Response Times in an Accident and Emergency Service Unit

Apurva Udeshi
aau03@doc.ic.ac.uk

Supervisor: Professor Peter Harrison
Second Marker: Dr. William Knottenbelt

June 14, 2006

www.doc.ic.ac.uk/~aau03
Abstract

Today’s world is such that it tends to be a bit of a rat – race in order to make our systems run faster. In systems such as healthcare, this has almost become a necessity, considering the fact that we are providing a service to patients where even a split – second makes a difference to their lives. The fact that there are conflicting objectives between patients trying to go through the hospital as fast as possible and the management trying to increase the utilisations of the staff makes this a very interesting problem to analyse.

Service and response time distributions are good measurements of measuring the quality of such a service accurately. Using patient data obtained from an A&E service unit of a major London hospital, this project looks into finding service and response time distributions of this system and trying to evaluate whether the quantiles of response time of the NHS have been achieved.

This project also looks into applying queuing theory in order to see the actual behaviour of such a system. The two main queues discussed in this project are the M/M/1 and M/M/m queue along with the calculation of the number of servers used when using the M/M/m queue. Response time and queuing time distributions of such queues at varying arrival rates are also studied using convolutions of service and queuing time distributions.

As an M/M/m queue doesn’t model the A&E service unit completely accurately, using the patient data, a Markovian network of queues is parameterised in order to model the A&E service unit. Nodes that prove to be bottlenecks in such a system are analysed further and where necessary, modified in order to cope with any increase in arrival rate of the A&E service unit. Furthermore, an additional node is added to such a system in order to see the effect on the response time and the bottlenecks in such a system.

Response time distributions of the different paths in such networks are calculated using Laplace transforms and compared to response time values obtained using Little’s Law.

Finally, the success of the project and areas of future work are discussed.
Acknowledgments

First and foremost, I would like to thank Professor Peter Harrison for all his help, advice and guidance throughout this project. His motivation has helped me develop a great interest in this area of queuing theory. If not for Pete’s constant replies to my e-mails and weekly meetings with me, (even when Molly was in his office and we had to use a different room as a result of my fear of dogs!), I doubt I would have been able to learn as much. Thanks a lot Peter, greatly appreciate it!

I would also like to thank Will Knottenbelt for agreeing to be my second marker, showing me some of the more practical aspects of the A&E service unit and always willing to help and discuss the progress of my project.

Furthermore, I would also like to thank Susanna Wau Men Au-Yeung for providing me with the patient’s data of the A&E service unit and constantly replying to my mails during the initial stages of my project.

On a more personal note, I would also like to thank my flat mate Akshay Hirdaramani for his constant moral support and last but not least my parents, who made it possible for me to study at Imperial College.
## Contents

1 Introduction ......................................................... 1  
1.1 Motivation ..................................................... 1  
1.2 Objectives ..................................................... 2  
1.3 Report Structure ............................................. 5  

2 Background ......................................................... 6  
2.1 Background to Accident and Emergency Service Unit .......... 6  
2.2 Methods used to Model Patient Flow in A&E Service Unit ..... 9  
2.3 History of Queuing Theory .................................... 9  
2.4 Background to Queuing Theory ................................ 10  
2.5 Useful Formulas ................................................ 12  
2.6 Steady State Theorem ......................................... 12  
2.7 Network of Queues ............................................. 13  
2.7.1 How a Network of Queues Work .......................... 13  
2.7.2 Type of Queues Used ...................................... 14  
2.7.3 Jackson’s Theorem .......................................... 15  
2.7.4 Bottlenecks in the System ................................. 15  
2.8 Response Time Distributions .................................. 16  
2.9 Laplace Transforms ............................................ 17  

3 Data Analysis ....................................................... 18  
3.1 Arrivals ......................................................... 18  
3.1.1 Inter-Arrival Time ........................................ 18  
3.2 Service Time ................................................... 19  
3.3 Response Time .................................................. 20  
3.3.1 Response Time of Patients for One Day ................. 22  
3.3.2 Response Time of Patients for One Week ............... 22  
3.4 Response Times of Different Classes of Patients .............. 23  
3.4.1 Response Times of Minors ................................. 23  
3.4.2 Response Times of Majors ................................ 25  
3.5 Quantiles of Response Times .................................. 26  
3.6 Treatment Profiles of Other Types of Patients ................. 27  
3.6.1 Proportion of Patients that left without being seen ..... 27  
3.6.2 Proportion of Patients that left without given treatment ........................................ 27  
3.6.3 Proportion of Patients that died whilst in the A&E ....... 27
4 Applying Queues to an A&E Service Unit

4.1 Modelling an A&E service unit using an M/M/1 Queue .......... 29
4.2 Modelling an A&E service unit using an M/M/m Queue .......... 31
  4.2.1 Number of servers when queue is always busy .......... 31
  4.2.2 Number of servers using M/M/m response time formula .. 32
4.3 Number of Jobs in an M/M/m queue .................................. 34
4.4 Distribution of Queuing and Response Time ..................... 37
  4.4.1 Queuing Time Distribution .................................. 37
  4.4.2 Response Time Distribution .................................. 38
4.5 Variations of Response Time distribution ......................... 41
  4.5.1 Response time distribution when arrival rate is increased .. 42
  4.5.2 Response time distribution when arrival rate is decreased ... 42
4.6 Application of laplace transforms to find queuing time distribution 43
4.7 Using Laplace Transforms to find the response time distribution .. 45
4.8 Summary ................................................................. 48

5 Modelling with a network of queues

5.1 Modelling the A&E service unit as simple network of queues .....50
  5.1.1 Explanation of Patient Flow in the simple network ........ 52
5.2 Response Time .......................................................... 53
  5.2.1 Number of jobs in the system ................................ 53
  5.2.2 Arrival rates for the system ................................ 53
  5.2.3 Response Time of the simple network ....................... 54
  5.2.4 Response time of simple network for varying arrival rates 55
5.3 Bottlenecks in the simple network .................................. 57
  5.3.1 Varying the number of servers at tests ..................... 58
  5.3.2 Varying the service rate at tests .......................... 60
5.4 Other bottlenecks in the system .................................... 62
  5.4.1 Varying the number of servers at Majors Doctor ............ 63
5.5 Effect of increase in the external arrival rate .................... 65
5.6 Response time distribution of paths in the simple network ...... 66
5.7 Mean response time for simple network .......................... 71
  5.7.1 Verifying mean response time using Little’s law .......... 71
  5.7.2 Approximating the model ..................................... 72
5.8 Conclusion for the simple network of the A&E service unit ......72
5.9 Adding the specialist doctor ......................................... 72
5.10 Bottlenecks in the extended network .............................. 75
  5.10.1 Proportion of majors & minors entering increased to 50% 75
  5.10.2 Proportion of majors increased to 55% minors still 25% ...76
5.11 Varying the number of servers at specialist doctor ............ 76
  5.11.1 Varying the service rate at specialist doctor ............. 78
5.12 Bottleneck Analysis at Majors Nurse .............................. 81
5.13 Response time distributions of the extended network ..........82
  5.13.1 Verifying response time using Little’s law .............. 84
5.14 Adding cycles to the extended network ........................... 85
6 Evaluation

6.1 Data Analysis ................................................................. 89
6.2 M/M/1 Queue ................................................................. 90
6.3 M/M/m Queue ............................................................... 90
6.4 Network of queues ......................................................... 90
6.4.1 Simple network of queues ........................................... 90
6.4.2 Extended network of queues ....................................... 91
6.5 Response time distributions ............................................ 92
6.6 Future Work ................................................................. 93

7 Conclusion

A Appendix
Chapter 1

Introduction

1.1 Motivation

Over the last two decades or so, we have seen a massive improvement in the overall speed at which systems work whether it is telecommunications processing, manufacturing equipment or even serving patients in a hospital A&E service unit. Although at the same time, the population of the world has increased and the ever changing demands of people have meant that their expectations of systems are far more than what is currently being given. As a result, it has become a constant rat-race to make our current systems faster. This brings about questions such as how do we measure such improvements? Is there a standard procedure? Are there any other alternatives in measuring such improvements?

In today’s world, a significant amount of research is taking place to make systems that provide life critical support to humans work faster. For example, in the UK itself, the Government has introduced performance specific targets for the NHS which state that 98% of the patients that enter an A&E service unit, should be treated in less than 4 hours (the response time). This value of 98% is known as the quantile of response time and they provide us with a good measurement of measuring the success of the system rather than the mean response time. This is because even though the mean response time could well have been low, this could have been achieved by the fact that some patients were treated very quickly and whilst the others had long response times.

As a result, the question now lies in the fact as to how we should achieve such targets. The first thing that comes into mind will be to increase the number of doctors and the speed of the equipment in the A&E service unit. This would be possible if we had unlimited amount of resources but, since this isn’t feasible in terms of the A&E, we would have to look at alternative ways of reaching these targets with the limited resources that the A&E has. In essence, the A&E could vary their limited number of staff at different departments according to the arrival rate and see what happens to the overall response time. However, this will prove to be time consuming and very expensive for the A&E and hence isn’t feasible.
This motivates us to find accurate models that will help us simulate and predict the behaviour of the A&E service unit. Various studies such as discrete event simulations, statistical modelling etc have been done in this area of performance analysis with all of them broadly based upon queuing theory. Since its inception by A.K Erlang in 1909, this has been the basis for modelling many different systems. As a result, we could use the various queuing models from the very basic M/M/1 queue upto a more complex network of M/M/m queues in order to model the A&E service unit. If modelled accurately, not only will such a model give managers an insight into optimising their resources, but will also show them which departments are bottlenecks when the arrival rate increases. Therefore, if they decided on making a change to a particular department, they could simply use the model and see what effect that this change will have on the response time of the system instead of trying it live and having to face the consequences of it failing if the decision was bad.

1.2 Objectives

Broadly speaking, this project looks into the application of different queuing models to model patient flow in the A&E service unit. Therefore, one of the first things that I intend on doing is carrying out a data analysis of patient arrivals at the A&E service unit of a major London hospital.

This data analysis will entail of:

1) Finding Arrival and service rates of patients in the A&E. This will enable us to parameterise our queuing models in a way as to simulate the current behaviour of the A&E service unit. Using the arrival and service times, I intend to plot a graph of the distribution of inter-arrival time and service time of patients. This is necessary as later on, when using the queues to model the A&E, we need to prove that the inter-arrival time and service time can be approximated as being exponentially distributed.

2) Finding response time and response time distributions of patients. This will give us a good idea as to how long a patient takes on average in order to be treated. Furthermore, variations of response times during the day and week too will be studied. This will give us a good insight as to when response times are high/low and to see if any patterns occur.

3) Classifying the patients as Majors/Minors and see their response time and response time distributions in order to compare them and see whether they have similar distributions to that of all patients.

4) Calculating the proportion of patients that are treated within the 4 hour target limit and see whether the 98% quantile is achieved.

5) Calculating the proportion of patients that died whilst in the A&E or left before being treated. This will give us a good idea of the quality of service of the A&E. If this proportion is large, it will suggest to us that the A&E isn’t treating its patients quickly enough and has long queuing times.
CHAPTER 1. INTRODUCTION

By this point, we would have a good idea of the values for the parameters of the queuing model, i.e. – the arrival rate, service time and response time. As a result, I firstly intend on modelling the A&E using the M/M/1 queue. The M/M/1 queue requires the inter-arrival time and service time of patients to follow an exponential distribution. Using the distributions of inter-arrival time and service time plotted in the data analysis section, we can see whether this is true and if so, try and model the A&E service unit using the M/M/1 queue. If it isn’t feasible to do so, I intend to calculate the service rate that the server (staff) should be working at for the steady state of the system to be preserved.

The second queue that I intend to study is the M/M/m queue. The M/M/m queue is an extension to the existing M/M/1 queue with the difference being that there are m servers in an M/M/m queue instead of only 1 with the M/M/1. This queue too requires that the inter-arrival time and service time follows an exponential distribution such that the Markov property can be applied to the queue. If this is true, then some would argue that we could have a very large number of servers to model this A&E system. However, since we have limited resources, I intend on finding the minimum number of servers that could be used in order to model such a system.

Using this value of m, I also intend on studying the queuing and response time distributions of such a queue. The response time distribution in particular, will be derived using a convolution of the queuing time and the service time. Not only will such an analysis give us an insight into the probability of having to queue for a certain amount of time, but using the response time distributions, the management can see how probable a certain response time is when the arrival rate is varied.

Even though we should be able to model the A&E using an M/M/m queue, it does seem to have some constraints. For example, it assumes that all the servers work at the same rate and are identical. However, as we know, all staff in an A&E are most certainly not going to work at the same rate. i.e – receptionist’s probably work faster than doctors and furthermore staff are most certainly different (nurses, doctors, receptionists etc). This motivates us into trying to model the A&E in other ways.

Hence, in this report I also intend on investigating the effect of modelling the A&E using a network of M/M/m queues. This overcomes the problem with the M/M/m queue as each type of staff (nurse, receptionist, doctor etc) can be modelled using a single M/M/m queue having its own service rate. Furthermore, it also helps us to understand patient flow in an A&E service unit much better since it allows us to estimate the proportion of patients going into each queue which isn’t possible using an M/M/m queue. Once I have parameterised the queuing network model to approximate it to the real patient data, I intend on looking at the bottlenecks in such a system as the arrival rate varies.

Bottlenecks are nodes in the system with the greatest utilisation. As a result, any increase in the arrival rate will cause the load at the bottleneck to increase greatly and may even prevent the steady state of the system from existing (i.e – number of jobs in the system tends towards infinity). By varying the parameters at the bottleneck (i.e – the number of servers, service rate etc) we can see the effect that this would have on the overall response time of the system when the arrival rate is increased/decreased. Such a bottleneck analysis will help the management of the A&E service unit
optimise their resource planning. For example, if we plotted a graph of the response time of the system as the arrival rate varied for different number of servers at the bottleneck, we could observe the maximum arrival rates for each value of \( m \). Furthermore, we could also see whether there are any marked differences in the response time when the number of servers is increased for a particular arrival rate. If this decrease in response time was small and not crucial, then the management need not add this extra staff at this node, hence saving useful resources.

I also intend on looking at the effect on the total response time of the system when an extra node (i.e – the specialist doctor) is added. This will help us find out as to whether this model (extended network) is a better representation of the A&E service unit. Furthermore, I also intend to investigate as to whether the bottlenecks in such a system change in comparison to the simple network and what happens when external arrival rate of the system is varied.

Finally, using the Markovian network queuing model, I intend to study response time distributions for the different patient flow paths through the system. As before, we could simply find the response time distribution at each node (using convolutions) and then find the convolution of the response time distributions between each node in the path in order to find the response time distribution of a path. However, since Laplace transforms determine the probability density functions uniquely, I intend on studying the response time distributions of such paths using Laplace Transforms.
1.3 Report Structure

Chapter 2 looks into the background of this project. It shows some of the queues and methods of queuing theory that I will be using.

Chapter 3 looks into the data analysis I have performed on patient arrival data in the A&E service unit of a major London hospital. The arrival rate, service rate and response times are all calculated along with their distribution graphs. Response times of Majors/Minors too are found along with their distributions in order to see if any inferences could be made of them. The chapter then looks to see whether the target of 98% of patients have actually been treated within 4 hours of entering the A&E service unit.

Using the results and distributions found in chapter 3, Chapter 4 looks into applying the M/M/1 and M/M/m queue to the A&E service unit. In the case of the M/M/m queue, the number of servers that would be used in such a system is calculated and this is verified using different methods. Furthermore, the chapter also looks at the queuing and response time distributions of a system using an M/M/m queue and discusses what happens to these distributions when the arrival rate is varied. Lastly, this chapter introduces the concept of Laplace transforms and shows how they can be used to calculate queuing and response time distributions within the M/M/n queue as well as within a network of queues (in Chapter 5).

Chapter 5 looks into modelling the A&E service unit as a Markovian network of queues. It firstly models the A&E using a simple network of queues calculating bottlenecks of such queues. Once a node is found to be a bottleneck, it does a bottleneck analysis at this node, causing the parameters to be varied in order to see the effect it will have on the response time of the whole model. Furthermore, using Laplace Transforms, response time distributions of paths within this network are calculated and plotted. This chapter then introduces the specialist doctor to the model (extended network) and looks into the effect that this has on the response time of the system. Here too, bottlenecks are found and at each bottleneck, a bottleneck analysis is done. Lastly, this chapter looks into the response time distributions of this extended network calculated using Laplace Transforms.

In chapter 6, the success of the project is evaluated along with areas where further work can be done. Finally, chapter 7 gives a conclusion of the project.
Chapter 2

Background

This chapter discusses the different methods that can be used when modelling patient flow in an A&E service unit. It also discusses some of the concepts of queuing theory and notation that I will be using in the project.

As this project is based around modelling patient flow in an A&E, one of the first things we need to look at is how the A&E actually works. Given below is a summary of the A&E service unit.

2.1 Background to an Accident and Emergency Service Unit

(Note: A lot of the information in this section has been obtained using [1] as well as conversations with my professor Peter Harrison and secondary marker William Knottenbelt.)

An accident and emergency service unit is a part of a hospital where patients come in with regard to the assessment and treatment of serious injuries and emergency treatment. Often referred to as ‘casualty’ departments, they are opened around the clock, around the year. According to government targets it is a requirement that a patient seeking service at an accident and emergency unit be attended to with four hours of registration at the reception. This is a very short time limit, especially during peak hours (late at night, weekends) and a split second can mean the difference between life and death. Hence the need for a competent queuing system that will enable maximum efficiency and time management, which is the prime focus of my project.

Research shows that patient arrivals can be categorised into 2 different categories:

1) Walk in: These patients are the ones that arrive at the A&E via their own transport.
2) Ambulance Arrivals: These patients arrive in the ambulance from the scene of an accident or if the ambulance has picked them up from the point of calling.

Once these patients arrive, each of them will go through a series of different steps in order to be treated according to their form of arrival.
1) Walk-in Patients

Once they enter the A&E, they register themselves with the receptionist who routes them to either one of three queues: minors queue, majors queue or nurse assessment depending on whether they have a minor injury/illness, major injury/illness or one that cannot be detected at that moment but is a suspect case of a minor illness. This initial filtering process is one derived from war times. On the battlefield, patients were categorised according to one criterion – if they were treated would they be able to hold a gun again. This process, referred to as ‘triage’ has now been refined and is the first step in the accident and emergency unit and determines the route taken by a patient on registering at accident and emergency.

2) Ambulance Arrivals

Patients arriving in an ambulance are firstly assessed by a nurse. Depending on the result of the assessment, the nurse either decides to send the patient to the reception so that they can be registered or to the majors queue.

Given below is a short description of what happens to each patient as they are routed to one of these three queues:

i) Minors Queue

Once a patient has entered a minors queue, he/she will have to wait for a cubicle to become free before being seen by the minors practitioner (doctor/nurse). After assessing the patient, the practitioner can then decide to do one of the following:

- Treat the patient and discharge to home or alternatively refer to the GP or pharmacy to pick up medication
- Request for a series of diagnostic tests to further evaluate the patient’s condition. Eg: blood tests, diagnostic imaging (x-ray, MRI)
- Refer the patient to a specialist doctor to perform a more thorough investigation
- Admit the patient to a surgical ward or to the MAU (Medical Assessment Unit).

ii) Majors Queue

Once a patient enters a majors queue, the majors nurse performs tests (x-ray, blood tests etc) on the patient whose results are then given to the majors doctor. Depending on the outcome of the tests and assessment, the majors doctor may then decide to:

- Refer the patient to a specialist doctor for further investigation
- Perform more tests
- Discharge the patient to their home, GP or pharmacy to pick up medication.

Note: If the patient’s health suddenly deteriorates, the patient is sent to the resuscitation bay.
iii) Nurse Assessment

Once an assessment room becomes free, the nurse assesses the type of injury/illness to a patient. According to this, the nurse then sends the patient to the minors queue, majors queue or discharges the patient to their home, GP or pharmacy to pick up medication.

Tests

The tests for both majors and minors are done in this same laboratory.

Specialist Doctor

A specialist doctor focuses in depth on a specific area of medicine contrary to a GP who does not specialise in a particular field but has an overall knowledge of medicine. The specialist doctors that a minor might have to visit are ENT, Gynaecologists or Orthopaedics. Majors, on the other hand are seen by a wider range of medical and surgical specialists.

