users

Analysis of a higher number of SUs can also be undertaken. This work can be extended to determine the maximum number of SUs that a particular network will be able to handle. The goal of the work is to improve spectrum utilisation and efficiency as much as possible. The results have indicated that adding an additional SU will certainly improve network utilisation. Therefore adding more than two SUs must improve the network

even further. However, this is not absolute. There is a limit that once it is reached, the network will cease to be beneficial to its users and long waiting times and very long queues will ensue. However, a technique can be developed to manage additional SUs. Game theory, for example, can be implemented to keep only two SUs active at a time among a pool of many users. The effects of having always present SUs can be investigated using the already developed algorithm.

6.3.3 Queue models

Other queue models can be introduced in attempting to solve the problem. One such model may involve the use of time-varying queue behaviour to model the service time. That implies that, the service time is observed as non-constant and allowed to change, depending on the activity of the SUs. The Mt/Mt/1 queuing model has great potential in progressing the models developed. The developed models can be defined as time-varying and will therefore fit the queueing model.

In addition the service is modelled as an exponential service because of the nature of the developed queue. The models can be extended to more realistic real life scenarios such as fading channels. This poses the immediate challenge of changing the service distribution into a general distribution and since there are two dependent queues, the observation point becomes difficult to attain. One suggestion may be to obtain the busy period distribution of the two queues. These may be used to formulate a single general distribution for the entire system. A departure may thus simply be observed as a departure from the system and not necessarily as a departure from a particular queue. This may very well solve the problem of finding the point to embed the queue.

REFERENCES

[1] International Telecommunications Union. (2016) ITU key 2005 - 2016 ICT data. [Online]. Available: http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

[2] J. Mitola and G. Q. Maguire, "Cognitive radio: making radios more personal," *IEEE Personal Communication*, vol. 6, no. 4, pp. 13–18, Aug 1999.

[3] E. Tragos, S. Zeadally, A. G. Fragkiadakis, and V. A. Siris, "Spectrum assignment in cognitive radio networks: A comprehensive survey," *IEEE Communication Surveys and Tutorials*, vol. 15, no. 3, pp. 1108–1134, Jan 2013.

[4] X. Li, T. Drive, and S. A. R. Zekavat, "Distributed channel assignment in cognitive radio networks," in *Proc. IEEE Conference on Wireless Communication and Mobile Computiong: Connecting the World Wirelessy*, Jun 2009, pp. 989–993.

[5] H. Shiang and M. van der Scaar, "Queueing-based dynamic channel selection for heterogenous multimedia application over cognitive radio networks," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 896–909, Jun 2008.

[6] L. Wang, C. Wang, and K. Feng, "A queueing-theorectical framework for QoS-enhanced spectrum management in cognitive radio networks," *IEEE Wireless Communication*, vol. 18, no. 6, pp. 18–26, Dec 2011.

[7] A. S. Alfa, *Queueing Theory for Telecommunications: Discrete Time Modelling of a Single Node System*, 1st ed. Springer, 2010.

[8] M. Zhang, S. Jiang, G. Wei, and H. wang, "Performance analysis of the cognitive radio network with a call level queue for secondary users," in *Proc. International Conference on Wireless Communications, Networking and Mobile Computing*, Oct 2009, pp. 1–4.

[9] T. M. C. Chu, H. Phan, and H. Zerperneck, "On the performance of underlay cognitive radio networks: An M/M/1 queue," *IEEE Communication Letters*, vol. 17, no. 5, pp. 876–879, Mar 2013.

[10] F. T. Ulaby, *Fundamentals of Applied Electromagnetics*, 7th ed. Pearson, 2015.

[11] Federal Communications Commission. (2016) Government affairs policy briefs on spectrum allocation. [Online]. Available: https://www.fcc.gov/engineering-technology/policy-and-rules-division/general/radio-spectrum-allocation

[12] B. P. Lathi and Z. Ding, *Modern Digital and Analog Communications*, 4th ed. Oxford, 2010.

[13] N. Orozco, H. Carvajal, G. Olmedo, R. Leon, and C. de Almeida, "UMTS/HSPA and LTE cellular systems: On the frequency bands and the bit error rate," in *Proc. 2016 IEEE Colombian Conference on Communications and Computing (COLCOM)*, Apr 2016, pp. 1–6.