As we can see, there are many paths of patient flow through the A&E service unit depending on a number of factors. For example: type of arrival, type of injury/illness, result of tests etc. Hence, it is a very difficult task to model it accurately. Given below are some of the methods that we could use in order to model such a system.
2.2 Methods used to model patient flow in an A&E service unit

Patient flow in an A&E service unit is an example of a system with discrete state spaces. Hence, we require our models to be stochastic transition systems with delays at each state. There are two main methods which were considered in order to model the patient flow through the A&E service unit. They are: (discrete event) simulation and analytical queuing theory. As with most methods, they have their advantages and disadvantages.

A simulation is a way of modelling a system by mimicking its behaviour. In essence, simulation uses randomly generated values to simulate the events of a system. Due to this randomness, results of simulations can vary and hence isn’t accurate at all times. Although, the effects of randomness of such results can be measured using statistical methods such as variance and confidence intervals, it means that a large number of computations need to be done in order to validate such results. In general, simulations require a large amount of data to mimic the behaviour of the system which in turn causes the computational time of such simulations to be very large. As a result, simulations aren’t as efficient (in general) as using analytical methods to model a system but can be used in order to verify some of the results obtained using analytical methods.

Analytical methods on the other hand, make use of formulas derived from mathematical analysis in order to calculate results. As an A&E service unit is effectively made up of a number of queues, we can use queuing theory to model the A&E service unit and make use of analytical methods to calculate results for such queues. Even though we would have to make assumptions in order to parameterise such queues, using analytical methods, results can be verified to model the actual system. In general, analytical methods are quick and accurate ways of modelling a system although, at times they can be infeasible when there are complex dependencies in the system. As a result, I decided to use analytical queuing theory to model the A&E service unit.

2.3 History of Queuing Theory

Queuing theory dates back to 1909 when a Danish mathematician known as A.K Erlang published his first work - "The Theory of Probabilities and Telephone Conversations". Whilst working for the Copenhagen telephone company, Erlang tried to find ways of calculating the number of circuits that were required to provide an acceptable telephone service. He realised that probability theory could be used to do so, and went on to find the number of operators that were required to handle a number of telephone calls. By doing so, this helped the telephone company assess the number of circuits required as well as the number of operators for a particular arrival rate.

Since Erlang’s first work was published, queuing theory has been applied to many other areas such as telecommunications, transport systems, scheduling jobs in a computer and even modelling patient flow in hospital systems.
2.4 Background to Queuing Theory

When modelling the A&E service unit using queuing theory, there are many different types of queues that we could use. Each queue is distinguished according to certain characteristics it may have. For eg: a queue may have m servers in comparison to a queue which has 1 server. In order to standardise these characteristics, David G.Kendall introduced an A/B/C queuing notation in 1953.

Each type of queue is represented as follows:

\[ A / S / M / C \]

where:

1) **A** - nature of the arrival process characterised by the inter-arrival time distribution

2) **S** - Probability distribution of the service time

3) **M** - number of servers in the queue

4) **C** – capacity of the queue. If this is omitted, \( c = \text{infinity} \)

In most cases, \( c \) is omitted and as a result, only three letters are used to represent the queue – \( A/S/m \). The probability distribution of the inter-arrival times and service times can be represented by:

**M** - Abbreviation for the inter-arrival time/service time to have the Markovian (memoryless) property. Since the exponential distribution has the Markov property, any queue whose inter-arrival/service time is an exponential distribution is said to have the Markov property. Furthermore, if its inter-arrival time fits an exponential distribution, the queue is said to have a Poisson distribution for its arrivals. The Markovian property is quite a useful property since random independent streams of requests seem to be Markovian.

**G** - The inter-arrival time and service time have a general distribution and as a result, any probability distribution could fit into such a model.

**D** - The inter-arrival and service time have a deterministic distribution and hence are constant.

**E_k** - The inter-arrival and service times have an Erlang distribution with \( k \) as the shape parameter.
In addition to these, there are also two other characteristics that can be used to
distinguish queues. They are:

1) The way jobs in a queue are served:

Most queues implement the FIFO (First-In-First-Out) procedure. i.e –
customers are treated according to the times at which they arrived. However,
there are other variations too, such as LIFO (Last-In-First-Out) which states
that the last customer that arrived is the first one to be served as well as SIRO
(Service in Random Order).

2) Size of population from which the customers arrive. This limits the arrival
rate.

As we can see, there are various queues that can be used to model the A&E service
unit using Kendall’s notation. Given below are some of the queues that we could use:

1) M/M/1 – the most basic queue and hence, simplest to analyse. This queue is a
FIFO (First-In-First-Out) queue and has patients arriving in batches of various
sizes. Therefore, patients are treated in order of their arrivals one after the
other. Here the inter-arrival times and service times follow an exponential
distribution. However, this queue has the capacity of only one server. i.e –
there is only one member of staff in the whole A&E and hence, only one
patient can be processed/dealt with at a time. Therefore the second patient in
line will have to wait till the first patient has been discharged from the
accident and emergency service unit to be served.

2) M/M/n – this is an extension of the M/M/1 queue with the difference being
that this has m servers instead of 1. Here too, the inter-arrival time and service
time follow an exponential distribution since it has the Markov property.

3) M/G/1 – this queue has its inter-arrival times following an exponential
distribution, but its service times are distributed generally. Since patients have
varying service times, it could be that the service time might have a general
distribution and hence, such a queue could be used to model the A&E service
unit. If more than 1 server is required, we could use the M/G/m queue.

4) G/M/1 – this queue has a general inter-arrival time distribution and an
exponential service time distribution. The fact that the inter-arrival time
distribution can be general is a helpful feature of this queue, although in our
model since it’s most likely that there might be processor sharing (i.e –
equipment might be shares between 2 patients’ tests), the M/G/1 queue is
preferred over this queue.

5) G/G/1 – this queue has a general inter-arrival time and service time
distribution. Hence, if we could model the A&E service unit using such a
queue, we wouldn’t have to worry about having to fit a specific distribution to
its inter-arrival and service times. In essence, due to its generality, one could
model any system using such a queue but since almost impossible to solve its
equations, we won’t consider modelling the A&E service unit using such a queue.

Given the large choice of queues which we could use in order to model the A&E service unit, since the M/M/1 is the simplest of them all, I have decided to firstly study the effect of modelling the A&E service unit using such a queue. Due to its limitations with the number of servers (m=1), I also study the effect of modelling the A&E service unit using the M/M/m queue.

2.5 Useful Formulas

In order to model the A&E service unit using the queues above, one of the first things we will have to do is to prove that the inter-arrival time and service time follow an exponential distribution. This enables us to verify that these queues have the Markovian property. We then need to look at the service and response times of such queues in order to see whether they are feasible. (For example, if we were using an M/M/1 queue and it had a service time of 50 minutes, then this wouldn’t be feasible as this effectively means that only after 50 minutes is it that the new patient can be looked into and hence, the queue will only grow larger as the time increases.) In the case of the A&E service unit, the service time is the time taken to treat the patient whereas the response time is total time taken for a patient from when he/she enters the A&E service unit to when he/she leaves it.

Although there are variations as to how the service time and response time of different queues are calculated, there are some general formulas we could use. One such formula is Little’s Law.

Little Law states:

\[ L = \lambda w \]

Where \( L \) = number of jobs in the system, \( \lambda \) = arrival rate of the system, \( w \) = response time of the system. i.e – the response time of the system is proportional to the mean number of jobs in the system. Hence, such a formula could be applied easily to a simple queue such as an M/M/1 queue but is slightly more complicated when applying to an M/M/m queue (as the number of jobs in the system is different). This is derived in chapter 4.

Note: The number of jobs in the system is sometimes also referred to as the mean queue length. The derivation of this formula is discussed in detail in [5].

2.6 Steady State Theorem

In order for us to apply any sort of analytical methods on a queue, we need to ensure that the steady state of the system holds. In essence the steady state means that the system has been running for some time and has settled down causing its behaviour to no longer exhibit any trends. The steady state of a system is sometimes also referred to as the statistical equilibrium of the system.
This theorem brings us to the utilisation law which has to hold for the stability of the system to be preserved. The utilisation law states that the total offered load on the system does not exceed the number of servers. Since the total load on a system is calculated as the proportion of the arrival rate to the service rate, using the utilisation law, it can be written that:

\[ \frac{\lambda}{\mu} \leq m \]  

(2.1)

Simplifying this equation, we get:

\[ U \leq 1 \]  

(2.2)

Where

\[ U = \frac{\lambda}{m\mu} \]  

(2.3)

Therefore, we can see that for any system to be stable, its utilisation has to be less than or equal to one. This property is used greatly throughout this project especially in chapter 5, when dealing with networks of queues in order to ensure that the steady state of each queue is preserved when a parameter of the system is changed. For example, when the external arrival rate of the system is changed, we need to ensure that steady state property is maintained for each M/M/m queue.

2.7 Network of Queues

Even though we can model the A&E service unit using an M/M/m queue, it has 2 major problems with it. Firstly, it assumes that all the staff in the A&E service unit are identical and hence work at the same service rate. In essence, this isn’t true since we have different kinds of staff (receptionists, nurses, doctors etc) all working at different service rates. Secondly, it assumes that all the staff are working perfectly in parallel. This too may not be the case in our A&E service unit since staff may well have to share resources and hence, can only work after the shared resource has freed up.

This causes us to look at different ways of modelling the A&E service unit. As we can see from the background of the A&E service unit, (Section 2.1) it is effectively a network of queues with each patient having to go from one node to another according to his/her treatment. Hence, in this report I study the effect of modelling the A&E service unit using a network of queues. By using such a model, not only can we have each node in the network represent a type of staff but also have individual service rates for the different kinds of staff.

2.7.1 How a Network of queues works

In essence, a queuing network can be thought of as a connected directed graph whose nodes (in the queuing system) serve the purpose of service centres. The arcs between the nodes are used to represent paths through which a job can move from node i to node j with a routeing probability of \( q_{ij} \) regardless of its history. The routeing probability gives us the proportion of jobs that move from node i to node j.
In general, there are two main features which we use in order to distinguish different queuing networks:

1) The shape of the underlying graph
2) The nature of the job population

The shape of the underlying graph is sometimes also referred to as the topology of the graph and there are three different topologies – open, closed and mixed. An open network is one in which no job can be trapped inside the system and not be allowed to leave it. i.e – there is at least one path through which a job can enter and leave the system. On the other hand, a closed network is one in which jobs circulate inside the system forever and hence there isn’t a path through which a job can leave the system. A system which has characteristics of both is said to be a mixed system.

Our A&E service unit has to be modelled using an open network since there never is a case of when a patient enters and circulates in the system forever. i.e – the patient finally leaves the A&E service unit through one of the channels described in the Section 2.1.

2.7.2 Type of queues used

As stated above, the queuing network consists of nodes which act as service centres. Hence, when modelling the A&E service unit, we need to decide on the type of queues we are going to use in our queuing network. There are two queues that we could look at:

1) M/G/m – This queue is used only in special circumstances in a network. For example, if there is a sharing of service (i.e – processor sharing) or Last-Come-First-Serve basis, it would be efficient to use this queue. However, when there is processor sharing, it means that the way the jobs in the system are treated changes, and is no more a First-Come-First-Serve basis. Although in our A&E service unit, there might be processor sharing of equipment, it causes the queuing discipline to change (as stated above) which might not be the way the queue should really work. Furthermore, if we were to model our A&E service unit using a network of M/G/m queues, we would find it very difficult to get response time distributions analytically. Since this project studies response time distributions, we would have to look at other ways of modelling the A&E service unit.

2) M/M/m – This queue is preferred over the M/G/m since it follows the First-Come-First-Serve basis and enables us to get accurate response time distributions of paths within the queuing network. Hence, I have decided to use M/M/m queues in order to model this queuing network.

Before we can obtain any sort of analytical methods on the queuing network, we need to look at a very important theorem in queuing networks called Jackson’s Theorem.
2.7.3 Jackson’s Theorem

If we wanted to find the solution to this network, i.e – find the probability that there are $n_1$ jobs at node 1, $n_2$ jobs at node 2, $n_3$ jobs at node 3…$n_n$ jobs at node N, then we would have to use Jackson’s theorem.

Jackson’s Theorem states that:

*If the arrival rates obtained from the traffic equations are such that $\rho_i < 1$ for all $i = 1, 2, \ldots N$ then steady state exists and*

$$p(n_1, n_2, \ldots, n_N) = \prod_{i=1}^{N} (1 - \rho_i) \rho_i^{n_i}$$ (2.4)

The fact that the joint distribution of $n_1, n_2, \ldots n_n$ can be written as a product of $N$ modified geometric distributions implies the following 2 statements:

1) The marginal distribution of the number of jobs at node $i$ is the same as for an isolated M/M/1 queue.
2) The number of jobs present at any moment at node $i$ is independent of the number of jobs at any of the other nodes in the network.

These two statements are particularly useful as it enables us to calculate many of the results that are used in queuing theory. For example, using the fact that the number of jobs at any node is independent of the number of jobs at any other node, we could simply add up all the jobs at each node in the network to find the total number of jobs in the system. Then, to find the response time of the network, we could use Little’s law and hence divide the total number of jobs in the system by the external arrival rate of the system.

As we can see, using Jackson’s theorem, we can find the response time of the queuing network. The analytical methods used to find the number of jobs and arrival rate at each node is derived in chapter 5.

2.7.4 Bottlenecks in the System

In this project we also study the effect of varying the parameters of a bottleneck. A bottleneck is the node(s) in the queuing network that has the highest utilisation. Hence, when performing a bottleneck analysis for the nodes, we will have to find the utilisation values for each node (using the utilisation formula derived above) and ensure that this value never exceeds 1 in order to preserve the steady state of the system. If it does, then we will have to look at varying the parameters of this node namely the service rate and the number of servers and see the effect that this change has on the overall response time of the system.
(Note: In general, with our queuing network, we vary the arrival rate from a small number to a very large number to see how the bottleneck moves from one node to another)

2.8 Response Time Distributions

In this report, I also study queuing and response time distributions of the M/M/n queue and the queuing network. These distributions show us the probability of a particular value \( t \) being the queuing/response time.

For example, if we plotted a graph of the cumulative frequency distribution of the queuing time, at any point \( t \), we would get the probability of a patient having to queue for a time less than or equal to \( t \). Hence, if a patient had minor injury and had access to these distributions, he/she could see the probability of having to queue for this less than or equal to this time and looking at this, could decide if it was worth going to the A&E unit to get his injury looked into.

In terms of the queuing time distribution, we simply need to differentiate the cumulative distribution function to get the probability density function. This derivation is shown in section 4.4.1.

However, when looking at response time distributions, since we need to find the probability density function of 2 independent variables (i.e – the queuing and service time) we need to use the method of convolutions. (The basics to this are derived below and applied to the system in chapter 4)

The method of convolution simply states that:

If we had the probability distribution function of 2 independent variables \( X \) and \( Y \) and we needed to find the distribution of the sum \( S \) of these 2 variables, then:

For \( S \) to take a value \( i \), \( X \) must take the values \( 0, 1, \ldots, i \) and \( Y \) must take the value \( i - X \). Using this proposition, we can write:

\[
P(X + Y = i) = \sum_{j=0}^{i} P(X = j, Y = i - j) = \sum_{j=0}^{i} p_j q_{i-j}
\]

(2.5)

Since we are working with continuous probability density function using the same variables above we would get:

\[
f_X(i) = \int_{j=0}^{i} f_X(j)f_Y(i - j)
\]

(2.6)

(2.7)

Hence, we can use this result to find the convolution of the distribution of the queuing and service time which is equal to the response time distribution. This project also looks into finding response time distributions of paths within the queuing network.
Hence, using the method shown above, we would have to find the convolutions of response time distribution between nodes in each path. If there are a lot of nodes in the path, it would require us to perform many integrals. This calculation is simplified considerably by the introduction of Laplace transforms.

### 2.9 Laplace Transforms

The Laplace transform of a probability function $f(x)$ is defined as:

$$f^*(s) = \int_0^\infty e^{-sx} f(x)dx$$

(2.9)

This transform enables the probability density function $f(x)$ to be determined uniquely. Hence, it follows that the Laplace transform of the sum of two independent variables is equal to the product of their Laplace transforms. i.e – if variable $X$ has a Laplace transform $f^*(s)$ and variable $Y$ has Laplace transform $g^*(s)$ and $S$, the sum of $X$ and $Y$ has Laplace transform $h^*(s)$ then:

$$h^*(s) = f^*(s)g^*(s)$$

(2.10)

Therefore, using the equation above, the response time distribution of a path can be found easily. All we would have to do is to find the Laplace transform for the response time distribution at a node and then multiply it by the number of nodes in the paths. (with each Laplace transform having its unique parameters of each node). This way, if we decide to extend a path, we could simply multiply the previous Laplace transform by the Laplace transform of the new node’s response time distribution.
Chapter 3

Data Analysis

As stated in the introduction section, one of objectives of this project was to carry out an analysis on patient data in an A&E service unit. I was very fortunate to have been able to get hold of one month’s worth of A&E patient arrival data for January 2005. This data consisted of 8458 records of individual patients, each patient having data such as his/her arrival time, kind of treatment, service times and discharge time. As the data was presented to me in Excel, it enabled me to write VB macros in order to calculate things such as the number of patients who had response times >240 minutes etc.

In this section, I will find arrival times, service times and response time variations of patients enabling me to use different kinds of queues based on the results found. I will first carry out an analysis on all patients arriving into an A&E and then break this down into the 2 classes of patients classified - Majors and Minors.

3.1 Arrivals

One of the first things that could be done when given such data about an A&E service unit is to find the arrival rate of patients. Since the unit time for the data given has been given in minutes, I have used minutes as my unit time.

The arrival rate was calculated using one day’s worth of data. The number of arrivals in a day was calculated to be 266 Patients. As there are 1440 Minutes in a day, the arrival rate is therefore:

\[
\frac{266}{1440} = 0.184722 \text{ Patients/Minute}
\]

This suggests that an average of 11.08 patients arrive every hour.

3.1.1 Inter-arrival time

In order to get a good understanding of the distribution of arrivals at an A&E service unit, we need to look at the inter-arrival time. The inter-arrival time is simply the time taken between successive arrivals. Furthermore, when we need to apply queues in
order to model the A&E service unit, we will have to prove that the inter-arrival time is exponential. As an exponential distribution has the Markov property, it will enable us to say that the arrivals have the Markov property. Given below is a histogram showing the inter-arrival time distribution.

From the histogram, we can see that the inter-arrival time for a large number of patients is very small and there are very few cases where the inter-arrival time is large. (i.e – very few instances where there is a long delay between 2 patients arriving). This seems to suggest that the distribution of the inter-arrival time seems near enough to make it a reasonable approximation of an exponential distribution and thus the arrivals can be said to have the Markov property.

### 3.2 Service Time

The service time is typically the treatment time of the patient. Here again, the average service time of a patient is simply found by dividing the total service time by the number of patients served. (Since a fair amount of the data had missing values in some of its fields, I was only able to calculate the average service time for 3222 patients). The total service time was 236022 Minutes and 3222 Patients were served in this time. This gave us an average service time of:

\[
\frac{236022}{3222} = 73.2532 \text{ Minutes}
\]

This gives us an average service rate of:

\[
\frac{1}{73.2532} = 0.01365 \text{ Patients/Minute}
\]
In order to get a better understanding of the service time distribution of the patients, given below is a histogram:

From the histogram, we can see that the service time of most patients is fairly low and a few of them have slightly high service times – denoted by the long tail in the histogram. This histogram suggests that the service time distribution is near enough to make it a reasonable approximation to state that it is exponentially distributed. This is an important finding as we will later see when using queues that a lot of the queues require the service time to be exponentially distributed.

3.3 Response Time

The response time is also known as the waiting time. This in effect is the time taken from when the patient arrived at the A&E service unit to when he/she was discharged. Once more, the average response time could be found by simply taking the total response time and dividing it by the number of patients seen. The total response time of 8458 patients was found to be 1157158 Minutes giving us an average response time of:

$$\frac{1157158}{8458} = 136.8122 \text{ Minutes}$$

This suggests that on average, each patient is treated in slightly over 2 hours. Given below is the graph showing the response time of each patient from the data collected:
From this graph, we can see that most of the patients have been treated with response times less than the critical response time of 4 hours (240 minutes). We can also see that the graph shows a few large spikes which suggests that some patients have very large response times. Each of these large spikes is followed immediately by another large spike. This is because if one patient gets delayed at a particular node in the system, then, the patient following it will also be delayed. This is brought back to normalcy by the ‘breach buster’ nurse. The breach buster nurse is a member of staff who simply helps out whenever there is a case of a patient whose response time goes over 240 minutes. By doing so, the number of staff increases and as a result, patient’s service time decreases. These large spikes will be of primary concern to the A&E and they should look into record details of such patients and see what could be done in order to bring down these large response times. This graph doesn’t show us the distribution of the response time. In order to do so, we would have to plot a histogram. Given below is what the histogram of response time of patients looks like:
CHAPTER 3. DATA ANALYSIS

The histogram suggests that a large amount of the patients have response times below the critical time of 240 minutes (4 hours) whereas a few still have large response times (indicated by the long tail). Furthermore, if we carefully look at the critical value, (at response time of 240) we can see a large sudden spike and then a sudden drop in the frequency which probably suggests that the A&E service unit is desperately trying to get the response time of most patients under 240 minutes. Fitting a distribution to this histogram would be very difficult although it is possible to state that it loosely fits a general distribution.

3.3.1 Response Time of Patients for One-Day

In order to further understand the variation in response times of patients, it might help in order to see the response time distribution of patients during one-day. Given below is this graph:

The graph suggests that there is no definite pattern of response times during the day, i.e – the response times of patients is not small or large in the early hours of the day in comparison to the latter part of the day. Hence suggestions about what the A&E service unit should do in order to decrease the response time during different parts of the day can’t be made accurately.

3.3.2 Response Time of Patients for one week

Even though it’s not possible to make accurate predictions of when response time will be high/low during the day, by looking at a graph that shows us the response time during one week, we will be able to see if there are any patterns of response time during that week. Given below is this graph:
The graph suggests that the average response time of patients go up significantly during mid-week (i.e – on Wednesday and Thursday) in comparison to the end of week. (i.e –Friday, Saturday and Sunday). This is as expected, because on Wednesday’s, 2 of the staff go for training. As a result, the A&E service unit is short of staff and thus, service times of patients are longer, which in turn increases their response times (queue lengths increase too as a result of the shortage of staff).

Using such graphs, the A&E service unit could look into increasing the number of staff serving on Wednesdays/Thursdays or even trying to look into ways of increasing the speed at which a customer is served (For example, by making their equipment faster). These will bring down the queue lengths significantly, thus bringing down response times.

3.4 Response Times of Different Classes of Patients

As stated earlier, once a patient arrives, he/she is classified as either a major or a minor depending upon the type of injury. One particular analysis of response times is to break down the arrival of each patient into the relevant category and find the response times of these two categories. Firstly, let’s look at the Minors.

3.4.1 Response Times of Minors

From the data given, there were 5430 patients who were classified as minors. The total response time of these patients was 647593 Minutes giving us an average response time of:

\[
\frac{647593}{5430} = 119.2621 \text{ Minutes}
\]

This suggests that each minor, on average, has a response time of approximately 2 hours – slightly less than the average response time of any patient – (136.2 minutes). This is rightfully so as minors have less injury to them and so, the service (treatment)
time is lower. Given below is a graph showing the response time of each minor as they arrive:

As we can see, the response time of minors is very small and there are very few large spikes which suggest that only a small number of the minors had large response times. Given below is a histogram which shows us the response time distribution of minors.

Like with the response time of all the patients, we can see that there is a sudden spike around the critical value of 240 minutes which again suggests that the A&E service unit is trying very hard in order to bring down the average response time of minors to less than the critical value. Looking at the histogram, it is very hard to fit a distribution, although we can state that the minor’s response time loosely fits a general distribution.
3.4.2 Response Times of Majors

The number of patients that were classified as majors from the data collected was 3028. The total response time of all majors was 509565 Minutes giving us average response time of:

\[
\frac{509565}{3028} = 168.2843 \text{ Minutes}
\]

This suggests that on average, each major is treated in slightly under 3 hours. Since they require more treatment time, their average response time is more that that of the minors. Given below is the response time of each major as they arrive into the A&E service unit.

Here, the histogram suggests that most majors are treated under the critical value of 240 minutes, but we see far more “large spikes” than found in the profile of minors. This is probably because majors require more attention – i.e – longer treatment times and hence, there is a greater chance of having a larger response time. Given below is the histogram showing the response time distribution of majors:
CHAPTER 3. DATA ANALYSIS

This histogram very closely matches the response time distribution of minors as well as the average response time of all patients in general. The only difference between the majors’ and minors’ response time distribution is that there seem to be far more majors whose response time is greater than the critical response time of 240 minutes as explained above.

3.5 Quantile of response times

As stated in the introduction section, one of key objectives of an A&E service unit is to make sure that 98% of its patients are treated in a response time of less than 240 minutes. Therefore, in order to see whether this was true, I wrote a small VB Macro that enabled me to count the number of patients whose response time was less than or equal to 240 minutes. This gave me a value of 7870 Patients. Therefore the proportion of patients:

\[
\frac{7870}{8458} = 93.05\%
\]

As we can see, the value of 98% of patients that have to be treated within a response time of 240 is not achieved as there are still approximately 5% more of those patients who have response times of more than 240 minutes. The A&E service unit should look into this matter and see where patients are spending the most times and take necessary action accordingly. i.e – if they realised that a particular test was taking a long time and as a result, patients had to wait for a result before they could see a doctor, they could look into increasing the number of people helping out with the test or even look into the possibility of getting equipment that would produce the results faster. They could also see whether any parallel processing could be done, by having two patients possibly looked into at the same time without having to wait to get the results of a particular patient.
3.6 Treatment Profiles of Other Types of Patients

A good way of measuring the quality of service of an A&E service unit is by seeing the proportion of patients who weren’t given the service that they required. This could be for eg: the number of patients who left without being seen, the number who left without being given treatment or even the number who died whilst in the A&E service unit. This was found pretty easily using the data given as each patient had a discharge code attached to him/her. This code simply signified the way the patient left the hospital. Hence if he went home, he simply had a code called “HOME” etc. Given below is a small analysis that finds the proportion of patients that fall under the category described above:

3.6.1 Proportion of patients that left without being seen

These patients were classified with the code “LEAB”. There were 163 patients with this code, hence giving us a proportion of:

\[
\frac{163}{8458} = 1.9272\%
\]

As we can see, there are a significant proportion of patients who have arrived at the A&E service unit and left without being seen. This is probably due to the fact that the queue lengths at the reception (where you have to register) were very long at the time of joining and the patients didn’t want to spend so much time waiting to be seen. On the other hand, most patients would have spent some time in the queue and then felt that it might take way longer for them to be seen and as a result would have left. Thus, in order for this proportion to decrease, the A&E service unit should look into ways of reducing the queue length. Suggestions like having more people serve patients at the reception counter might help with this.

3.6.2 Proportion of patients that left without being given treatment

These patients were classified with the code “LEAB”. There were 152 patients with this code, hence giving us a proportion of:

\[
\frac{152}{8458} = 1.7977\%
\]

These patients are ones who have seen a doctor who might have recommended some treatment. However, it could possibly have been that the queuing time to receive treatment might have been really long and as a result patients might have left.

3.6.3 Proportion of patients that died whilst in the A&E Service Unit

From the data given, these patients have been classified as “DID” (Died in Department) or “MORT” (patients sent to the mortuary). There was only one instance of a patient dying in the department and 24 instances of patients being sent to the mortuary. This gives us a proportion of
\[ \frac{25}{8548} = 0.00292 \% \]

As we can see, only a small proportion of patients actually die whilst in the A&E service unit which could suggest that the doctors are of high standard and are able to save a lot of the patient’s lives.
Chapter 4

Applying Queues to an A&E Service Unit

As stated in the introduction section, one of the objectives of this project is to find credible queuing models that can be used to represent an A&E service unit. In order to do this, we would have to match known queues to the A&E service unit, and try to find the models that fit the results we have found in the previous chapter the best. For example, if the service rate of a particular system was low and its inter-arrival time and service time had exponential distribution, then we could suggest that it fits an M/M/1 model.

This chapter firstly looks into trying to fit the A&E service unit to 2 queues - the M/M/1 queue and the M/M/m. It then suggests ways of trying to approximate the number of servers (m) there are in the M/M/m queue and verifies this value using different methods. Furthermore, it discusses the queuing time and response time distributions when an M/M/m queue is used and also looks at what happens to the response time distribution when the arrival rate is varied. Lastly, it introduces Laplace transforms and shows how such transforms can be applied to find response time distributions

4.1 Modelling an A&E service unit using an M/M/1 Queue

When trying to model any system with a queue, one of the simplest queues to start off with would be to use an M/M/1 queue. This queue follows the FIFO principle and has one server serving its customers. For this queue to be used, the inter-arrival time and service time of the system has to be proven to have an exponential distribution so that the system can be assumed to have the Markov property (memoryless).

From the graph showing the patient’s inter-arrival time (Section 3.1.1), we can see there are more patients with short inter-arrival times than long inter-arrival times. This seems to suggest that it is near enough to make an approximation that the inter-arrival times are exponentially distributed. The same can be said about the service time graph in section 3.2. Hence, we can go ahead and try modelling our system with an M/M/1 queue.

The second step we would have to look at when trying to use an M/M/1 queue would be to see whether the observed service time is feasible. From the data given, the
service time has been calculated as 73.2532 Minutes/Patient. (refer section 3.2 for calculation) As we can see, this is ridiculously high for an M/M/1 queue as it suggests that every patient (if modelled using an M/M/1 queue) would take 73.2532 minutes to be served (since there is only one server) and as a result, only after this time will it be that the next patient will be served. This means that a steady state will never be reached and the queue will grow infinitely long.

If we wanted to use an M/M/1 model, then we could calculate the service rate (and in effect the service time) the server should work at in order for a steady state to be achieved using the arrival and response time values we have got. Given below is this service rate:

The queue length for an M/M/1 queue is defined as:

\[ L = \frac{\rho}{1 - \rho} \]  \hspace{1cm} (4.1)

Using Little’s Law, \( L = \lambda w \)

Therefore:

\[ w = \frac{\rho}{\lambda (1 - \rho)} \]

\[ w = \frac{1}{\mu (1 - \frac{\lambda}{\mu})} \]

\[ w = \frac{1}{\mu - \lambda} \] \hspace{1cm} (4.2)

Re-arranging the formula so we find a value for \( \mu \):

\[ \mu = \frac{1}{w} + \lambda \] \hspace{1cm} (4.3)

Using the values of \( w = 136.8122 \) and \( \lambda = 0.184722 \)

\[ \mu = 0.19203 \text{ Patients/Minute} \]

Since:

\[ \text{Service Time} = \frac{1}{\mu} \]

\[ \text{Service Time} = 5.207 \text{ Minutes/Patient} \]

This means that if an M/M/1 queue were to be used, the service time of each patient should be 5.207 minutes and the server should work at a speed of 0.19203 Patients/Minute. Only then will a steady state be reached.
As we can see, this value of approximately 5 minutes is too low for the service time and furthermore it assumes that there is only 1 person working in the whole A&E service unit! As we know that this isn’t true and there are several staff working in an A&E service unit, this queue can’t be used to model the A&E service unit. With several staff working in an A&E, individual service times increase causing the service rate to decrease. As a result, this gives us the same combined service rate which leads us into having to look at other queues – particularly the M/M/m queue.

4.2 Modelling an A&E service unit using an M/M/m Queue

The M/M/n queue is an extension of the M/M/1 queue as effectively, the only difference is the number of servers that are used to service the patients (in this case m). As with the M/M/1 queue, the M/M/m queue requires the inter-arrival time and the service times to follow an exponential distribution. Since we have already proved in the section above, we could say that this is true. We then need to look into the number of servers that are used in order to model this system. Since we don’t have a value for this, I have decided to approximate it using the following:

\[
\text{The number of servers (staff) working at any given time} = \frac{\text{Total service time of all patients}}{\text{Time period over which the patients were served}}
\]

From the data, the total service time taken to serve 3222 patients was 236022 minutes. This took a time of 11.8 days (11.8 * 24 * 60 = 17003.232 minutes). As a result, the number of servers:

\[
\frac{236022}{17003.232} = 13.88 \Rightarrow 14 \text{ servers}
\]

As we can see, this is a very rough approximation to calculating the servers. Shown in the next two sections are ways we can calculate the number of servers more accurately.

4.2.1 Number of servers when queue is always busy

For any steady state M/M/m queue, according to the utilisation law, (refer section 2.6), the utilisation of the queue has to be \( \leq 1 \). Taking the worst case where the queue is always busy, i.e – utilisation = 1

Then:

\[
U = \frac{\lambda}{m \mu} = 1
\]

Using values of \( \lambda = 0.184722 \) and \( \mu = 0.01365128 \):

\[
m = 13.53 \text{ servers} \Rightarrow 14 \text{ servers}
\]

This means that there has to be at least 13.53 servers in order for there to be a steady state. Any number of servers less than this value will make the utilisation value rise above 1 and hence the steady state property won’t be preserved.
In fact, using the calculation of 14 for the number of servers, gives an utilisation of 0.96 which means that the queue is busy at almost all times. Therefore, we can say that using 14 servers to model this A&E service unit as an M/M/14 queue is credible.

4.2.2 Number of servers using M/M/m response time formula

Another good way of checking our value of 14 servers for the M/M/m queue is by using the basic M/M/m response time formula. In essence, what we have to do is use the values of \( \lambda \) and \( W \) from the data analysis carried out and by varying \( m \), we can find the service rate for each \( m \) value. This will give us a good approximation of the number of servers (\( m \)) required in order for the A&E service unit to function at the service rate of 0.01365 Patients/Minute calculated in section 3.2. Given below is the derivation of the response time formula of an M/M/m queue.

For any system of queues: Response Time = Mean queuing time + Service time

The service time is \( \frac{1}{\mu} \) whereas the queuing time requires slightly more calculation.

The mean queuing time in an M/M/m queue is calculated as follows:

\[
P ( Q_1 < t ) = F_q (t) = 1 - q e^{-(m \mu - \lambda) t}
\]

where \( q \) is the probability that the job has to wait. i.e – it has to queue.

Therefore:

\[
q = 1 - (p_0 + p_1 + p_2 + \ldots + p_{n-1})
\]

When: \((i \leq m)\),

\[
p_i = \frac{\lambda^i}{i! \mu^i} p_0
\]

and,

\[
p_0 = \frac{1}{\sum_{i=0}^{m-1} \frac{\lambda^i}{i! \mu^i} + \frac{\rho^m}{(m-1)! (m-\rho)}}
\]

Therefore,

\[
q = 1 - \frac{\sum_{i=0}^{m-1} \frac{\lambda^i}{i! \mu^i}}{\sum_{i=0}^{m-1} \frac{\lambda^i}{i! \mu^i} + \frac{\rho^m}{(m-1)! (m-\rho)}}
\]

This is known as Erlang’s delay formula.
Now, the mean queuing time can be written as:

\[
\text{Queuing Time} = \frac{q}{m \mu - \lambda}
\]  (4.10)

Inserting this into the response time formula gives:

\[
w = \frac{1}{\mu} + \frac{q}{m \mu - \lambda}
\]  (4.11)

Re-arranging the formula to give \( \mu \) in terms of \( W \) and \( \lambda \) (so that service rate can be found):

\[
w = \frac{m \mu - \lambda + \mu q}{\mu (m \mu - \lambda)}
\]

\[
w = \frac{\mu (m + q - \frac{\lambda}{\mu})}{\mu (m \mu - \lambda)}
\]

when \( \rho = \frac{\lambda}{\mu} \):

\[
w = \frac{m + q - \rho}{m \mu - \lambda}
\]

\[m\mu = \frac{m+q-\rho}{w} + \lambda\]

Therefore:

\[
\mu = \frac{m+q-\rho + \lambda w}{mw}
\]  (4.12)

As we can see, given above is the formula for finding the service rate in terms of \( m \), \( q \), \( \lambda \) and \( w \). The formula for \( q \) requires an initial value for the service rate, but since this is the value we want to find, we would have to approximate it. A sensible approximation was to divide the service rate for an M/M/1 queue by the number of servers (\( m \)).

Using \( w = 136.8122 \) Minutes and \( \lambda = 0.184722 \) Patients/Minute (as calculated in chapter 3), the formulas were input into Mathematica in order to obtain the result of \( \mu \) for each corresponding value of \( m \). For every \( m \) value, the result of \( \mu \) that was output was fed back into the new value of \( \mu \) and the process repeated until the values converged. i.e – both sides of the equation shown above balanced.

The results obtained are shown below:
As we can see from the results, as the number of servers double, the service rate is approximately halved. For example, from 4 – 8 servers, the service rate halves from 0.048 to 0.024. This is rightfully so as now, since there are more servers, each server works slower. In order to understand this better, we could use the utilisation formula where

$$U = \frac{\lambda}{m \mu}$$

Since the arrival rate remains constant, if $m$ (the number of servers) increases, then $\mu$ has to decrease. Furthermore, if we look at the service rate for when $m=14$, we can see that it is very close to the service rate of 0.01365 Patients/Minute that we have calculated in section 3.2. Hence, this proves that using an M/M/14 queue for the A&E service unit is very credible.

### 4.3 Number of jobs in an M/M/m queue

If we were to use Little’s law in order to calculate the response time of this system, we would have to find the number of jobs in the system and divide this by the external arrival rate. The number of jobs at any node/in a system is calculated as follows:

Let us denote the limiting probability of state $i$ by $p_i (i = 0,1,2...) \ i.e \ the \ probability \ that \ the \ system \ has \ i \ jobs \ in \ it$.

Therefore:

$$L = \sum_{i=1}^{\infty} p_i i$$ \hspace{1cm} (4.13)

The general form for $p_i$ is one of 2 cases:

1) when $i < m$; as shown in equation (4.7)

$$p_i = \frac{\lambda^i}{m!m^{i-m} \mu} p_0$$

where $p_0$ is shown in equation (4.8)
Therefore, when both these cases are combined, we get the general equation of:

\[ L = \sum_{i=1}^{m-1} \frac{\lambda^i}{i! \mu^i} p_0 i + \frac{m^m}{m!} p_0 \sum_{i=m}^{\infty} \left( \frac{\lambda}{m \mu} \right)^i \]  

(4.14)

As we can see, the first summation can easily be plugged into Mathematica and the value derived although if the second one is plugged into Mathematica, since the series doesn’t converge, Mathematica doesn’t give a result for the second summation. Hence we will have to derive this summation so that it can be input into Mathematica.

Given below is the derivation of this summation using the formula for a geometric progression series:

When \( i >= m \):

\[ L = \frac{m^m}{m!(x + y)} \sum_{i=m}^{\infty} \left( \frac{\lambda}{m \mu} \right)^i \]  

(4.15)

Where:

\[ p_0 = \frac{1}{x + y} \]

Let \( i = j + m \), Then the summation can be written as:

\[ = \sum_{j=0}^{\infty} \left( \frac{\lambda}{m \mu} \right)^{j+m} (j + m) \]

\[ = \sum_{j=0}^{\infty} \left( \frac{\lambda}{m \mu} \right)^i \left( \frac{\lambda}{m \mu} \right)^m (j + m) \]

Since, \( \left( \frac{\lambda}{m \mu} \right)^m \) is a constant:

\[ \left( \frac{\lambda}{m \mu} \right)^m \sum_{j=0}^{\infty} \left( \frac{\lambda}{m \mu} \right)^j (j + m) \]

\[ \left( \frac{\lambda}{m \mu} \right)^m \left[ \sum_{j=0}^{\infty} j \left( \frac{\lambda}{m \mu} \right)^j + m \sum_{j=0}^{\infty} \left( \frac{\lambda}{m \mu} \right)^j \right] \]

Let \( a = \frac{\lambda}{m \mu} \)

Therefore:

\[ = a^m \left[ \sum_{j=0}^{\infty} j(a)^j + m \sum_{j=0}^{\infty} a^j \right] \]  

(4.16)
Given below is the derivation of the first summation above:
The terms when the first summation is expanded are: \(0 + a + ax^2 + ax^3 + ax^4 + \ldots\)

This can be written as:

\[
= a(1 + 2a + 3a^2 + \ldots)
\]

\[
= a \left[ \frac{d}{da} \left( a + a^2 + a^3 + \ldots \right) \right]
\]

Using the formula for geometric progression, where \(\sum_{i=1}^{\infty} x^i = \frac{x}{1-x}\)

\[
= a \left[ \frac{dy}{dx} \left( \frac{a}{1-a} \right) \right]
\]

\[
= a \left( \frac{1}{1-a} + \frac{a}{(1-a)^2} \right)
\]

Simplifying the equation:

\[
= \frac{a}{(1-a)^2} \quad (4.17)
\]

The second derivation in equation (4.16) above is a simple geometric progression
where \(\sum_{j=0}^{\infty} a^j = \frac{1}{1-a}\). Hence the equation can now be written as:

\[
= a^m \left[ \frac{a}{(1-a)^2} \frac{m}{1-a} \right]
\]

\[
= a^m \left[ \frac{a+m(1-a)}{(1-a)^2} \right]
\]

\[
= a^m \left[ \frac{a + m - am}{(1-a)^2} \right]
\]

Therefore, when \(i \geq m\),

\[
L = \frac{m^m}{(x+y)m!} a^m \left( \frac{a + m - am}{(1-a)^2} \right) \quad (4.18)
\]

This formula can now be input into Mathematica. Hence, if this equation is added to
the first part of equation (4.14), we get the total number of jobs in this queue.
4.4 Distributions of Queuing and Response Time

Since we have a good idea of the number of servers in an A&E service unit, we can use this value in order to calculate the queuing time and response time distributions for an M/M/m queue. These distributions will give us an insight into how probable it would be to have a queuing/response time of value \( t \) and whether such times are within acceptable limits. This is particularly useful when say for example, a patient had a minor injury and he had access to these distribution graphs. If he was willing to queue for a maximum of \( x \) minutes, using the distribution graph, he could see how probable that \( P(\text{Queuing Time} \leq x) \) was and using this, could decide whether or not to go to the A&E service unit.