[14] Indepedent Communications Authority of South Africa. (2015, Mar) Radio frequency spectrum regualations 2015. [Online]. Available: https://www.icasa.org.za/LegislationRegulations/FinalRegulations/MiscellaneousRegulations

[15] Akamai. (2017) State of the internet report 2017. [Online]. Available: https://www.akamai.com/us/en/about/our-thinking/state-of-the-internet-report/

[16] H. Beyranvand, M. Levesque, M. Maier, J. A. Salehi, C. Verikoukis, and D. Tipper, "Toward 5G: FiWi enhanced LTE-A HetNets with reliable low-latency fiber backhaul sharing and WiFi offloading," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 690–707, Apr 2017.

[17] M. Ruffini, "Multidimensional convergence in future 5G networks," *Journal of Lightwave Technology*, vol. 35, no. 3, pp. 535–549, Feb 2017.

[18] D. Moongilan, "5G wireless communications (60 GHz band) for smart grid; An EMC perspective," in *Proc. 2016 IEEE International Symposium on Electromagnetic Compatibility (EMC)*, Jul 2016, pp. 689–694.

[19] X. Ge, H. Cheng, G. Mao, Y. Yang, and S. Tu, "Vehicular communications for 5G co-operative small-cell networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 7882–7894, Oct 2016.

[20] Z. Wang, B. Hu, X. Wang, and S. Chen, "Interference pricing in 5G ultra-dense small cell networks: a Stackelberg game approach," *IET Communications*, vol. 10, no. 15, pp. 1865–1872, Oct 2016.

[21] S. Din, A. Paul, A. Ahmad, and S. Rho, "Emerging mobile communication technologies for healthcare system in 5G network," in *Proc. 2016 IEEE 14th International Conference on Dependable, Autonomic and Secure Computing, 14th International Conference on Pervasive Intelligence and Computing, 2nd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, Aug 2016, pp. 47–54.

[22] H. Yang, B. C. Seet, S. F. Hasan, P. H. J. Chong, and M. Y. Chung, "Radio resource allocation for D2D-enabled massive machine communication in the 5G Era," in *Proc. 2016 IEEE 14th International Conference on Dependable, Autonomic and Secure Computing, 14th International Conference on Pervasive Intelligence and Computing, 2nd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, Aug 2016, pp. 55–60.

[23] D. Treeumnuk and D. C. Popescu, "Enhanced spectrum utilisation in dynamic cognitive radios with adaptive sensing," *IET Signal Processing*, vol. 8, no. 4, pp. 339–346, Jun 2014.

[24] Indepedent Communications Authority of South Africa. (2017, Mar) 2nd report on the state of ict sector in South Africa. [Online]. Available: https://www.icasa.org.za/LegislationRegulations

[25] Institute of Race Relations. (2017) South Africa survey 2017: Communications. [Online]. Available: http://irr.org.za/reports-and-publications/south-africa-survey/south-africa-survey-2017

[26] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Prentice Hall PTR, 2012.

[27] A. Jaziri, R. Nasri, and T. Chahed, "Congestion mitigation in 5G networks using drone relays," in *Proc. 2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Sep 2016, pp. 233–238.

[28] A. Saeed, M. Ibrahim, K. A. Harras, and M. Youssef, "Toward dynamic real-time geo-location databases for TV white spaces," *IEEE Network*, vol. 29, no. 5, pp. 76–82, Sep 2015.

[29] F. Marino, L. Paura, and R. Savoia, "On spectrum sensing optimal design in spatial and temporal domain for cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 10, pp. 8496–8510, Oct 2016.

[30] O. Holland, S. Ping, A. Aijaz, J. M. Chareau, P. Chawdhry, Y. Gao, Z. Qin, and H. Kokkinen, "To white space or not to white space: That is the trial within the OFCOM TV white spaces pilot," in *Proc. 2015 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, Sep 2015, pp. 11–22.

[31] Y. Xu, J. Chen, P. Kang, X. Zhu, and J. Ding, "Quantifying the availability of TV white spaces for cognitive LTE systems in China," in *Proc. 2015 IEEE 6th International Symposium on Microwave, Antenna, Propagation, and EMC Technologies (MAPE)*, Oct 2015, pp. 779–783.

[32] R. Kennedy, K. George, O. Vitalice, and W. Okello-Odongo, "TV white spaces in Africa: Trials and role in improving broadband access in Africa," in *Proc. IEEE AFRICON 2015*, Sep 2015, pp. 1–5.