4.4.1 Queuing time Distribution

In order to see the probability that a patient has to queue for \( \leq t \) minutes, we would have to plot the cumulative frequency distribution graph. As stated above, in any M/M/m queue:

\[
P(\text{Queuing Time} \leq t) = F_q(t)
\]

where \( t \) is the queuing time

\[
F_q(t) = 1 - q e^{-(m \mu - \lambda) t}
\] (4.19)

Using the parameters of \( m=14, \lambda=0.184722, \mu=\frac{0.192031}{m} \) as obtained in chapter 3, the graph was plotted using equation (4.19) shown above and the result obtained as shown below:

Like all cumulative density functions, the larger the queuing time, the greater the chance that that the patient is served after time \( t \). Furthermore, if we look at the queuing time of approximately 800 minutes, the probability = 1. Since this is effectively the maximum value that the probability can be, it suggests that the
maximum queuing time is about 800 minutes and you wouldn’t have to wait any longer in order to be served.

If we differentiated the cumulative frequency density function shown in equation (4.19), we would get the probability density function of the queuing time distribution. This would show us how probable a particular queuing time was. Therefore:

\[
 f_q = \frac{dF}{dt} = q(m \mu - \lambda)e^{-(m \mu - \lambda)t} \quad (4.20)
\]

Using the parameters of \( M, \lambda, \mu \) as above, the graph was plotted and the result is shown below:

From the graph we can see that the probability of queuing for time \( t \) decreases greatly as the queuing time increases. i.e – there is a very high chance that the patient will have to queue for a short amount of time. Furthermore, we can see that at approximately a queuing time of 800 minutes, the probability of queuing is almost 0 which seems to suggest that a patient will almost never have to queue for such a long time.

4.4.2 Response time Distribution

Like with the queuing time, the response time distribution for the A&E service unit will show us the probability of being treated by a time \( t \). Although, as the response time is the summation of the queuing time and the service time, we will have to use the convolution of these 2 distributions in order to get the response time distribution. Shown below is the derivation for the response time distribution:
CHAPTER 4. APPLYING QUEUES TO AN A&E SERVICE UNIT

Since

\[ f_R = f_Q \ast f_S \]  

where \( f_X \) = probability density function of X

\& \( \ast \) represents convolution.

As the queuing time and service time are continuous random variables, we will have to integrate the product of their distributions in order to find the convolution of these 2 distributions. Furthermore, if the service time is distributed with parameter \( x \), then the queuing time is distributed with parameter \((t-x)\).

Therefore, the response time distribution can be written as:

\[ f_R (t) = \int_0^t f_Q (t-x) f_S (x) \, dx \]  
\[ \text{(4.21)} \]

As stated in equation (4.20) the queuing time distribution with parameter \((t-x)\) is:

\[ f_Q (t-x) = q (m \mu - \lambda) e^{-(m\mu - \lambda)(t-x)} \]  
\[ \text{(4.22)} \]

And the service time distribution with parameter \( x \) is:

\[ f_S (x) = \mu e^{-\mu x} \]  
\[ \text{(4.23)} \]

Therefore:

\[ f_R (t) = \int_0^t q (m \mu - \lambda) e^{-(m\mu - \lambda)(t-x)} \cdot \mu e^{-\mu x} \, dx \]

\[ = \int_0^t q \mu (m \mu - \lambda) e^{-(m\mu - \lambda)t} \cdot e^{x(m\mu - \lambda - \mu)} \, dx \]

\[ = q \mu (m \mu - \lambda) e^{-(m\mu - \lambda)t} \left[ e^{x(m\mu - \lambda)} - 1 \right] \]

\[ = \frac{q \mu (m \mu - \lambda)}{(m-1) \mu - \lambda} \left[ e^{-\mu t} - e^{-(m\mu - \lambda)t} \right] \]  
\[ \text{(4.24)} \]

As we can see, this formula can now be inputted into Mathematica with the parameters the same as the ones used for the queuing time distribution. (i.e- \( m = 14, \lambda = 0.184722, \mu = \frac{0.192031}{m} \)). Since Mathematica doesn’t converge for \( t = \infty \), \( t \) was set to 1400. The graph plotted for this response time distribution is shown below:
The graph shows us that there is a very high probability that the response time is approximately 150 minutes. This is rightfully so because the response time that we calculated in chapter 3 for the A&E service unit is 136.8122 Minutes/Patient and therefore, we would expect that the highest probability would be around this response time. The graph also shows us that as the response time increases after 200 minutes, the probability of the particular response time \( t \) decreases and there is a very small probability that the patient’s response time \( t > 800 \) minutes.

Using the utilisation formula \( U = \frac{\lambda}{m \mu} \), we get a value of 0.962 for the utilisation value of the A&E service unit, i.e – it is very busy. As a result, if we look at the probability of the response time being high, e.g: 400 minutes, we can see that this is a fairly high value whereas if we saw the response time distribution of a system whose utilisation was lower, we would expect this probability value to be lower. This is because, in essence, since such a system will be less utilised, there will be a greater chance that a patient in such a system will have a smaller response time. As a result, the whole peak of the graph of such a system will be shifted towards the left and will have a longer tail to show the probability of having a high response time is low. Furthermore, the narrowness of the curve of a lower utilised system is signified by the fact that the since the area under such a graph is 1, if the graph has a longer tail and accommodates a large amount of the area, then in essence, the curve at the beginning has to be pushed up and hence is narrower.
### 4.5 Variations of Response time distribution

From the equation derived above for the response time distribution, we can see that the denominator of the constant is \((m-1)\mu - \lambda\). Therefore, an interesting case is to see what happens to the response time distribution when \(\lambda = (m-1)\mu\). This causes the denominator to be equal to 0. Plugging in this value of \(\lambda\) to the response time distribution integral in equation (4.24) causes the nature of the formula to change and gives us:

\[
f_R(t) = \int_0^t q \mu (m \mu - (m-1)\mu) \left( e^{-(m \mu - (m-1)\mu)x} \cdot e^{x(m \mu - (m-1)\mu - \mu)} \right) dx
\]

Therefore:

\[
f_R(t) = \int_0^t q \mu^2 \left( e^{-\mu t} \cdot e^{x.0} \right) dx
\]

Plotting this graph using Mathematica, we get the following graph:

![Graph showing the response time distribution](image)

As we can see, the graph has the same shape as the general response time distribution we have plotted for the A&E system above, the only difference being the arrival rate of the system. This decrease in the arrival rate causes this utilisation of such a system to be 0.929 which is slightly lower than the utilisation of the graph shown above. Hence, as expected, the probability of the response time being high (e.g., 400 minutes) in this system is lower than the system above. Furthermore, we can see that the tail in this graph is longer than the graph above for the reason explained in the previous section.

As we can see, the arrival rate (which affects the utilisation) has an effect on the distribution of the response time. In order to see the patterns of various arrival rates more clearly, I have plotted some graphs below:
4.5.1 Response time Distribution when Arrival Rate is increased

Given below is a graph showing the response time distribution of an A&E service unit when the arrival rate is increased to \((m - 0.5)\mu\). In such a system, the utilisation is 0.964, which is slightly higher than the system showing the general response time graph. As a result, we would expect a slightly broader curve with a smaller tail than the general response time graph above:

![Probability Density Function](image)

As expected, the curve is slightly broader and the tail is slightly longer as the system is more utilised.

4.5.2 Response time Distribution when Arrival Rate is decreased

Given below is the graph showing the response time distribution of an A&E service unit when the arrival rate is decreased to \((m - 1.5)\mu\). In this system, the utilisation is 0.893 and hence, is less utilised than the general system. Due to the reasons given above, we would expect such a system to have a much narrower curve and a long tail.
As we can see, the response time distribution has a narrow curve and a long tail, as expected.

### 4.6 Application of Laplace Transforms to find Queuing Time Distribution

Even though we could use the method of convolutions described above to find the distribution of a sum of two independent variables, as stated in the background section, Laplace transforms enable us to find such distributions easily and also define the function \( f(x) \) uniquely. In this section we look at the application of Laplace transforms to find \( q \) shown in equation (4.20) which gives us the probability density function of having to queue.

For any M/M/m queue, 
\[
q = P(\text{Have to queue})
\]

Therefore:

With probability \( 1-q \),

\[
\text{Queuing time} = 0
\]

With probability \( q(1 - \rho)\rho^{i-1} \), queuing time comprises of \( i \geq 1 \) exponential delays with parameter \( m\mu \). Therefore, in order to find the queuing time of the \( i^{th} \) patient in the queue, we will have to find the Laplace transform of the service time distribution of server and of all patients in the queue upto the \( i^{th} \) patient.
The Laplace transform of the density function is of the form:

\[ s_i^*(\theta) s_2^*(\theta) \ldots s_i^*(\theta) \]

by the memoryless property where 

\[ S^*(\theta) \] is the Laplace Transform of the service time distribution of \( x \)

When:

\[ s(t) = m \mu e^{-m\mu t} \]

Then:

\[ s^*(\theta) = \frac{m \mu}{m \mu + \theta} \] (4.26)

Therefore:

\[ f_q^*(\theta) = \sum_{i=1}^{\infty} q(1 - \rho) \rho^{i-1} (s^*(\theta))^i + (1- q) \] (4.27)

\[ = q(1 - \rho) s^*(\theta) \sum_{i=1}^{\infty} (\rho s^*(\theta))^{i-1} \]

\[ = \frac{q(1 - \rho) s^*(\theta)}{1 - \rho s^*(\theta)} \]

\[ = \frac{q(m \mu - \lambda)}{(m \mu - \lambda) + \theta} \] (4.28)

Since \( \frac{\alpha}{\alpha + \theta} \) is the Laplace transform of \( \alpha e^{-\alpha t} \), the queuing time distribution function can now be written as:

\[ = q(m \mu - \lambda) e^{(m \mu - \lambda) t} \] (4.29)

As we can see, this is the same as what was shown in equation (4.20). This shows us that Laplace transforms are very useful for finding distributions between independent variables. Shown in the next section is how Laplace transforms are applied to find response time distributions.
4.7 Using Laplace Transforms to find the response time distribution

As stated in the introduction section, one of the objectives of this project is to find response time distributions of queues. Although, we could use the method of convolutions described above, we can also use Laplace transforms to do so. In particular, Laplace transforms are extremely helpful when we have to find response time distributions within paths of networks. If we used the convolution method, we would have $2^n$ integrals in this case which would be very tedious to compute if we had a large number of nodes in the system. Shown below is the Laplace transform of the response time distribution for an M/M/m queue.

In general,

\[ \text{Response Time} = \text{Queuing Time} + \text{Service Time} \]

Therefore, using Laplace Transforms:

\[ r^*(\theta) = q^*(\theta)s^*(\theta) \] \hspace{1cm} (4.30)

Hence, we need to find the Laplace transforms of the queuing and service time - \(q^*(\theta)\) and \(s^*(\theta)\) respectively.

Firstly, finding \(q^*(\theta)\):

\[ q(x) = q(m\mu - \lambda)e^{-(m\mu - \lambda)x} \]

Therefore:

\[ q^*(\theta) = \int_0^\infty e^{-\theta x} q(x) \]

\[ = \frac{q(m\mu - \lambda)}{(m\mu - \lambda) + \theta} + (1 - q) \] \hspace{1cm} (4.31)

Likewise, finding \(s^*(\theta)\)

\[ s(x) = \mu e^{-\mu x} \]

Therefore:

\[ s^*(\theta) = \int_0^\infty e^{-\theta x} s(x) \]

\[ = \frac{\mu}{\mu + \theta} \] \hspace{1cm} (4.32)

Using equation (4.30)

\[ r^*(\theta) = \left[ \frac{q(m\mu - \lambda)}{(m\mu - \lambda) + \theta} + (1 - q) \right] \frac{\mu}{\mu + \theta} \] \hspace{1cm} (4.33)
Hence, using this formula, we can find the response time distribution of an M/M/m queue. (The result of using this will be the same as using the convolution method described above.)

The next chapter describes a Markovian queuing network to model the A&E service unit. In order to find the response time distribution for each path in the system, we would need to find the response time distributions between each of the nodes in the path. In order to do so, using the result derived above in equation (4.33) and the Laplace transform theorem in section 2.9, we could simply find the product of the Laplace transforms at each node. i.e:

\[ r^* (\theta) = r_1^* (\theta) r_2^* (\theta) \ldots r_n^* (\theta) \]  

where \( n \) = number of nodes \hspace{1cm} (4.34)

As we can see, this method allows us to calculate the response time distribution of a path through the queuing network very easily. Furthermore, if we decided to add another node to the system, all we would have to do, is find the Laplace transform of the response time distribution at that node, and multiply it to the Laplace transform of the previous path (without the extra node).

In order to find the solution to \( r^* (\theta) \), we need to use the method of partial fractions in order to split up equation (4.34).

Hence, using equation (4.34) and equation (4.33):

\[ r^* (\theta) = r_1^* (\theta) r_2^* (\theta) \ldots r_n^* (\theta) = \prod_{i=1}^{n} \left[ \frac{q_i (m \mu_i - \lambda_i)}{m_i \mu_i - \lambda_i + \theta} + (1 - q_i) \frac{\mu_i}{\mu_i + \theta} \right] \hspace{1cm} (4.35) \]

Using the method of partial fractions, where:

\[ \frac{1}{(A + \theta)(B + \theta)} = \frac{1}{B - A} \left[ \frac{1}{A + \theta} - \frac{1}{B + \theta} \right] \]

And \( A_i = (m_i \mu_i - \lambda_i) \) and \( B = \mu_i \) we can write that:

\[ \prod_{i=1}^{n} \left[ q_i (A_i) \right] \frac{\mu_i}{A_i + \theta} + (1 - q_i) \frac{\mu_i}{\mu_i + \theta} = \sum_{j=1}^{n} \frac{\alpha_j}{\mu_j + \theta} + \frac{\beta_j}{A_j + \theta} \hspace{1cm} (4.36) \]

When we multiply both sides by: \( \prod_{i=1}^{n} (A_i + \theta)(\mu_i + \theta) \); The equation becomes:

\[ \prod_{i=1}^{n} \mu_i \prod_{j=1}^{n} [q_i (A_i) + (1 - q_i)(A_i + \theta)] = \alpha_i (A_i + \theta)(\mu_2 + \theta)(A_2 + \theta)\ldots + \beta_i (\mu_i + \theta)(\mu_2 + \theta)(A_2 + \theta)\ldots \hspace{1cm} (4.37) \]
The left hand side of the equation is simplified to:

\[ \prod_{i=1}^{n} \mu_i \prod_{i=1}^{n} (A_i - (1-q_i)\theta) \]

Hence, in order to solve this solution, we need to firstly find the constants \( \alpha_1, \beta_1, \alpha_2, \beta_2 \ldots \)

In order to do so, we will have to vary \( \theta \) so that everything, but the constant remains. Therefore, in order to find the \( \alpha \) terms, we can use the following equation:

Let \( \theta = -\mu_k \) : (where \( k = 1,2,3, \ldots n \) where \( n \) is the number of nodes)

Then, we can use the general equation where:

\[ \prod_{i=1}^{n} \mu_i \prod_{i=1}^{n} (A_i - (1-q_i)\mu_k) = \alpha_k \prod_{j \neq k}^{n} (\mu_j - \mu_k) \prod_{j=1}^{n} (A_j - \mu_k) \]

Therefore:

\[ \alpha_k = \prod_{i=1}^{n} \mu_i \prod_{i=1}^{n} (A_i - (1-q_i)\mu_k) \]

\[ \prod_{j \neq k}^{n} (\mu_j - \mu_k) \prod_{j=1}^{n} (A_j - \mu_k) \]  \hspace{1cm} (4.38)

Likewise, in order to find \( \beta_k \):

Let \( \theta = -A_k \)

Then:

\[ \prod_{i=1}^{n} \mu_i \prod_{i=1}^{n} (A_i - (1-q_i)A_k) = \beta_k \prod_{j=1}^{n} (\mu_j - A_k) \prod_{j \neq k}^{n} (A_j - A_k) \]

Therefore:

\[ \beta_k = \prod_{i=1}^{n} \mu_i \prod_{i=1}^{n} (A_i - (1-q_i)A_k) \]

\[ \prod_{j=1}^{n} (\mu_j - A_k) \prod_{j \neq k}^{n} (A_j - A_k) \]  \hspace{1cm} (4.39)

Therefore using these 2 formulas for \( \alpha \) and \( \beta \) we can find the constants. Now, in order to convert it into a probably density function \( f(x) \), we simply need to invert the Laplace Transforms. Inverting these Laplace Transforms gives us the following function:

\[ r(x) = \alpha_1 e^{-\mu_1 x} + \beta_1 e^{-A_1 x} + \alpha_2 e^{-\mu_2 x} + \beta_2 e^{-A_2 x} \ldots \]  \hspace{1cm} (4.40)

(Note: the constants are found for as many nodes there are in the system).
As we can see, by plotting a graph using the equation above, we can find the response time distribution of a path in a queuing network. Response time distributions for paths in the queuing network are found in the Chapter 5 and hence, I will keep referring to the derivations shown above in Chapter 5.

4.8 Summary

This chapter suggests that we use an M/M/14 queue in order to model our A&E service unit given the data that we have about each patient arriving at the A&E. Even though it is credible to model it using such a queue, there are 2 main problems with this:

1) All the staff in the A&E aren’t identical as suggested by the M/M/m queue. i.e – there are doctors, nurses and receptionists who each have different service rates.
2) All the staff aren’t working perfectly in parallel as suggested by the M/M/m queue.

This motivates us to model the A&E service unit differently taking these problems into consideration. A good way of doing so is by modelling it using a network of M/M/m queues. Using such a model, each queue can be used to represent a type of staff (eg: a receptionist, nurse etc) hence overcoming the problem of all the staff having the same service rate. (now, each type of staff can have a different service rate.) Furthermore, this also enables us to model patient flow in an A&E service unit more accurately. Chapter 5 looks into modelling this A&E service unit using a network of queues.
Chapter 5

Modelling with networks of queues

As stated in chapter 4, even though we model the A&E service unit using an M/M/1 queue, in reality, an A&E service unit is a network of queues. By this we mean a group of M/M/m queues which are linked together in order to show the patient flow in an A&E service unit.

This chapter firstly looks into modelling the A&E service unit as a very simple network of queues. i.e – parameterises the network in such a way that it can be used to approximate it to the A&E service unit. It then looks at the bottlenecks in this system – i.e - the nodes that seem to have the maximum utilisation and sees the effect of changing the number of servers/service rate at such nodes for different values of the arrival rate. This gives us a good indication of how many servers we would need when the arrival rate increases over a certain value in order for a steady state to exist. Furthermore, this will also provide us with an insight as to whether adding any extra servers or increasing the service rate at the bottleneck will have any significant effect to the overall response time. Response time distributions calculated using Laplace Transforms of paths in this network too is studied in this chapter and verified using Little’s Law.