[33] A. A. L. et al, "First large TV white spaces trial in South Africa: A brief overview," in *Proc. 2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Oct 2014, pp. 407–414.

[34] P. Johnson, "New research lab leads to unique radio receiver," *E-Systems Team*, vol. 5, no. 4, pp. 6–7, May 1985.

[35] I. F. Akyildiz, W. Lee, M. C. Mohanty, and S. Vuran, "Next generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer Networks*, vol. 50, no. 13, pp. 2127–2159, May 2006.

[36] A. Ahmad, S. Ahmad, M. H. Rehmani, and N. U. Hassan, "A survey on radio resource allocation in cognitive radio sensor networks," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 888–917, Feb 2015.

[37] A. Plummer, "Distributed spectrum assignment for cognitive networks with heterogeneous spectrum opportunities," *Wireless Communications and Mobile Computing*, vol. 11, no. 9, pp. 1239–1253, Sep 2011.

[38] M. Pischella and D. L. Ruyet, "Cooperative allocation for underlay cognitive radio systems," in *Proc. 2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Jun 2013, pp. 245–249.

[39] J. Oh and W. Choi, "A hybrid cognitive radio system: A combination of underlay and overlay approaches," in *Proc. 2010 IEEE 72nd Vehicular Technology Conference - Fall*, Sep 2010, pp. 1–5.

[40] E. Biglieri, A. J. Goldsmith, L. J. Greestein, N. B. Mandayan, and H. Vincent, *Principles of Cognitive Radio*, 1st ed. Cambridge University Press, 2012.

[41] T. Chen, H. Zhang, G. M. Maggio, and I. Chlamtac, "Topology management in cogmesh: A cluster-based cognitive radio mesh network," in *Proc. 2007 IEEE International Conference on Communications*, Jun 2007, pp. 6516–6521.

[42] H. Wang, J. Ren, and T. Li, "Resource allocation with load balancing for cognitive radio networks," in *Proc. 2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, Dec 2010, pp. 1–5.

[43] W. Wang, K. G. Shin, and W. Wang, "Joint spectrum allocation and power control for multihop cognitive radio networks," *IEEE Transactions on Mobile Computing*, vol. 10, no. 7, pp. 1042–1055, Jul 2011.

[44] J. Wang and Y. Huang, "A cross-layer design of channel assignment and routing in cognitive radio networks," in *Proc. 2010 3rd International Conference on Computer Science and Information Technology*, vol. 7, Jul 2010, pp. 542–547.

[45] T. H. Kim and T. J. Lee, "Spectrum allocation algorithms for uplink sub-carriers in OFDMA-based cognitive radio networks," in *Proc. 2007 Innovations in Information Technologies (IIT)*, Nov 2007, pp. 51–54.

[46] R. Hou, K. S. Lui, and J. Li, "Routing in multi-radio multi-channel multi-hop wireless mesh networks with bandwidth guarantees," in *Proc. 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, May 2011, pp. 1–5.

[47] G. Cheng, W. Liu, Y. Li, and W. Cheng, "Joint on-demand routing and spectrum assignment in cognitive radio networks," in *Proc. 2007 IEEE International Conference on Communications*, Jun 2007, pp. 6499–6503.

[48] S. Gao, L. Qian, and D. Vaman, "Distributed energy efficient spectrum allocation based on spectrum difference," in *Proc. 2008 IEEE Wireless Communications and Networking Conference*, 2008, pp. 1442–1447.

[49] Y. Li, Z. Wang, B. Cao, and W. Huang, "Impact of spectrum allocation on connectivity of cognitive radio ad-hoc networks," in *Proc. 2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, Dec 2011, pp. 1–5.

[50] S. S. Byun, I. Balasingham, and X. Liang, "Dynamic spectrum allocation in wireless cognitive sensor networks: Improving fairness and energy efficiency," in *Proc. 2008 IEEE 68th Vehicular Technology Conference*, Sep 2008, pp. 1–5.

[51] C. Peng, H. Zheng, and B. Y. Zhao, "Utilization and fairness in spectrum assignment for opportunistic spectrum access," *Mobile Networks and Applications*, vol. 11, no. 4, pp. 555–576, May 2006.