We then move onto the case as to what happens when a specialist doctor is added into the network of queues. (The extended network) The specialist doctor is a doctor that any patient (major/minor) could see after he has done his/her test. Hence, only a proportion of patients coming out of ‘tests’ go to such a doctor. Here again, the response time and bottlenecks of such a network are found. As with the simple network (without the specialist doctor) a bottleneck analysis is carried out on each bottleneck, causing the bottleneck to move from one node to the other. We then look at the case of what would happen if there were to be a sudden increase in the arrival rate. eg – a terrorist attack and how the A&E service unit should cope with such a problem – i.e – where they should add more doctors/where the service rate should increase etc. Here too, response time distributions of paths within the extended network are studied and verified using Little’s Law.
5.1 Modelling the A&E service unit as a simple network of queues

In order for us to make any calculations in a simple network of queues, one of the first things we will have to do is to draw out the network of queues. This in effect should be a simplified version of how an accident service unit actually works i.e – the way the different staff are inter-linked with each other/ Number of staff in each department etc. Shown on the next page is a simplified version of the A&E service unit.
Figure 5.1:

- Reception
  - Majors - Nurse
    - M1 $\mu_1$
    - M2 $\mu_2$
    - M3 $\mu_3$
  - Minors - Nurse
    - M6 $\mu_6$
    - M5 $\mu_5$

- Tests
  - 1.0
    - Majors - Doctor
      - M4 $\mu_4$
    - Minors - Doctor
      - 0.8

- EXIT
  - 0.2
    - 2/17
  - 0.6
    - 15/17
  - $\lambda$
**Code:** The diagram below signifies that whenever a node is drawn (in the figure above) as shown on the left, it effectively means that there are m servers working in parallel as shown on the right.

![Diagram](image)

5.1.1 Explanation of Patient Flow in the Simple Network (Fig. 5.1)

Looking at Figure 1, we can see that the A&E service unit has been simplified such that there are only 5 nodes – i.e – 5 queues which are inter-linked. Furthermore, each node has been assigned as “M1/M2...” along with its corresponding service rate $\mu_1, \mu_2...$. The numbers simply signify the node value, i.e – node 1 is receptionist, node 2 is minors-nurse etc and the M and $\mu$ values are the number of servers/service rate of each node.

As shown in the figure, one of the first things that a patient does when he/she enters an A&E is to register with the reception. The Arrival rate of the patient is symbolised with the $\lambda$ symbol. Once the patient has registered with the receptionist, according to the injury (major/minor), he/she is sent to the majors/minors nurse. The values alongside each arrow correspond to the proportion of patients that go to that particular node. Hence, when the arrow from receptionist to minors nurse has the value of 0.6, it simply means that 60% of the total patients that arrive at the A&E are classified as minors and hence have to visit the minors nurse. Therefore, the arrival rate into any node in the network is calculated as: Proportion of Patients * Arrival rate. Using the minors as an example, the arrival rate into minors nurse will therefore be $0.6 \lambda$. Once the patient has visited the corresponding nurse, he/she will have to move onto visiting the next department. In the case of majors, all the majors have to go for tests (signified by the proportion = 1) whereas with minors, only 0.2 of the minors go for tests and the remaining 0.8 go to visit the minors doctor. If the patient goes to tests, then he/she will then have to go to his/her respective doctor to show results of the test. (Note: the proportion of patients leaving tests for minors and majors is the same proportion of minors and majors that arrived at tests). Once this is done, the patient exits the A&E service unit.

(Note: values for M1, M2....and $\mu_1, \mu_2...$ haven’t been given since it changes. For example, if the node was a bottleneck, its M values would change in order to see the effect that it would have on the total response time of the network.)
5.2 Response Time

In order to test whether or not our network of queues is suitable to model the A&E service unit, we have to calculate the response time of such a model and compare it to the response time we have obtained in section 3.3. In order to do so, we would have to use little’s law which states that \( L = \lambda w \)

Therefore, by dividing \( L \) by the external arrival rate, we can find the response time of the system. As we can see, the response time of any system, can be varied by simply varying the arrival rate and the number of jobs in the system, which in turn is varied by the number of servers and the service rate of the servers. Hence, by varying these parameters, any model can be tweaked in order to fit the response time we obtained in section 3.3. Shown below is the way we calculate \( L \) and \( \lambda \).

5.2.1 Number of jobs in the system

The average number of jobs in the system is found by summing up the average number of jobs over all the nodes:

\[
L = \sum_{i=1}^{N} L_i
\]  

(5.1)

Since all the nodes are M/M/m queues, we can use the equation (4.18) to calculate the number of jobs at each node.

5.2.2 Arrival Rates for the system

When using little’s law to calculate the response time of the entire network, we simply use the external arrival rate of the patients as our value for \( \lambda \).

If we look at the formula for calculating \( L \) at each node, we can see that it requires a value for \( \lambda \). This value for \( \lambda \) isn’t the external arrival rate and in fact requires us to calculate traffic rates for that node. Shown below is the way we calculate traffic rates:

Suppose we had a steady network with \( N \) nodes. Let \( \lambda_i \) be the arrival rate at node \( i \). (i.e – no. of jobs arriving at node \( i \) per unit time). Then, any job arrival at node \( i \) could be an external arrival or an internal arrival. External arrivals are classified as arrivals coming from out of the system and have an arrival rate of \( \gamma_i \). Internal arrivals are classified as arrivals coming from within the system i.e – from other nodes in the system. Hence, if \( \lambda_j \) jobs leave node \( j \) per unit time and a fraction of these \( q_{ji} \) go to node \( i \), the arrival rate into node \( i \) from node \( j \) is \( \lambda_j q_{ji} \). Since node \( j \) can be any node in the network, the total internal arrival rate is the summation of arrival rates from all the nodes in the network. i.e

\[
\sum_{j=1}^{N} \lambda_j q_{ji}
\]  

(5.2)
Since the total arrival rate at each node is the summation of the external and internal traffic rates:

\[ \lambda_i = \gamma_i + \sum_{j=1}^{N} \lambda_j q_{ji} \quad \text{where } i = 1, 2, \ldots, N \]  

(5.3)

This is known as the ‘traffic equation’ for the node. Hence, if a particular network had cycles (i.e. had feedback from nodes), using this we can find the arrival rates at each node in the network. However, since our network is an open feed forward network and there are no cycles, we won’t need to use the traffic equations as such. Therefore, the arrival into each node \( i \) will be as shown in equation (5.2) where \( j \) is a node in the network and doesn’t form a cycle with node \( i \).

### 5.2.3 Response Time of the Simple Network

As we have derived the formula for the number of jobs in a system, we can now proceed and apply it to our model of the A&E service unit. As our model has 6 nodes, in order to find the number of jobs in the whole system, we would have to find the number of jobs at each node and then sum these values up. This number can then be divided by the arrival rate at the receptionist in order to give the total response time of the A&E service unit. Shown in the table below are the initial values of the number of servers and the service rate at each node in the network. Furthermore, the traffic rates (arrival rates) at each node are calculated too, using the formula shown above:

<table>
<thead>
<tr>
<th>Number of Servers (M)</th>
<th>Receptionist</th>
<th>Minors Nurse</th>
<th>Majors Nurse</th>
<th>Tests</th>
<th>Minors Doctor</th>
<th>Majors Doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Service Rate ( \mu ) (Patients/Minute)</th>
<th>12</th>
<th>6</th>
<th>4</th>
<th>2</th>
<th>3</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traffic Rate ( \lambda )</td>
<td>0.4 ( \lambda )</td>
<td>0.6 ( \lambda )</td>
<td>(((0.6<em>1)+(0.4</em>0.2)) = 0.68 ( \lambda )</td>
<td>0.4 ( \lambda )</td>
<td>0.6 ( \lambda )</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: Parameters and Traffic Rates at each node

Using the parameters above, the traffic rates and \( \lambda = 0.184722 \) Patients/minute (as obtained for the A&E service unit in section 3.1) the formula to calculate \( L \) for each node was inputted into Mathematica. The results are shown below:

<table>
<thead>
<tr>
<th>Number of Jobs</th>
<th>Receptionist</th>
<th>Minors Nurse</th>
<th>Majors Nurse</th>
<th>Tests</th>
<th>Minors-Doctor</th>
<th>Majors Doctor</th>
<th>Whole System</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.17398</td>
<td>0.752762</td>
<td>2.03304</td>
<td>17.9955</td>
<td>1.51952</td>
<td>6.54561</td>
<td>30.0204</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2: Number of Jobs at each node
(Note: Arrival rate into each node is calculated as the proportion of patients*arrival rate to A&E service unit.). As we can see, the total number of jobs in the A&E system is 30.0204 giving us a response time of:

\[
\frac{30.0204}{0.184722} = 162.517 \text{ Minutes}
\]

As we can see, this value is very close to the value of 136.122 minutes/patient that we got as the response time in section 3.3 for our A&E service unit. Hence, the parameters of M and \( \mu \) we have chosen for each node in our model seem to be good approximations. As stated earlier, this model can easily be tweaked by varying the parameters for each node in order to get a response time closer to the one calculated in section 3.3 eg, as we can see, the majors doctor seems to have approximately 6.5 jobs in its system. In comparison to the other nodes (apart from tests), this seems to be fairly high. Hence, if we were to add another doctor to this node and make the total number of doctors serving this queue as 5, we see that the number of jobs goes down to 3.98. As a result, the total number of jobs in the system decreases to 27.4548 giving us a response time of 148.627 Minutes/patient. In such ways, we can match the model almost perfectly to the A&E service unit.

### 5.2.4 Response Time of Simple Network for Varying Arrival Rates

As shown above, we have only calculated the number of jobs in the system and the response time of the model when \( \lambda = 0.184722 \) Patients/Minute. In order to see the performance of the system, we would have to calculate the number of jobs and response times over varying arrival rates. Using the same method as shown above, the number of jobs in the model was found for varying arrival rates and plotted in a graph as shown below:

![Graph showing the relationship between arrival rate and number of jobs in the system](image)

From the graph we can see that as the external arrival rate increases, the number of jobs in the system increases. Furthermore, after approximately \( \lambda = 0.19 \) Patients/minute, the number of jobs begins to increase rapidly as it moves towards
infinity. This seems to suggest that there is an asymptote slightly after this value. In fact, calculations of utilisation at the nodes in the network seem to show that at $\lambda = 0.196$ Patients/Minute, node ‘tests’ begins to saturate and its utilisation = 1. Hence, queue length at this value is very large which in turn causes the queue length of the system to be very large. Therefore, this graph enables us to see that if we were to have an arrival rate of $\lambda \geq 0.196$ Patients/minute, our system wouldn’t be able to cope with this arrival rate and we have to look into ways of bringing down the queue lengths at ‘tests’. i.e. – add more servers/increase the service rate at ‘tests’.

Shown below is how the response time of the system varies with the arrival rate:

![Response Time Graph](image)

Since $L \propto W$, the shape of the response time graph is the same as the graph showing the number of jobs in the system varying with the arrival rate. Therefore, like with the previous graph, at $\lambda = 0.1960$ Patients/minute, there is an asymptote. This causes the response time to increase rapidly after $\lambda = 0.19$ Patients/Minute as ‘tests’ becomes saturated. Furthermore, if we look at the response time of the system at $\lambda = 0.184722$ Patients/minute, we can see that it is approximately 160 minutes – as calculated in the previous section.

As we have an idea of what happens to the performance of the system when it reaches a certain arrival rate, we can now proceed and see what could be done in such circumstances.
5.3 Bottlenecks in the Simple Network

When looking at increasing the performance of any system, one of the first things that is looked into, are the bottlenecks in the system. Bottlenecks are nodes that have the greatest utilisation. i.e – the greatest load. Once these nodes have been identified, changes to them could be made in order to see whether there is any difference to the overall performance of the system. eg: in the case of the A&E service unit, if we found out that the utilisation of the “test” node was the greatest, we could look into increasing the number of servers/service rate of tests. Once these changes are made, we could calculate the response time of the whole system again and see whether there is any significant change in the overall performance of the system.

In order to calculate the utilisation for our system, we need to use the utilisation formula where:

\[ U = \frac{\lambda}{m\mu} \]

Using the same parameters as given above, shown in the table below are the utilisation values for each node:

<table>
<thead>
<tr>
<th></th>
<th>Receptionist</th>
<th>Minors Nurse</th>
<th>Majors Nurse</th>
<th>Tests</th>
<th>Minors Doctor</th>
<th>Majors Doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilisation</td>
<td>0.461805</td>
<td>0.246296</td>
<td>0.554166</td>
<td>0.9420825</td>
<td>0.369445</td>
<td>0.83125</td>
</tr>
</tbody>
</table>

Table 5.3: Utilisations at each node

(Note: In order for a steady state to exist, \( U \leq 1 \).)

As we can see, “tests” seems to be the bottleneck in this system due to its large utilisation value. Since the utilisation value is so high, any increase in the arrival rate over a certain value, will cause “test” to get saturated and its utilisation rise above 1. This will mean that there won’t be a steady state any more and for the steady state to be preserved, we will have to add more servers to “test” to bring down its utilisation value. As a result, we will have to perform a bottleneck analysis on this node and see what happens to the overall response time of the system when its parameters are varied.
5.3.1 Varying the Number of Servers at Tests

One of the parameters that we could vary at tests is the number of servers (M). At the moment, “tests” has 4 servers. This section looks into the effect of what happens to the overall response time when the arrival rate is varied for M = 1,2,4 or 8. In order to make comparisons clearly, a graph is plotted.

Let’s take the case when M = 1:

Since we need the steady state of the system to be preserved, U <= 1. When U=1, the maximum arrival rate into the system can be found using the formula:

\[ U = \frac{\lambda}{m\mu} \]

Therefore,

\[ \frac{0.68 \lambda}{1 \times \left(\frac{2}{60}\right)} = 1 \]

Giving:

\[ \lambda = 0.04901 \]

(Note: The arrival rate (traffic rate) is 0.68 \( \lambda \) as shown in table 5.1)

Hence, after this value, “tests” becomes saturated and as a result, another server needs to be added to it to bring down its utilisation value and preserve its steady state. Therefore, when plotting the graph, the range of values for which we can vary the arrival rate is from 0.01 – 0.049. In essence, \( \lambda = 0.04901 \) is the asymptote of the graph.

The varying arrival rates were plugged into Mathematica for each node (keeping all the parameters the same except tests where M=1) and its queue length recorded for the corresponding arrival rate. This was then tabulated using Excel in the form below:

<table>
<thead>
<tr>
<th>Arrival Rate</th>
<th>Receptionist</th>
<th>Minors Nurse</th>
<th>Majors Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0040</td>
<td>0.0900</td>
<td>0.2527</td>
<td>0.0800</td>
<td>0.1800</td>
<td>0.6567</td>
<td>65.6717</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0080</td>
<td>0.1801</td>
<td>0.6892</td>
<td>0.1600</td>
<td>0.3601</td>
<td>1.4975</td>
<td>74.8775</td>
</tr>
<tr>
<td>0.03</td>
<td>0.1508</td>
<td>0.1200</td>
<td>0.2703</td>
<td>1.5773</td>
<td>0.2401</td>
<td>0.5404</td>
<td>2.8989</td>
<td>96.6298</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>4.4348</td>
<td>0.4802</td>
<td>1.0897</td>
<td>6.7276</td>
<td>168.1904</td>
</tr>
<tr>
<td>0.048</td>
<td>0.2435</td>
<td>0.1921</td>
<td>0.4337</td>
<td>47.0769</td>
<td>0.3841</td>
<td>0.8674</td>
<td>49.1977</td>
<td>1024.9521</td>
</tr>
</tbody>
</table>

Table 5.4: Number of jobs at each node along with the response time of the system for varying arrival rates

(For every arrival rate in the range, the table shows the number of jobs at each node. The field ‘Total L’ is simply the total number of jobs in the system (i.e – the summation of jobs at each node).
From the table, we can see that as the arrival rate increases, the response time increases. Furthermore, if we look at the response time for the arrival rate of 0.048, we can see that it is a very large number. This is because the arrival rate is approaching its asymptote and as a result the value of the response time moves towards infinity. Using the values of the arrival rate (on the x-axis) and response time (on the y-axis) we can plot the graph using Mathematica.

The same procedure is carried out for M = 2, M = 4 and M = 8 (tables shown in the appendix – A.1, A.2, A.3) although each of them will have different asymptotes. (i.e. maximum values of arrival rate after which tests will saturate and won’t preserve its steady state any more.)

The asymptotes are as follows:

<table>
<thead>
<tr>
<th>Number of Servers (M)</th>
<th>Maximum Arrival Rate - (Patients/Minute)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>$U = \frac{0.68 \lambda}{2 \times \left(\frac{2}{60}\right)} = 1$ =&gt; $\lambda = 0.098$</td>
</tr>
<tr>
<td>4</td>
<td>$U = \frac{0.68 \lambda}{4 \times \left(\frac{2}{60}\right)} = 1$ =&gt; $\lambda = 0.196$</td>
</tr>
<tr>
<td>8</td>
<td>$U = \frac{0.68 \lambda}{8 \times \left(\frac{2}{60}\right)} = 1$ =&gt; $\lambda = 0.3921$</td>
</tr>
</tbody>
</table>

Table 5.5: Maximum Arrival Rates for varying number of servers at tests

Once the response times for varying arrival rate were calculated for each of the M values, they were plotted on the same graph in order for comparisons to be made easily. Shown below is the graph:
CHAPTER 5. MODELLING WITH NETWORK OF QUEUES

The graph above shows us the following:

1) All the lines have the same shape as the response time distribution graph shown above. This is true because in essence, the only difference is the number of servers at tests which in turn affects the maximum arrival rate and response time of the system.

2) All the curves on the graph (except M=8) seem to be bounded by the asymptotes calculated above as their response times are moving towards infinity.

3) The graph enables us to see the minimum number of servers that are needed for a particular arrival rate. For example, if we had an arrival rate of 0.8 Patients/Minute, we would only need 4 servers in order to cope with this arrival rate. This is because the graph seems to suggest that there is no difference in the response time of the system when the number of servers is increased to 8. Therefore, we wouldn’t need to waste our resources and add more staff to tests at this arrival rate. Although, if the arrival rate increased to 0.18 Patients/Minute, then we can see that there is a large increase in the response time when 4 servers are used in comparison to when 8 servers are used at tests. Hence, it will be up to the management staff to decide whether or not this increase in response time is acceptable taking into account the limited amount of resources that they have.

(Note: the asymptote for “M=8” cannot be seen because it occurs when \( \lambda = 0.392 \) Patients/Minute and as a result of this large arrival rate, a lot of the other nodes will have to increase their capacity in order to preserve the steady state of the system. This has been discussed later.)

5.3.2 Varying the Service Rate at Tests

In addition to varying the number of servers at this bottleneck, we can also see what happens when we vary the service rate of tests and how it affects the response time of the system. In terms of the A&E service unit, varying the service rate would mean that we change the speed at which the equipment works. For example, if we were to increase the service rate, we would have to get equipment that would serve the customers faster. (Since there isn’t really a way in which the staff at tests could work any faster). The model above has parameters of M=4 and \( \mu = 2/60 \) Patients/Minute. Keeping M the same, it would interesting to find out what happened when \( \mu \) was varied between \( \mu = 1/60 \) and \( \mu = 4/60 \) Patients/Minute for varying arrival rates.

In order to do this, we have to use the same method as above (varying M for varying arrival rates) except this time we will be varying \( \mu \) at tests instead of M. Therefore, we will have to calculate the queue lengths at each node, sum them up and then using little’s law find the response time of that particular arrival rate. (The tables of the queue lengths and response time for varying \( \lambda \) are given in the appendix – A.4, A.5, A.6, A.7). As with M, the asymptotes of the graphs (maximum arrival rate) for different values of \( \mu \) are calculated using the utilisation formula as follows:
As we vary $\mu$ at the bottleneck, shown below is the graph of the response time of the system for varying arrival rates.

The graph above shows us the following:

1) The larger the service rate, the smaller the response time for all arrival rates from $\lambda = 0.01$ to $\lambda = 0.2$ Patients/Minute.

2) When ‘tests’ has $\mu = 1/60$ & $\mu = 2/60$ Patients/Minute, we can see that their response times tend to infinity as they reach their asymptotes.

3) There doesn’t seem to be much of a difference in the response times between $\mu = 3/60$ & $\mu = 4/60$ at tests. As a result, we can say that if the arrival rate of the A&E is below 0.2 Patients/Minute, then it will be sufficient for $\mu = 3/60$ unless having the lowest response time possible is a necessity. This saves us the need to have the current equipment running faster than a service rate of 3/60 Patients/Minute.
5.4 Other bottlenecks in the System

In order to find other bottlenecks in this system, and carry out a bottleneck analysis on them, we would have to vary the arrival rate and see what happens to the utilisation at the nodes. Given below is a table that shows the utilisation values at each node for the arrival rate of 0.01-0.2 Patients/Minute.