[52] T. Zhang, B. Wang, and Z. Wu, "Spectrum assignment in infrastructure based cognitive radio networks," in *Proc. IEEE 2009 National Aerospace Electronics Conference (NAECON)*, Jul 2009, pp. 69–74.

[53] L. B. Le and E. Hossain, "Resource allocation for spectrum underlay in cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 12, pp. 5306–5315, Dec 2008.

[54] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems," Hudson, MA, Tech. Rep. DEC-TR-301, Sep 1984.

[55] A. B. Sediq, R. H. Gohary, and H. Yanikomeroglu, "Optimal tradeoff between efficiency and Jain's fairness index in resource allocation," in *Proc. 2012 IEEE 23rd International Symposium on Personal, Indoor and Mobile Radio Communications - (PIMRC)*, Sep 2012, pp. 577–583.

[56] L. Akter and B. Natarajan, "Modeling fairness in resource allocation for secondary users in a competitive cognitive radio network," in *Proc. 2010 Wireless Telecommunications Symposium (WTS)*, Apr 2010, pp. 1–6.

[57] N. Li, L. Gong, and S. Li, "Price-based spectrum-allocation relay routing in cognitive radio networks," in *Proc. 2009 7th International Conference on Information, Communications and Signal Processing (ICICS)*, Dec 2009, pp. 1–5.

[58] H. Li and L. Qian, "Enhancing the reliability of cognitive radio networks via channel assignment: Risk analysis and redundancy allocation," in *Proc. 2010 44th Annual Conference on Information Sciences and Systems (CISS)*, Mar 2010, pp. 1–6.

[59] H. He, H. Shan, A. Huang, and L. Sun, "SMDP-based resource allocation for video streaming in cognitive vehicular networks," in *Proc. 2015 IEEE/CIC International Conference on Communications in China (ICCC)*, Nov 2015, pp. 1–6.

[60] H. A. B. Salameh, M. Krunz, and O. Younis, "Cooperative adaptive spectrum sharing in cognitive radio networks," *IEEE/ACM Transactions on Networking*, vol. 18, no. 4, pp. 1181–1194, Aug 2010.

[61] A. Mackenzie and L. DaSilval, *Game Theory for Wireless Engineers*, 4th ed. Morgan and Claypool, 2006.

[62] A. T. Hoang and Y. C. Liang, "Downlink channel assignment and power control for cognitive radio networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 8, pp. 3106–3117, Aug 2008.

[63] R. Diestel, *Graph Theory: Springer Graduate Text*, 4th ed. Springer, 2012.

[64] T. Shu and M. Krunz, "Exploiting microscopic spectrum opportunities in cognitive radio networks via coordinated channel access," *IEEE Transactions on Mobile Computing*, vol. 9, no. 11, pp. 1522–1534, Nov 2010.

[65] P. Kaur, M. Uddin, and A. Khosla, "Fuzzy based adaptive bandwidth allocation scheme in cognitive radio networks," in *Proc. 2010 Eighth International Conference on ICT and Knowledge Engineering*, Nov 2010, pp. 41–45.

[66] Z. Zhao, Z. Peng, S. Zheng, and J. Shang, "Cognitive radio spectrum allocation using evolutionary algorithms," *IEEE Transactions on Wireless Communications*, vol. 8, no. 9, pp. 4421–4425, Sep 2009.

[67] X. Yu and M. Gen, *Introduction to evolutionary algorithms*, 1st ed. Springer, 2010.

[68] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, 2nd ed. John Wiley and Sons Inc, 1985.

[69] A. A. Markov, "Extension of limit theorems of probability theory to sum of variables connected in a chain," *The Notes of the Imperial Academy of Sciences of St Petersburg VIII Series, Physio-Mathematical College XXII*, no. 9, Dec 1907.

[70] X. L. Huang, X. Wu, J. Wu, Y. Xu, and M. Wang, "Queuing theory based spectrum allocation in cognitive radio networks," in *Proc. 2014 Sixth International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct 2014, pp. 1–6.

[71] A. K. Farraj, S. L. Miller, and K. A. Qaraqe, "Queue performance measures for cognitive radios in spectrum sharing systems," in *Proc. 2011 IEEE GLOBECOM Workshops (GC Wkshps)*, Dec 2011, pp. 997–1001.