<table>
<thead>
<tr>
<th>Arrival Rate</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.025</td>
<td>0.013</td>
<td>0.030</td>
<td>0.051</td>
<td>0.020</td>
<td>0.045</td>
</tr>
<tr>
<td>0.02</td>
<td>0.050</td>
<td>0.027</td>
<td>0.060</td>
<td>0.102</td>
<td>0.040</td>
<td>0.090</td>
</tr>
<tr>
<td>0.03</td>
<td>0.075</td>
<td>0.040</td>
<td>0.090</td>
<td>0.153</td>
<td>0.060</td>
<td>0.135</td>
</tr>
<tr>
<td>0.04</td>
<td>0.100</td>
<td>0.053</td>
<td>0.120</td>
<td>0.204</td>
<td>0.080</td>
<td>0.180</td>
</tr>
<tr>
<td>0.05</td>
<td>0.125</td>
<td>0.067</td>
<td>0.150</td>
<td>0.255</td>
<td>0.100</td>
<td>0.225</td>
</tr>
<tr>
<td>0.06</td>
<td>0.150</td>
<td>0.080</td>
<td>0.180</td>
<td>0.306</td>
<td>0.120</td>
<td>0.270</td>
</tr>
<tr>
<td>0.07</td>
<td>0.175</td>
<td>0.093</td>
<td>0.210</td>
<td>0.357</td>
<td>0.140</td>
<td>0.315</td>
</tr>
<tr>
<td>0.08</td>
<td>0.200</td>
<td>0.107</td>
<td>0.240</td>
<td>0.408</td>
<td>0.160</td>
<td>0.360</td>
</tr>
<tr>
<td>0.09</td>
<td>0.225</td>
<td>0.120</td>
<td>0.270</td>
<td>0.459</td>
<td>0.180</td>
<td>0.405</td>
</tr>
<tr>
<td>0.10</td>
<td>0.250</td>
<td>0.133</td>
<td>0.300</td>
<td>0.510</td>
<td>0.200</td>
<td>0.450</td>
</tr>
<tr>
<td>0.11</td>
<td>0.275</td>
<td>0.147</td>
<td>0.330</td>
<td>0.561</td>
<td>0.220</td>
<td>0.495</td>
</tr>
<tr>
<td>0.12</td>
<td>0.300</td>
<td>0.160</td>
<td>0.360</td>
<td>0.612</td>
<td>0.240</td>
<td>0.540</td>
</tr>
<tr>
<td>0.13</td>
<td>0.325</td>
<td>0.173</td>
<td>0.390</td>
<td>0.663</td>
<td>0.260</td>
<td>0.585</td>
</tr>
<tr>
<td>0.14</td>
<td>0.350</td>
<td>0.187</td>
<td>0.420</td>
<td>0.714</td>
<td>0.280</td>
<td>0.630</td>
</tr>
<tr>
<td>0.15</td>
<td>0.375</td>
<td>0.200</td>
<td>0.450</td>
<td>0.765</td>
<td>0.300</td>
<td>0.675</td>
</tr>
<tr>
<td>0.16</td>
<td>0.400</td>
<td>0.213</td>
<td>0.480</td>
<td>0.816</td>
<td>0.320</td>
<td>0.720</td>
</tr>
<tr>
<td>0.17</td>
<td>0.425</td>
<td>0.227</td>
<td>0.510</td>
<td>0.867</td>
<td>0.340</td>
<td>0.765</td>
</tr>
<tr>
<td>0.18</td>
<td>0.450</td>
<td>0.240</td>
<td>0.540</td>
<td>0.918</td>
<td>0.360</td>
<td>0.810</td>
</tr>
<tr>
<td>0.19</td>
<td>0.475</td>
<td>0.253</td>
<td>0.570</td>
<td>0.969</td>
<td>0.380</td>
<td>0.855</td>
</tr>
<tr>
<td>0.20</td>
<td>0.500</td>
<td>0.267</td>
<td>0.600</td>
<td>1.020</td>
<td>0.400</td>
<td>0.900</td>
</tr>
</tbody>
</table>

Table 5.7: Utilisation at each node for arrival rates between 0.01 & 0.2 Patients/Minute

As we can see, for each arrival rate from $\lambda = 0.01$ Patients/Minute to approximately $\lambda = 0.2$ Patients/Minute, tests has the highest utilisation value amongst all the nodes. It is at the value of 0.196 Patients/Minute (calculated in table 5.6) that the utilisation of tests (M=4) becomes 1. This can be seen by the fact that $U$ (tests) at $\lambda = 0.2$ is 1.020. Hence, in order to preserve the steady state of the system and bring down its utilisation, the number of servers at tests is increased to 8. After this change, we can see which nodes become bottlenecks (if at all) when we increase the arrival rate further from $\lambda = 0.2$ Patients/Minute. Given below are the utilisation values of these nodes:

<table>
<thead>
<tr>
<th>Arrival Rate</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.21</td>
<td>0.525</td>
<td>0.280</td>
<td>0.630</td>
<td>0.535</td>
<td>0.420</td>
<td>0.945</td>
</tr>
<tr>
<td>0.22</td>
<td>0.550</td>
<td>0.293</td>
<td>0.660</td>
<td>0.561</td>
<td>0.440</td>
<td>0.990</td>
</tr>
<tr>
<td>0.23</td>
<td>0.575</td>
<td>0.307</td>
<td>0.690</td>
<td>0.586</td>
<td>0.460</td>
<td>1.035</td>
</tr>
</tbody>
</table>

Table 5.8: Utilisation at each node for arrival rates between 0.2 & 0.23 Patients/Minute
The table above shows that once the number of servers at tests is increased to 8, the utilisation drops immediately to 0.51 after $\lambda = 0.2$ Patients/Minute. At this value, major’s doctor proves to be the bottleneck since it has the highest utilisation value. Hence, we should carry out a bottleneck analysis as before in order to see the effect of varying the parameters at this bottleneck.

(Note: The range of the arrival rate is only from 0.2 – 0.23 since at 0.23, the utilisation of Majors Doctor goes over 1, i.e – it becomes saturated and the steady state isn’t preserved any more. As a result, we will have to increase the capacity at this node by either increasing the service rate or increasing the number of servers.)

Shown below is the bottleneck analysis done on Majors Doctor:

### 5.4.1 Varying the Number of Servers at Majors Doctor

At the moment, the major’s doctor node has 4 doctors serving it and has a $\mu = 2/60$ Patients/Minute. As a result, if we varied the number of doctors (servers) at this bottleneck, we could see what happens to the response time of the system for varying arrival rates. As with the previous bottleneck, the method to do this is the same except this time, we will be varying $M$ at Doctors Major instead of tests. I have chosen to vary $M$ from 4 – 7 doctors.

The calculations of the total number of jobs in the system and response time for the corresponding arrival rates varying values of $M$ are shown in the appendix. (A.8, A.9, A.10, A.11). As stated earlier, each of these $M$ values will have maximum arrival rates after which the node will get saturated and won’t preserve its steady state property. Given in the table below are the rates for each $M$ value.

<table>
<thead>
<tr>
<th>Number of Servers (M)</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maximum Arrival Rate</strong> - (Patients/Minute)</td>
<td>$U = \frac{0.6 \lambda}{4 \times \left(\frac{2}{60}\right)} = 1$</td>
<td>$U = \frac{0.6 \lambda}{5 \times \left(\frac{2}{60}\right)} = 1$</td>
<td>$U = \frac{0.6 \lambda}{6 \times \left(\frac{2}{60}\right)} = 1$</td>
<td>$U = \frac{0.6 \lambda}{7 \times \left(\frac{2}{60}\right)} = 1$</td>
</tr>
<tr>
<td>$\Rightarrow \lambda = 0.222$</td>
<td>$\Rightarrow \lambda = 0.277$</td>
<td>$\Rightarrow \lambda = 0.333$</td>
<td>$\Rightarrow \lambda = 0.392$</td>
<td></td>
</tr>
</tbody>
</table>

**Table 5.9: Maximum Arrival Rates for varying the number of doctors at Majors Doctor**

Using the arrival rates and response times calculated in the tables listed above the following graph was plotted using Mathematica.
The graph above shows us the following:

1) As with all the response time graphs, each curve has the same shape as the general response time graph varying with $\lambda$. Furthermore, each curve is bounded by the asymptotes calculated above.

2) For arrival rates from 0.2 – 0.25, whether there are 5, 6 or 7 doctors at Majors Doctor doesn’t seem to make much of a difference to the response time of the overall system. Hence, if a small increase in response time is acceptable to the A&E, for such arrival rates, it will be sufficient to have 5 doctors treating the patients at Doctor’s Major.

3) For arrival rates above $\lambda = 0.25$, we could use 5 doctors although the response time when this is so, is very large because it approaches its asymptote at 0.2777. If such response times are acceptable the A&E could still use 5 doctors until $\lambda = 0.2777$. After this value, we would have to increase the number of doctors to either 6 or 7, both who have a marginally small difference until approximately $\lambda = 0.3$ when the “M=6” curve starts moving towards infinity as it approaches its asymptote.

4) When the arrival rate increases above 0.3 (i.e. – 18 patients/minute – in cases of terrorist attacks, epidemic flu etc), it would be more advisable to use 7 doctors at this node in order to cope with the increase in the arrival rate.
5.5 Effect of Increase in the External Arrival Rate

The table showing the utilisation values (table 5.8) only shows us the utilisation values of the nodes until $\lambda = 0.23$. If there were a further increase in the arrival rate, for example, there was a terrorist attack in the country and there was a sudden increase in the arrival rate, it would be interesting to see which of the nodes becomes the bottleneck and what we could do at such nodes in order to help the A&E plan for such events. Shown below are the utilisation values.

(Note: the number of doctors at Doctors Major has been increased to 5 since using 4 doctors; it gets saturated at $\lambda = 0.222$ as shown above.)

<table>
<thead>
<tr>
<th>Arrival Rate</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.23</td>
<td>0.575</td>
<td>0.307</td>
<td>0.690</td>
<td>0.5865</td>
<td>0.460</td>
<td>0.828</td>
</tr>
<tr>
<td>0.24</td>
<td>0.600</td>
<td>0.320</td>
<td>0.720</td>
<td>0.612</td>
<td>0.480</td>
<td>0.864</td>
</tr>
<tr>
<td>0.25</td>
<td>0.625</td>
<td>0.333</td>
<td>0.750</td>
<td>0.6375</td>
<td>0.500</td>
<td>0.9</td>
</tr>
<tr>
<td>0.26</td>
<td>0.650</td>
<td>0.347</td>
<td>0.780</td>
<td>0.663</td>
<td>0.520</td>
<td>0.936</td>
</tr>
<tr>
<td>0.27</td>
<td>0.675</td>
<td>0.360</td>
<td>0.810</td>
<td>0.6885</td>
<td>0.540</td>
<td>0.972</td>
</tr>
<tr>
<td>0.28</td>
<td>0.700</td>
<td>0.373</td>
<td>0.840</td>
<td>0.714</td>
<td>0.560</td>
<td>1.008</td>
</tr>
</tbody>
</table>

Table 5.10: Utilisation at each node for arrival rates between 0.23 & 0.28 Patients/Minute

From the table above, we can see that Doctor Major still proves to be the bottleneck between $\lambda = 0.23$ and $\lambda = 0.28$ (approximately). At this value, the utilisation goes above 1 since it has reached its asymptote of ($\lambda = 0.2777$) and once again, we will have to increase the number of doctors to 6 at this node when $\lambda > 0.2777$. Once this change is made, we would have to look at the utilisations of the nodes again to see whether the bottleneck has moved. Shown in the table below are the utilisations:

<table>
<thead>
<tr>
<th>Arrival Rate</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.28</td>
<td>0.700</td>
<td>0.373</td>
<td>0.840</td>
<td>0.714</td>
<td>0.560</td>
<td>0.84</td>
</tr>
<tr>
<td>0.29</td>
<td>0.725</td>
<td>0.387</td>
<td>0.870</td>
<td>0.7395</td>
<td>0.580</td>
<td>0.87</td>
</tr>
<tr>
<td>0.30</td>
<td>0.750</td>
<td>0.400</td>
<td>0.900</td>
<td>0.765</td>
<td>0.600</td>
<td>0.9</td>
</tr>
<tr>
<td>0.31</td>
<td>0.775</td>
<td>0.413</td>
<td>0.930</td>
<td>0.7905</td>
<td>0.620</td>
<td>0.93</td>
</tr>
<tr>
<td>0.32</td>
<td>0.800</td>
<td>0.427</td>
<td>0.960</td>
<td>0.816</td>
<td>0.640</td>
<td>0.96</td>
</tr>
<tr>
<td>0.33</td>
<td>0.825</td>
<td>0.440</td>
<td>0.990</td>
<td>0.8415</td>
<td>0.660</td>
<td>0.99</td>
</tr>
<tr>
<td>0.34</td>
<td>0.850</td>
<td>0.453</td>
<td>1.020</td>
<td>0.867</td>
<td>0.680</td>
<td>1.02</td>
</tr>
</tbody>
</table>

Table 5.11: Utilisation at each node for arrival rates between 0.28 & 0.34 Patients/Minute

The table above shows us a very interesting result in that there is no specific bottleneck in the system when the arrival rate is between 0.28 and 0.33. It turns out to be that there are two bottlenecks in this range - the majors nurse and majors doctor as they have the same utilisation values. At $\lambda = 0.333$, (calculated as an asymptote for Majors Doctor), the majors nurse and majors doctor have utilisations of 1 and hence, the number of staff at these nodes have to be increased if the arrival rate increases. Increasing the number of nurses at Majors nurse from 3 to 4, and the number of
doctors at Majors Doctor from 6 to 7, we get the following utilisation values when $\lambda > 0.333$.

<table>
<thead>
<tr>
<th>Arrival Rate</th>
<th>Receptionist</th>
<th>Nurse Minor</th>
<th>Nurse Major</th>
<th>Tests</th>
<th>Doctor Minor</th>
<th>Doctor Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.34</td>
<td>0.850</td>
<td>0.453</td>
<td>0.765</td>
<td>0.867</td>
<td>0.680</td>
<td>0.874</td>
</tr>
<tr>
<td>0.35</td>
<td>0.875</td>
<td>0.467</td>
<td>0.788</td>
<td>0.893</td>
<td>0.700</td>
<td>0.900</td>
</tr>
<tr>
<td>0.36</td>
<td>0.900</td>
<td>0.480</td>
<td>0.810</td>
<td>0.918</td>
<td>0.720</td>
<td>0.926</td>
</tr>
<tr>
<td>0.37</td>
<td>0.925</td>
<td>0.493</td>
<td>0.833</td>
<td>0.944</td>
<td>0.740</td>
<td>0.951</td>
</tr>
<tr>
<td>0.38</td>
<td>0.950</td>
<td>0.507</td>
<td>0.855</td>
<td>0.969</td>
<td>0.760</td>
<td>0.977</td>
</tr>
<tr>
<td>0.39</td>
<td>0.975</td>
<td>0.520</td>
<td>0.878</td>
<td>0.995</td>
<td>0.780</td>
<td>1.003</td>
</tr>
</tbody>
</table>

Table 5.12: Utilisation at each node for arrival rates between 0.34 & 0.39 Patients/Minute

As we can see, between arrival rates of 0.34 and 0.39, Doctor Major seems to be the bottleneck, although the utilisations of receptionist and tests too seem to be very high. Hence, even though the steady state property is preserved using these parameters, the management of the A&E might look into bringing down these utilisations by adding more servers or increasing the service rate.

### 5.6 Response Time Distribution for Paths in the Simple Network

As stated previously, one of the objectives of this project was to find response time distributions of paths within a network. A path is simply a route through which a patient may pass through an A&E service unit. As shown in section 4.7, this can be done with the use of Laplace transforms. Hence if we use equation (4.40) in chapter 4, we can uniquely define the probability density function of each path in the network and using this, find the mean response time of the path.(shown below). This section also looks at finding the mean response time of all the paths in the system and compares it to the mean response time value we obtained using Little’s law in section (5.2.3). In essence, both these values should be the same.

In our model of the A&E, since it’s a one-class model, any patient entering tests is not classified uniquely as a majors or minors. Therefore, in essence, a minor could go to the majors doctor and likewise with the majors, they could go to minors doctor. Hence, we can see that there are 5 paths through which a patient can go through in our model.

**Path 1:** Receptionist -> Majors Nurse -> Tests -> Majors Doctor -> Exit (4 nodes)

**Path 2:** Receptionist -> Minors Nurse -> Tests -> Minors Doctor -> Exit (4 nodes)

**Path 3:** Receptionist -> Minors Nurse -> Minors Doctor -> Exit (3 nodes)

**Path 4:** Receptionist -> Majors Nurse -> Tests -> Minors Doctor -> Exit (4 nodes)

**Path 5:** Receptionist -> Minors Nurse -> Tests -> Majors Doctor -> Exit (4 nodes)
Path 1:

This path is the path that any patient classified as a major will take whilst in the A&E service unit. In order to calculate the response time distribution of this path, we can use equation (4.37) with number with the number of nodes set to 4 and the parameters at these nodes set to the ones shown in section 4.7. The calculations of all the constants and the probability density function were performed using Mathematica and time distribution graph for the following path was plotted as shown below.

(Note: since 2 nodes in this path – (i.e – tests and Doctor Major) have the same service rate (\( \mu \)) as \( \frac{2}{60} \), the service rate at tests was changed to 2.01 in order to prevent getting 0 as the denominator and hence causing the constants \( \alpha \) and \( \beta \) to be \( \infty \).)

If this graph is compared to the response time distribution graphs shown in section 4, (refer section 4.4.2 for further explanation) we can see that the shape of the graph is approximately the same and there is a high probability of the response time being approximately 170 minutes. This response time value is higher than the average response time value (as shown in section 3.3). This is because this graph shows us the response time distribution of a major’s patient and on average, majors have high response times since they require more attention. We can also see the tail of the graph flattening out at approximately 900 minutes. i.e – the probability of having a response time of 900 minutes, is 0.
Using the equation of the probability density function \( f(x) \) derived, we can find the mean response time of the path using the formula shown below:

\[
\text{Mean Response Time} = \int_{0}^{\infty} x f(x) \, dx
\]  

(5.4)

Hence, using Mathematica, this equation was calculated and the mean response time of this path was 216.696 Minutes.

**Path 2:** Receptionist -> Minors Nurse -> Tests -> Minors Doctor -> Exit (4 nodes)

This path is taken by any minor that visits the tests node. The response time distribution was found in the same way as that for path 1 above using the same parameters as shown in table 5.1. The calculations of the constants and the response time distribution equation were performed using Mathematica. Given below is a graph of the response time distribution function plotted:

![Graph of response time distribution function](image)

The graph above shows us that the probability of having a response time of approximately 130 minutes is very high. This is lower than the response time of the majors, and this is as expected since the minors take a shorter amount of time through the A&E service unit. In this graph, the tail starts flattening out much lower than that of the majors at a value of approximately 800 minutes.

The mean response time of this path was calculated using equation (5.4) which yielded a value of 170.047 Minutes.

**Path 3:** Receptionist -> Minors Nurse -> Minors Doctor -> Exit (3 nodes)

This path is taken by a minors patient who doesn’t have to visit the tests node. The response time distribution function was calculated as with the previous paths. Using Mathematica, the response time distribution was plotted as shown below:
The graph above shows us that there is a very high chance of having a very small response time (approximately 30 minutes) if patients took this path. This is significantly smaller than the previous 2 paths. This is because a patient using this path doesn’t visit tests and as a result, spends less time in the A&E service unit. Furthermore, we can also see the graph flattening out at approximately 160 minutes. This suggests that, in essence, the probability of having a response time (using this path) of over 160 minutes is 0. i.e – a patient will almost never have a response time over 160 minutes.

The mean response time calculated for this path was: 37.1082 Minutes.

Path 4: Receptionist -> Majors Nurse -> Tests -> Minors Doctor -> Exit (4 nodes)

This is one of the additional paths that result due to the fact that our model is a one class model and therefore, a majors patient could visit a minors doctor after tests. Using Mathematica, the following response time distribution was plotted
As we can see, this response time distribution graph has its peak occurring at a response time of approximately 140 minutes which is slightly lower than that of a majors patient using Path 1. This is because, in essence, this is the typical path that a majors patient would take except this time, he/she visits the minors doctor instead of the majors one.

The mean response time calculated for this path was: 178.202 minutes which is significantly lower than the mean response time calculated for path 1 as expected. This is because as stated above, in this path the patient visits the minors doctor node where there are more doctors each of whom works faster than the majors doctors. Hence, response times at this node are low which brings down the overall response time of the system.

**Path 5:** Receptionist -> Minors Nurse -> Tests -> Majors Doctor -> Exit (4 nodes)

This is the second additional path that a minors patient could take using our model due to the fact that he visits the majors doctor instead of the minors doctor. Using Mathematica, the following response time distribution was plotted as shown below:

As expected, this graph is very similar to the graph plotted for path 2 with the slight difference being that it has its highest probability is when the response time is approximately 160 minutes. i.e – the most probable value for the response time is 160 minutes. The reason for this increase in comparison to path 2 is because in this path, a minors patient visits a majors doctor. There are fewer majors doctor at the majors doctor node each of whom work slower. Hence response times at this node are large thus causing the overall response time of the system to increase.

The mean response time of this path was 208.54 minutes.

Using the mean response times for each path, we can calculate the mean response time for the whole network as shown in the next section.
5.7 Mean Response Time for Simple Network

The mean response time for the network can be found by summing the weighted response time of all three paths. In order to find the weighted response time, we will have to find the probability of each path and then multiply this by the response time of that path.

The probabilities of the different paths are found by finding the proportion of patients that use that path. Hence, following each path as shown in Figure 1, the probabilities of the paths are as follows:

<table>
<thead>
<tr>
<th>Path</th>
<th>Probability of Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6<em>1</em>(15/17)</td>
</tr>
<tr>
<td></td>
<td>= 0.529</td>
</tr>
<tr>
<td>2</td>
<td>0.4<em>0.2</em>(2/17)</td>
</tr>
<tr>
<td></td>
<td>= 0.00941</td>
</tr>
<tr>
<td>3</td>
<td>0.4*0.8</td>
</tr>
<tr>
<td></td>
<td>= 0.32</td>
</tr>
<tr>
<td>4</td>
<td>0.6<em>1</em>(2/17)</td>
</tr>
<tr>
<td></td>
<td>= 0.071</td>
</tr>
<tr>
<td>5</td>
<td>0.4<em>0.2</em>(15/17)</td>
</tr>
<tr>
<td></td>
<td>= 0.071</td>
</tr>
</tbody>
</table>

Table 5.13: Probabilities of each path within Simple Network (5 Paths)

Hence, finding the weighted response time:

\[
(0.6*(15/17)*216.696) + (0.08*(2/17)*170.047) + (0.32*37.108) + \\
(0.6*(2/17)*178.202) + (0.08*(15/17)*208.54) \\
= 155.496 	ext{ Minutes}
\]

5.7.1 Verifying Mean Response Time using Little’s Law

In order to verify our value calculated for the mean response time above, we can use Little’s law. As shown in section 5.2.3, a value of 162.517 Minutes was calculated as the response time using the parameters shown in table 5.1 for the model of the A&E service unit. Although, as stated in section 5.5, the service rate at tests was varied to \( \mu = \frac{201}{60} \). Hence, in order to take into account this change, the number of jobs was calculated at tests keeping all the other parameters the same.

With this service rate, the number of jobs at tests was 16.6986 giving us a total number of jobs in the system equal to 28.7235. Using Little’s law, the response time of the system is:

\[
= \frac{28.7235}{0.18472} = 155.496 \text{ Minutes}
\]

As we can see, this value is the same as the value calculated using the response time distributions above. Hence, we can confirm that the response time distributions calculated in the previous section were correct.
5.7.2 Approximating the Model

If we approximated our model and assume that all majors go to the majors doctor and all minors go to the minors doctor, then we would only get 3 paths. i.e – Path 1, 2 and 3. If we were to find the mean response time using such a path, the probabilities of these paths would change in order to compensate for the assumption made above. Hence, since a proportion of 60% of all patients that enter tests are majors, 60% of all patients leaving tests to majors doctor are majors. And for minors, 8% leaving tests to minors doctor are minors. This gives us the following Path probabilities.

<table>
<thead>
<tr>
<th>Path</th>
<th>Probability of Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6<em>1</em>1=0.6</td>
</tr>
<tr>
<td>2</td>
<td>0.2<em>0.4</em>1=0.08</td>
</tr>
<tr>
<td>3</td>
<td>0.4*0.8=0.32</td>
</tr>
</tbody>
</table>

Table 5.14: Probabilities of each path within Simple Network (3 Paths)

Using these probabilities to calculate the mean response time of the network:

\[
\text{Mean Response Time} = (0.6*216.696) + (0.08*170.047) + (0.32*37.108)
\]

\[
= 155.496 \text{ Minutes}
\]

As we can see, this model too matches the response time value calculated using Little’s Law and suggests that assumption made above is reasonable.

5.8 Conclusion for the simple network of the A&E

As we can see, the results from the A&E provide a good insight into the response time at different arrival rates. It also shows us the bottleneck’s in the system and sees the effect of varying the parameters at such nodes. This sort of analysis will be of great help to the management of an A&E service unit. Since the number of resources (i.e – staff, equipment) etc are always limited, using such an analysis, the management can see what the minimum number of staff at each node can be whilst preserving the steady state theorem. For example, if there was a terrorist attack and the arrival rate increased to 0.3 Patients/Minute, the management could see that the minimum number of staff stay the same as showed in table 5.11 but the number of staff at tests has to increase to 8, and Majors Doctor changes to 6 in order for the steady state to be preserved. Hence, this system helps us allocate staff to different nodes in the system.

5.9 Adding the specialist doctor

As stated in the introduction section of this chapter, we need to look at the effect on the response time of the system when a specialist doctor is added. In essence, this type of doctor, (as described in the background section 2.1, looks into the injuries/illnesses of majors/minors once they have been referred to by the majors/minors practitioner. Examples of such doctors are orthopaedics, Gynaecologists, ENT’s etc. Shown on the next page is the extended queuing model with the specialist doctor included.
Figure 5.2:
(Note: Since a patient only sees the specialist doctor after seeing a minors/majors practitioner, the specialist doctor is added after tests.)

From the diagram, we can see that the specialist doctor is the seventh node in the system. We can also see that the proportion of majors/minors that visit the specialist doctor after tests is 25% of all majors and minors that enter the ‘tests’ node. This leaves 75% of all majors/minors to visit the majors/minors doctor accordingly and hence arrival rates at these nodes vary.

One of the first things we could do with such a system is to calculate its response time. As stated previously, we would need to find the number of jobs at each node and sum them up to find the total number of jobs in the system. Then, using little’s law, we need to divide this value by the external arrival rate. Shown in the table below are the parameters of this model. (i.e – the number of servers, service rate and arrival rate for each node).

<table>
<thead>
<tr>
<th>Receptionist</th>
<th>Minors Nurse</th>
<th>Majors Nurse</th>
<th>Tests</th>
<th>Minors Doctor</th>
<th>Majors Doctor</th>
<th>Specialist Doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Servers (M)</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Service Rate $\mu$ (Patients/Minute)</td>
<td>12</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Traffic Rate $\lambda$</td>
<td>$0.4\lambda$</td>
<td>$0.6\lambda$</td>
<td>$(0.6<em>1)+(0.4</em>0.2)$ = $0.68\lambda$</td>
<td>$(0.4<em>0.8)+(0.08</em>0.75)$ = $0.38\lambda$</td>
<td>$0.6*0.75$ = $0.45\lambda$</td>
<td>$(0.6<em>0.25)+(0.08</em>0.25)$ = $0.17\lambda$</td>
</tr>
</tbody>
</table>

Table 5.16: Parameters and Traffic Rates at each node in Extended Network

As we can see, the traffic rates at both the doctors have reduced since a proportion of the patients coming out of tests now go to the Specialist Doctor. These parameters were plugged into Mathematica and the number of jobs at each node was found as follows:

<table>
<thead>
<tr>
<th>Receptionist</th>
<th>Minors Nurse</th>
<th>Majors Nurse</th>
<th>Tests</th>
<th>Minors Doctor</th>
<th>Majors Doctor</th>
<th>Specialist Doctor</th>
<th>Whole System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Of Jobs</td>
<td>1.17398</td>
<td>0.752762</td>
<td>2.03304</td>
<td>17.9955</td>
<td>1.43678</td>
<td>3.01982</td>
<td>1.21072</td>
</tr>
</tbody>
</table>

Table 5.17: Number of Jobs at each node

Hence, using Little’s Law with an arrival rate of 0.184722 Patients/Minute (as calculated in section 3.1), we get a response time of:

$$w = \frac{27.6226}{0.184722}$$

$$= 149.536\text{ Minutes}$$
Firstly, we can see that this response time value is less than the one calculated for the simple network (i.e – without the specialist doctor). This is because, in essence, there is another “service centre” i.e – a new node which processes the patients. We can also see that the response time value of this model is closer to the actual response time of 136.8122 Minutes/Patient (calculated in section 3.3) than the response time value calculated for the simple network. This means that this model (with the specialist doctor) is a more realistic and accurate model of the A&E service unit. As with all such queuing models, this model can be tweaked in order to get the response time value closer to the actual response time.

5.10 Bottlenecks in the Extended Network

In order to see the effect of the addition of the specialist doctor on the overall performance of the system, we will have to see what happens to the utilisations of each node. Given below are the utilisation values calculated using the parameters declared above:

<table>
<thead>
<tr>
<th></th>
<th>Receptionist</th>
<th>Minors Nurse</th>
<th>Majors Nurse</th>
<th>Tests</th>
<th>Minors Doctor</th>
<th>Majors Doctor</th>
<th>Specialist Doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utilisation</td>
<td>0.461805</td>
<td>0.246296</td>
<td>0.554166</td>
<td>0.9420825</td>
<td>0.35097</td>
<td>0.623436</td>
<td>0.471041</td>
</tr>
</tbody>
</table>

Table 5.18: Utilisations at each node in extended network

As we can see from the utilisations, (with the arrival rate of 0.184722 Patients/Minute), tests still seems to be the bottleneck. Hence, the addition of the specialist doctor doesn’t change the position of the bottleneck at this arrival rate. The bottleneck analysis on this node would be very similar to the ‘tests’ node done in the previous section with the small difference being that the response time of the whole network will be lower for varying arrival rates.

Since the proportion of patients arriving at the specialist doctor have been approximated to be 25% of each class – Majors and Minors, we need to look at what happens to the overall performance of the system when this proportion is increased for the current arrival rate. For example, if there is a day when a lot of majors/minors need to visit the specialist doctor, this proportion will increase the arrival rate at the specialist doctor and hence we need to see at what proportion of majors/minors will the specialist doctor become the bottleneck in the system. Given below are some “what if” situations that calculate the utilisation of the specialist doctor if the proportion of majors and minors entering it are varied.
5.10.1 Proportion of majors and minors entering increases to 50% and 50%

When the proportion of majors and minors coming out of tests is increased to a half, the arrival rate at specialist doctor increases causing the utilisation to become:

\[ U = \frac{(0.6 \times 0.5) + (0.08 \times 0.5) \times 0.184722}{2 \times \frac{2}{60}} = 0.9420822 \]

As we can see, although, the utilisation of the specialist doctor increases greatly, it is still very marginally smaller than that of tests proving tests to still be the bottleneck. We now look at another “what if” situation

5.10.2 Proportion of majors increases to 55% and minors still 25%

In this case, I have kept the original approximation of the number of minors visiting the specialist doctor to 25% and increased the proportion of the majors to 55%. This seems to be a more realistic value as there will be a larger proportion of majors visiting the specialist doctor than minors. This gives us an utilisation value of:

\[ U = \frac{(0.6 \times 0.55) + (0.08 \times 0.25) \times 0.184722}{2 \times \frac{2}{60}} = 0.9698 \]

As we can see, with these proportions of majors and minors, the utilisation of the specialist doctor increases even more and now turns out to be the bottleneck of the system. Hence, we will have to perform a bottleneck analysis at this node in order to see the effect it has on the response time of the system.

5.11 Varying the number of servers at Specialist Doctor

The way this bottleneck analysis is done is similar to the ones done previously. In this case, we vary the number of servers at specialist doctor and see the effect that this has on the system keeping the proportion of majors and minors as 55% and 25% arriving into the specialist doctor. This causes the traffic rate at the specialist doctor to be equal to \( (0.6 \times 0.55) + (0.08 \times 0.25) = 0.35 \lambda \).

At the moment, the number of specialist doctors working in A&E service unit is 2. In this analysis, I have decided to vary this number from 1 to 4. The way the total number of jobs at each arrival rate (for varying arrival rates) is calculated as before and these calculations are shown in the appendix. (A.12, A.13, A.14, A.15)

As with all the analysis’, each value of M (i.e – the number of doctors) has an asymptote after which the utilisation of the specialist doctor increases over 1 and the steady state of the system is no longer preserved. Shown below are the asymptotes for corresponding m value.
Table 5.19: Maximum Arrival Rates for varying number of servers at specialist doctor

Using the calculations of the response time for corresponding arrival rates, the following graph was plotted when the number of specialist doctors was varied from 1 to 4.

The graph above shows us the following:

1) As with all the response time graphs, each curve has the same shape as the general response time graph varying with $\lambda$. Furthermore, each curve is bounded by the asymptotes calculated above.

2) For arrival rates between 0.05 and 0.13 whether there are 2, 3 or even 4 specialist doctors doesn’t make much of a difference to the response time of the system. Infact there is hardly any difference between using 3 or 4 doctors.

3) Above the rate of 0.13 patients/minute, the response time of the system gradually increases when the number of specialist doctors in the system is equal to 2 as it slowly approaches the asymptote (maximum arrival rate) of 0.190476. Hence, after this value, we will have to either have 3 or 4 specialist doctors.
4) One of the most interesting conclusions we can draw from this graph is that there is hardly any difference between using 3 and 4 specialist doctors since both lines follow each other almost identically. i.e – the response times for the system are the same whether there are 3 or 4 specialist doctors. Hence, using this graph, the management of the A&E service unit can see that they wouldn’t need to use an extra specialist doctor and hence save a resource.

(Note: The response times of M = 2, 3 and 4 are low since they still aren’t nearing their asymptotes. If the arrival rate were to increase, then some of the other nodes’ parameters would have to change and hence, the y-axis is only shown up to an arrival rate of 0.2 Patients/Minute.)

### 5.11.1 Varying the service rate at Specialist Doctor

As with the previous bottlenecks, another change we could make at the specialist doctor is to vary the service rate of each doctor. i.e – we could see what happens if the equipment that the specialist doctor uses works faster or slower hence causing the total service rate at which the doctor works at, to increase/decrease. At the moment, the service rate is \( \frac{2}{60} \). In this analysis, I have decided to vary the service rate from \( \frac{1}{60} \) to \( \frac{4}{60} \) and see the effect that this has on the response time of the system.

The response time (after this change was made) of the system was calculated for varying arrival rates as shown in appendix (A.16, A.17, A.18, A.19). The maximum arrival rate for each service rate value is shown below:

<table>
<thead>
<tr>
<th>Maximum Arrival Rate (Patients/Minute)</th>
<th>( U = \frac{0.35\lambda}{2 \times \left( \frac{1}{60} \right)} = 1 )</th>
<th>( U = \frac{0.35\lambda}{2 \times \left( \frac{2}{60} \right)} = 1 )</th>
<th>( U = \frac{0.35\lambda}{2 \times \left( \frac{3}{60} \right)} = 1 )</th>
<th>( U = \frac{0.35\lambda}{2 \times \left( \frac{4}{60} \right)} = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \frac{1}{60} )</td>
<td>( \Rightarrow \lambda = 0.0952 )</td>
<td>( \Rightarrow \lambda = 0.1905 )</td>
<td>( \Rightarrow \lambda = 0.2857 )</td>
<td>( \Rightarrow \lambda = 0.3809 )</td>
</tr>
<tr>
<td>( \frac{2}{60} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \frac{3}{60} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \frac{4}{60} )</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.20: Maximum Arrival Rates for varying service rate at specialist doctor

As we can see, the asymptotes of this analysis are the same as for the previous one. This is because in essence, we are now just varying the service rate – keeping m the same and in the previous analysis, it was doing the opposite.

Using the response time values calculated the following graph was plotted using Mathematica.
The graph above shows us the following:

1) All the lines i.e – different values of $\mu$ are bounded by their asymptotes and follow a similar structure to the graph shown for varying $M$ values at the specialist doctor.

2) Between an arrival rate of 0.05 and 0.12 Patients/Minute, there seems to be hardly any difference in the response time of the system when the specialist doctor works at $\mu = \frac{2}{60}, \frac{3}{60}, \frac{4}{60}$ . Hence if the management of the A&E wanted to save some resources (and not buy new equipment), for these arrival rates, they could have the specialist doctors working at a service rate of $\mu = \frac{2}{60}$.

3) After an arrival rate of 0.13 Patients/Minute, the response time of the system, when specialist doctors work at $\mu = \frac{2}{60}$ seems to gradually increase. If this increase in response time is acceptable for the management of the A&E service unit, then they could still work at this service rate.

4) There doesn’t seem to be much of a difference in the overall response time of the system when the specialist doctor works at $\mu = \frac{3}{60}$ or $\frac{4}{60}$ . Hence the management need not buy new equipment to make it run faster and the doctors can work at treating 3 patients per hour. i.e - $\mu = \frac{3}{60}$.

As we can see, this analysis at the specialist doctor is only carried out when the arrival rate of the A&E service unit is 0.184722 Patients/Minute. This gives us an utilisation
value of 0.9698 at the specialist doctor. Therefore, for any arrival rate less than this, the specialist doctor will still be the bottleneck. An interesting analysis to carry out will be to increase this arrival rate and see which nodes become the bottlenecks. This could well be the case in the A&E on particular days especially if there is an emergency such as a terrorist attack etc. Such an analysis would help the management of the A&E service unit as it will give them an idea of which nodes will need to most amount of resources/ how many will be needed etc.

As shown in table 5.20, the maximum arrival rate of the specialist doctor using the initial parameters is 0.19047 Patients/Minute. Hence, if the arrival rate changes to a value slightly higher than this, we will have to increase the number of specialist doctors to 3 to preserve the steady state of the system (as explained earlier). When this is done and the utilisation values calculated, the ‘tests’ node seems to be the bottleneck up to an arrival rate of 0.196 Patients/Minute.

Table 5.21: Utilisation at each node for arrival rates between 0.2 & 0.29 Patients/Minute

<table>
<thead>
<tr>
<th>Arrival Rate</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.500</td>
<td>0.267</td>
<td>0.600</td>
<td>0.510</td>
<td>0.380</td>
<td>0.405</td>
<td>0.700</td>
</tr>
<tr>
<td>0.21</td>
<td>0.525</td>
<td>0.280</td>
<td>0.630</td>
<td>0.536</td>
<td>0.399</td>
<td>0.425</td>
<td>0.735</td>
</tr>
<tr>
<td>0.22</td>
<td>0.550</td>
<td>0.293</td>
<td>0.660</td>
<td>0.561</td>
<td>0.418</td>
<td>0.446</td>
<td>0.770</td>
</tr>
<tr>
<td>0.23</td>
<td>0.575</td>
<td>0.307</td>
<td>0.690</td>
<td>0.587</td>
<td>0.437</td>
<td>0.466</td>
<td>0.805</td>
</tr>
<tr>
<td>0.24</td>
<td>0.600</td>
<td>0.320</td>
<td>0.720</td>
<td>0.612</td>
<td>0.456</td>
<td>0.486</td>
<td>0.840</td>
</tr>
<tr>
<td>0.25</td>
<td>0.625</td>
<td>0.333</td>
<td>0.750</td>
<td>0.638</td>
<td>0.475</td>
<td>0.506</td>
<td>0.875</td>
</tr>
<tr>
<td>0.26</td>
<td>0.650</td>
<td>0.347</td>
<td>0.780</td>
<td>0.663</td>
<td>0.494</td>
<td>0.527</td>
<td>0.910</td>
</tr>
<tr>
<td>0.27</td>
<td>0.675</td>
<td>0.360</td>
<td>0.810</td>
<td>0.689</td>
<td>0.513</td>
<td>0.547</td>
<td>0.945</td>
</tr>
<tr>
<td>0.28</td>
<td>0.700</td>
<td>0.373</td>
<td>0.840</td>
<td>0.714</td>
<td>0.532</td>
<td>0.567</td>
<td>0.980</td>
</tr>
<tr>
<td>0.29</td>
<td>0.725</td>
<td>0.387</td>
<td>0.870</td>
<td>0.740</td>
<td>0.551</td>
<td>0.587</td>
<td>1.015</td>
</tr>
</tbody>
</table>

Note: If we look at utilisation of tests in table 5.18, it marginally follows specialist doctor and hence, this too at an arrival rate of 0.196 reaches its maximum value.