[72] B. Zaman, Z. H. Abbas, and F. Y. Li, "Spectrum occupancy and residual service analysis in CRNs using a multi-server queueing model," in *Proc. 2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015, pp. 1–5.

[73] J. Zhang and E. J. Coyle, "The transient solution of time-dependent M/M/1 queues," in *Proc. First International Workshop on th Numerical Solution of Markov Chains*, Jan 1990, pp. 655–658.

[74] L. Chouhan and A. Trivedi, "Priority based MAC scheme for cognitive radio network: A queuing theory modelling," in *Proc. 2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN)*, Sep 2012, pp. 1–5.

[75] G. de Veciana and G. Kesidis, "Bandwidth allocation for multiple qualities of service using generalized processor sharing," in *Proc. 1994 IEEE GLOBECOM. Communications: The Global Bridge*, vol. 3, Nov 1994, pp. 1550–1554.

[76] R. Szabo, P. Barta, J. Biro, F. Nemeth, and C. G. Perntz, "Nonrate-proportional weighting of generalized processor sharing schedulers," in *Proc. Global Telecommunications Conference, 1999. GLOBECOM '99*, vol. 2, Dec 1999, pp. 1334–1339.

[77] A. G. Konheim, I. Meilijson, and A. Melkman, "Processor sharing of two parallel lines," *Journal of Applied Probability*, vol. 18, no. 4, pp. 952–956, Dec 1981.

[78] F. Beutler, "Mean sojourn times in Markov queueing networks: Little's formula revisited," *IEEE Transactions on Information Theory*, vol. 29, no. 2, pp. 233–241, Mar 1983.

[79] I. F. Akyildiz, W. Lee, M. C. Mohanty, and S. Vuran, "A survey on spectrum management in cognitive radio networks," *IEEE Communication Magazine*, vol. 50, no. 4, pp. 40–48, Apr 2008.

[80] K. Lange, *Optimization*, 2nd ed. Springer, 2013.

[81] A. S. Alfa, B. T. Maharaj, S. Lall, and S. Pal, "Mixed-integer programming based techniques for resource allocation in underlay cognitive radio networks: A survey," *Journal of Communications and Networks*, vol. 18, no. 5, pp. 744–761, Oct 2016.

[82] S. Boyd and C. Vandeberghe, *Convex Programming*, 1st ed. Cambridge University Press, 2004.

[83] H. Zhao, L. Chen, and W. Feng, "A signal detection scheme for wireless sensor networks based on convex optimization," in *Proc. 2016 IEEE Sensors*, Oct 2016, pp. 1–3.

[84] C. E. Shannon, "A mathematical theory of communications," *Bell Systems Technical Journal*, vol. 27, no. 4, pp. 623–656, Oct 1948.

[85] H. M. Tsimba, B. T. Maharaj, and A. S. Alfa, "Increased spectrum utilisation in a cognitive radio networks: An M/M/1-PS queue approach," in *Proc. Wireless Communication and Network Conference*, Mar 2017.

[86] M. F. Neuts, *Matrix-geometric Solutions in Stochastic Models*, 1st ed. John Hopkins University Press, 1981.

[87] D. R. Miller, "Computation of steady-state probabilities for M/M/1 priority queues," *Journal of Applied Probability*, vol. 29, no. 5, pp. 945–958, Oct 1981.

[88] W. L. Winston and M. Venkatarama, *Introduction to Mathematical Programming*, 4th ed. Cengage, 2002.

[89] O. L. Mangasarian, "Pseudo-convex functions," *Journal of the Society for Industrial and Applied Mathematica Series A Control*, vol. 3, no. 3, pp. 281–290, 1965.

[90] W. Whitt, *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues*. Springer, 2002.

[91] IEEE 802.22 Wireless RAN. Functional requirements for the 802.22 WRAN Standard, IEEE 802.22-05/0007r46, Oct 2005.

ADDENDUM A FURTHER QUEUEING THEORY

A.1 QUEUE NOTATION

Table A.1. Meaning of Various Queue Notation

Symbol	Meaning
M	Exponential/Poisson
D	Deterministic
E_k	Erlang type
HK	Hyper-exponential type
PH	Phase type
G	General
MAP	Markovian arrival process
FCFS/FIFO	First come, first served/ First in, first out
LCFS	Last come, first served
RSS	Random selection for service
PR	Priority
PS	Processor sharing
t	time varying (Mt - Poisson time varying)