After, an arrival rate of 0.196 Patients/minute, the number of servers at tests is increased to 8. Shown in the table below are the utilisation values of the nodes after an arrival rate of 0.2 Patients/Minute (Specialist doctor has increased to M=3 and tests, M=8)

As we can see in the table above, even when there are 3 specialist doctors, this node still proves to be the bottleneck in the system up to an arrival rate of 0.2857 Patients/Minute. Hence, any increase in the arrival rate after this value, will require the number of specialist doctors to increase to 4 in order to preserve the steady state of this system.

If this is done and we wish to see which node now becomes the bottleneck when the arrival rate increases further, we will have to calculate the utilisation values at each node.
### Table 5.22: Utilisation at each node for arrival rates between 0.3 & 0.34 Patients/Minute

As shown in the table above, Majors nurse is the bottleneck for any arrival rate between 0.3 and 0.34 Patients/Minute. Hence, we will need to perform a bottleneck analysis at this node and see how the response time of the system varies when the parameters at this node are varied.

#### 5.12 Bottleneck Analysis at Majors Nurse

At the moment, there are 3 majors nurses serving the A&E service unit. This analysis looks at what happens to the response time of the system when the number of majors nurses is kept at 3 and then increased to 4.

The maximum arrival rate that the Majors Nurse node can cope with 3 nurses is 0.333 Patients/Minute. This can be seen from the fact that in the table above, between the arrival rate of 0.33 and 0.34 the utilisation at Majors nurse rises above 1 and hence the steady state is no longer preserved. When there are 4 nurses at this node, the maximum arrival rate at Majors Nurse is 0.381 Patients/Minute.

The number of jobs at each node was calculated as before and is shown along with its response time values for varying arrival rates as shown in the appendix (A.20, A.21). Using these values, the following graph was plotted using Mathematica.
The graph above shows us:

1) Both the lines are approaching their asymptotes and their curves are similar to the general response time curve plotted in section 5.2.4

2) There seems to be a response time difference of approximately 30-40 minutes when 3 nurses are used instead of 4 for arrival rates between 0.3 and 0.32 Patients/Minute

3) After 0.32, since $M = 3$ approaches its asymptotes, the response time of the system increases drastically and hence, the management should look into increasing the number of nurses to 4 at this node after this value.

5.13 Response Time distributions of the extended network

As stated in the introduction section, one of the objectives of this project was to find response time distributions within paths of the network. Since we have introduced the extended network and established the fact that it models the A&E service unit more accurately, we can find response time distributions of paths in this network. The procedure to find these response time distributions is the same as that done for the simple network. However, due to the extra node in this system, i.e – the specialist doctor, we have 2 more paths than the general 3 paths that any patient could take. (Refer section 5.8)

Using the initial arrival rates, the response time distributions were calculated for paths 1, 2 and 3. The graphs of these paths would look almost identical to the ones shown in section 5.5. The only difference in these graphs would be that they would be marginally narrower and the tail will be slightly longer. This is because in all these paths, since the majors doctor and minors doctor have a slightly smaller arrival rate (as a proportion of patients go to the specialist doctor) these nodes will be less utilised, causing the overall response time to be smaller in that path. Hence, there is a greater chance of the response time being lower causing the shape of the graph to be narrower and the tail to be longer (marginally). This is explained in more detail in section 4.3.2.

The additional two paths are as follows:

Path 4: Receptionist -> Majors Nurse -> Tests -> Specialist Doctor -> Exit (4 nodes)

Path 5: Receptionist -> Minors Nurse -> Tests -> Specialist Doctor -> Exit (4 nodes)

Shown below are the response time distribution graphs for these graphs along with their mean response times.
Path 4:

This is the path that any major visiting the specialist doctor will take. Using the response time distribution, the following graph was plotted:

The graph above shows us that there is a high probability of patient having a response time of approximately 160 minutes. It also shows us that the tail flattens out at approximately 800 minutes. Using the method of calculating the mean response time, the mean response time of this path was found to be 196.192 minutes.

Path 5:

This is the path that any minor visiting the specialist doctor will typically take. Using the response time distribution, the following graph was plotted.
As we can see, the graph shows us that there is a high probability of the response time being approximately 150 minutes. Furthermore, in this graph too, the tail flattens out at approximately 800 minutes suggesting that it is almost unlikely that a patient using this path will ever have a response time of more than 800 minutes. The mean response time of this path was 188.036 Minutes.

Hence, we can use the method described in section 5.6 to calculate the mean response time of the whole network. Shown in the table below are the path probabilities along with the mean response time of each path.

<table>
<thead>
<tr>
<th>Path</th>
<th>Probability of Path</th>
<th>Mean Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.6<em>1</em>0.75)=0.45</td>
<td>193.966</td>
</tr>
<tr>
<td>2</td>
<td>(0.4<em>0.2</em>0.75)=0.06</td>
<td>170.047</td>
</tr>
<tr>
<td>3</td>
<td>(0.4*0.8)=0.32</td>
<td>37.0117</td>
</tr>
<tr>
<td>4</td>
<td>(0.6<em>1</em>0.25)=0.15</td>
<td>196.192</td>
</tr>
<tr>
<td>5</td>
<td>(0.4<em>0.2</em>0.25)=0.02</td>
<td>188.036</td>
</tr>
</tbody>
</table>

Table 5.23: Probabilities of each path and mean response time within Simple Network (5 Paths)

Therefore, the mean response time in this system is found by multiplying the path probabilities by the mean response time of each path giving us:

\[
\text{Mean Response Time} = (0.45 \times 193.966) + (0.06 \times 170.04696) + (0.32 \times 37.0117) + (0.15 \times 196.1919) + (0.02 \times 188.0364)
\]

\[
= 142.551 \text{ Minutes}
\]

5.13.1 Verifying Response Time using Little’s Law

In order to verify the value above, we can apply Little’s Law to this system. The response time value using Little’s Law as shown in section 5.8 was 149.536 Minutes. However, this was calculated with the service rate of tests equal to \( \frac{2}{60} \). In order for the response time distribution to be calculated, (explained in section 5.5), this value was changed to \( \mu = \frac{2.01}{60} \). Taking this into consideration, the number of jobs at tests decreased to 16.6986. This caused the total number of jobs in the system to decrease to 26.3257 from 27.6226.

Therefore, the modified response time of this system is:

\[
\frac{26.3257}{0.184722} = 142.5152 \text{ Minutes}
\]

As we can see, this value is the same as the response time value calculated above. Hence we can say that the response time distributions were calculated properly.
(Note: In our actual model, there will in fact be 9 paths due to the fact that our model is a one-class model. Therefore, a more accurate measurement would be to take all these 9 paths into consideration when calculating the mean response time of the network. The mean response time of this system should equal the value shown above)

5.14 Adding Cycles to the Extended Network

As described in section 2.1, a patient visiting the specialist doctor might have to visit the minors/majors node again depending on the outcome of the assessment by the specialist doctor. This means that any model of the A&E service unit should try and have cycles in it in order to model the A&E service unit accurately. If we look at Figure 5.2, we can see that the extended network is a feed forward network with no cycles in it. This section looks into finding the Laplace transforms of the paths with cycles. Shown in the figure below is what the network with cycles looks like:
Figure 5.3: Network of queues model for a hospital emergency department.
From the figure above we can see that the proportion of majors/minors who go back to the majors/minors doctor is $\frac{1}{4}$ of the majors and minors that enter the specialist doctor. This leaves $\frac{3}{4}$ of the majors and minors to exit the system after they have visited the specialist doctor. Furthermore, we can also see that there are 5 paths in this system:

Path 1: Receptionist -> Majors Nurse -> Tests -> Majors Doctor -> Exit (4 nodes)

Path 2: Receptionist -> Minors Nurse -> Tests -> Minors Doctor -> Exit (4 nodes)

Path 3: Receptionist -> Minors Nurse -> Minors Doctor -> Exit (3 nodes)

Path 4: Receptionist -> Majors Nurse -> Tests -> Specialist Doctor -> Exit (4 nodes)

Path 5: Receptionist -> Minors Nurse -> Tests -> Specialist Doctor -> Exit (4 nodes)

However, only paths 4 and 5 may have cycles. i.e – a patient may go through this path one or more times. As a result, if we had to calculate the Laplace transforms of the response time distributions of these 2 paths, we need to take into account the probability of each cycle occurring. Since the proportion of majors and minors that can cycle is $\frac{1}{4}$, the derivation of Laplace transforms for both these 2 paths are the same. Shown below is how the Laplace transforms of these two paths are calculated:

Taking the case of Path 4 where a majors patient may cycle 1 or many times:

$$P(i \text{ cycles}) = \frac{3}{4} \left( \frac{1}{4} \right)^{i-1} \text{ where } i = 1, 2, \ldots, n$$

(5.5)

Using Laplace transforms, the Laplace transform for one cycle is

$$f_1^*(\theta)f_2^*(\theta)\ldots f_i^*(\theta) \quad \text{where } i = \text{no. of nodes}$$

(5.6)

Therefore, the Laplace Transform of Path 4 occurring 1 or many times is:

$$= \sum_{i=1}^{\infty} p(i \text{ cycles}) \left( \frac{L^*}{L^*(\theta)} \right)^i$$

$$= L^*(\theta) \sum_{i=1}^{\infty} \left( \frac{3}{4} \right) \left( \frac{L^*}{4} \right)^{i-1}$$

$$= \left( \frac{3}{4} \right) \left( \frac{L^*}{4} \right) \frac{1}{1 - \frac{L^*}{4}}$$

$$= \frac{3L^*}{4 - L^*(\theta)}$$

(5.7)
Therefore, using the formula shown above, we can find the Laplace Transform of the response time distribution of a path in the network with cycles. As we can see, the distribution is not of general form of:

\[
L'(\theta) = \sum_{i=1}^{\infty} \frac{\alpha'_i}{A_i + \theta}
\]  

(5.8)

Therefore, if we had to find the response time distribution of this Laplace Transform, it would require slightly more calculation. i.e – you would have to expand in partial fractions like before but this time, factorise a polynomial of degree 2n for a cycle of length n. Had there been more time available, I would have attempted to solve this.
Chapter 6

Evaluation

As stated in the introduction section, this project looks into the application of different queuing models in order to model patient flow in an A&E service unit. It also investigates response times and response time distributions of such models and sees whether these distributions are acceptable in order to model an A&E service unit. This chapter evaluates the success of the different queuing models and methods that were used in this project.

6.1 Data Analysis

Chapter 3 carries out a data analysis using patient arrival data from an A&E service unit of a major London hospital. The values calculated in this chapter such as the arrival rate, response time, service rate etc are very useful as they give us a good idea of what these values actually are in an A&E service unit. They also help us parameterise of the queuing models that are studied in this project.

The analysis on response times during the week enables us to see on which days the A&E service unit has large response times. (For example, using our graph, we can clearly see that on average, a patient visiting the A&E service unit on Wednesdays will have the largest response times) Therefore, using such graphs, the management of the A&E could plan and optimise their resources in order to bring these response times down.

This chapter also investigates as to whether 98% of patients that are treated within the required response time of 240 minutes as explained in the introduction section. It was found that only 93.05% of all patients are treated within this time. Therefore, this A&E service unit doesn’t seem to be hitting the target set by the UK government and should look into ways of bringing down their response times.
CHAPTER 6. EVALUATION

6.2 M/M/1 Queue

One of the first queues I tried modelling the A&E service unit is the M/M/1 queue. This meant that in essence, the A&E service unit had only 1 staff working in it. As expected, this suggested ridiculously long service times of 73.2532 Minutes and hence only after this time, will the next patient in the queue be treated. As explained in section 4.1, such large response times would cause the queue to grow infinitely long and therefore, this queue can’t be used in order to model the A&E service unit. In fact, if this queue were to be used, the one member of staff working, would have to work at a speed of 0.19203 Patients/Minute (i.e - 11.52 Patients/hour) so that each patient will have a service time of 5.203 minutes.

6.3 M/M/m Queue

Using an M/M/m queue to model the A&E service unit gave us surprisingly good results. From the calculations done in chapter 4, it seems that we could model the A&E service unit using an M/M/14 queue. Using 14 servers with this queue, our investigation suggests that these servers will almost always be busy with an utilisation of approximately 0.96. Hence, using this queue, we can see that we are making the most of the limited resources of staff that we have in the A&E service unit. Even though we can numerically model the A&E service unit using an M/M/14 queue, there were 2 main problems of using such a queue. Firstly, it assumed that all staff are identical and hence worked at the same service rate. This is not so since there are various types of staff (eg receptionist, nurse, doctors etc) in an A&E service unit, all working at different service rates. Secondly, it assumes that all the staff are working perfectly in parallel. As this isn’t such the case in the A&E, it motivated us to model the A&E service unit using a Markovian queuing network.

6.4 Network of Queues

The network of queues overcomes the problem of modelling the A&E service unit using a single M/M/m queue. In such a model each node in the network can represent one type of staff (eg a receptionist) and hence have its own service rate. Furthermore, this network also enables us to model patient flow through an A&E service unit more accurately as we can clearly see the paths through the network that any patient may take.

6.4.1 Simple Network of Queues

The A&E service unit was firstly modelled using a network of queues having 6 nodes in it. Each node uniquely identified a type of staff in the A&E service unit. Once the nodes were parameterised, it gave us a response time of 162.11 minutes. This is very close to the actually response time of the A&E service unit calculated in section 3.3. As stated in section 5.2.3, this model can easily be tweaked in order to give us exact value of the response time calculated in section 3.3. Hence, this model seems to be a good representation of the A&E service unit. One slight problem with this model was the fact that since it’s a one-class model, all patients entering the ‘tests’ node are treated as one instead of being distinguished as majors and minors. This causes there
to be 5 paths (in our model) instead of 3. Hence, adjustments could be made to such a model in order to model majors and minors classes more accurately and have only 3 paths in the system through which a patient can go through.

This project also looked into the bottlenecks when the A&E service unit is modelled using the Markovian queuing network described above. Our investigation shows that using the arrival rate found in section 3.1, the ‘tests’ node seems to be the bottleneck. A bottleneck analysis was carried out by varying the number of servers and service rate at this node and for each change in the parameters, the maximum arrival rate that the node could handle (whilst preserving the steady state of the system) was calculated. Graphs were plotted for this. As expected, as the number of resources at the node increased, the response time of the system decreased. Hence, by looking at these graphs, the management of the A&E service unit could see the minimum number of servers/service rate that they would need at the bottleneck for a particular arrival value. Furthermore, the graphs also show us the difference in response times at a particular arrival rate when different values of M and $\mu$ are used. For example, the difference in response time of the system when M=3 or M=4 at arrival rate x. Hence, using these graphs, the management of the A&E service rate could decide whether the increase in staff at that node compensates for the fact that there is a decrease in the response time of the system. This kind of analysis is particularly useful when response times are large and small decreases in response time are vital in order to get 98% of all patients entering an A&E service treated within 4 hours.

Furthermore, this project also investigates the effect on the bottlenecks in the system when the arrival rate is increased. Such a circumstance may occur if there is a terrorist attack, epidemic flu or accident related events in the region. This sort of analysis helps the management identify which nodes have to be looked into at various arrival rates helping them plan and optimise their resources. A bottleneck analysis is carried out at each node to see the effect on the response time of the system when the parameters at the bottleneck are varied.

6.4.2 Extended Network of queues

As stated in the section 2.1, any minors or majors practitioner can refer a patient to a specialist doctor. Hence, in order to model the A&E service unit more accurately, the specialist doctor was added to the simple model (explained above) making the total nodes in the model now 7. Using Little’s Law, the response time for a patient was calculated for this extended network and was found to be 149.536 minutes. This value is closer to the actual response time value calculated in section 3.3 than the response time value of the simple network model described above. This suggests that this extended network is a better representation of the A&E service unit than the simple network. Like with the simple network, since the extended is modelled as a one class model, our model gives us additional paths as there is no clear distinction of majors and minors entering tests. Hence, a minor for example could go to the Majors Doctor. Therefore, a refinement to this model could be a 2 class model.

Furthermore, this project also looks at varying the proportion of majors and minors visiting the specialist doctor after they have visited tests. It also finds the proportion of these patients that would make the specialist doctor the bottleneck in the system. It
was found that if 55% of the majors and 25% of minors exiting tests visited the specialist doctor, the specialist doctor becomes the bottleneck in the system. Here too, a bottleneck analysis was carried out. Lastly, an analysis to investigate the effect of varying the arrival rate on the bottlenecks in the system was carried out in order to see how the bottlenecks move from node to node.

### 6.5 Response Time Distributions

One of the objectives stated in the introduction section was to find response time distributions of the different queuing models that I used. Response time distributions are particularly useful since they enable us to see which response times are most likely with each queuing model. For example, in our simple network of queues, if we saw the path of a minors patient who doesn’t visit tests, we would expect it to have high probabilities for low response times. On the other hand, a patient who does visit tests, would have high probabilities for high response times. This is because by visiting tests, the patient has to visit an extra node and hence the overall response time goes up. Especially when that node is the bottleneck

In this project, I firstly studied response time distributions of an M/M/14 queue, using the method of convolutions. Since the M/M/14 queue was a reasonable approximation to modelling the A&E service unit, this graph gave insight into what the response time distributions actually look like in the A&E service unit. Furthermore, the external arrival rate was varied and response time graphs were plotted. The external arrival rate caused the utilisation of the queue to vary and in general, our investigation found that once the arrival rate was increased (i.e. utilisation increased) the shape of the curve widened and had a longer tail. On the other hand, when we decreased the arrival rate and the system was less utilised, the curve had a narrower shape and a shorter tail. This is because, when a system is highly utilised, there is a very low probability of a patient having small response times. Hence, large response times in such a graph have relatively high probabilities. This causes the curve of the graph to widen and hence the tail lengthens. Furthermore, in general, the response time distribution graphs have a peak in them which isn’t present in an exponential distribution.

Although the method of convolutions is good in order to find distributions of independent variables, when we need to find distributions between a large number of independent variables, this process becomes more complicated as the number of integrals to integrate increases. Hence, this project also studied the use of Laplace transforms in order to find distributions of independent variables. The advantage of Laplace transforms is that we can simply use the property that the distribution of the sum of two independent variables is the convolution of their distributions, which has Laplace transform equal to the product of their Laplace transforms. Therefore, when finding response time distributions between the paths in a network, if we had a large number of nodes in the path, all we would have to do is find the product of their Laplace transforms and then apply the inverse Laplace transform operation to find the response time distribution of the path. Hence, using Laplace transforms, the response time distributions graphs of the different paths in the simple network were plotted. Using this, the mean response time for each path was calculated which in turn was used to calculate the weighted response time of the whole network. This value was compared to the response time found using Little’s Law. Our investigation showed
that both these values were exactly the same as expected. Hence, this suggested that
the response time distributions calculated were correct.

The response time distributions through paths in the extended network too were
calculated and verified using Little’s Law.

6.6 Future Work

This section discusses some of the investigations that could have been done in order
to model the A&E service unit more accurately had there been more time.

Our model is a one-class model of the A&E service unit. Since patients are classified
as majors or minors, in order to prevent the extra paths caused by the one class model,
and model the A&E service unit accurately, we should investigate the effect of
modelling it using a two-class model, in which each class has its own routing
probabilities as well as node service times.

Although I have implemented cycles in the extended system and calculated the
Laplace transforms of the paths with cycles, we should find the response time
distributions of these Laplace transforms and see the effect that of the cycles on the
total response time of the network.

Another approximation made by our model is that we have an adequate amount of
passive resources in the A&E service unit. These resources include beds, cubicles,
assessment rooms etc that a practitioner or specialist doctor might use in order to treat
a patient. But, in the real A&E service unit, there is a limit to the amount of passive
resources that it has. Therefore, even though a practitioner/specialist doctor may be
free, if there aren’t any resources, he/she won’t be able to serve the patient and hence,
the patient will have to queue for longer. i.e – wait in an extra queue before he/she can
even enter the A&E service unit as modelled here. Hence, one extension to our model
could be to take this into account when the patient is queuing.

Sometimes, patients can’t walk due to an injury that they may have. As a result,
porters are used in order to move the patient from one department to another. This
hasn’t been represented in my system and therefore a more accurate model of the
A&E service unit would have to take this into consideration.

In our model, we have assumed that a particular type of staff (eg. receptionist, nurse,
doctor) only performs a single task i.e - an assessment on the patient. In a real A&E
service unit, this is not the case. For example if we take nurses, they are trained to
perform assessments and treatments. Therefore, in some circumstances, a nurse may
even treat a patient. Hence, if we were to improve the model, we would have to take
this into consideration too.

Even though a simulation may not provide as accurate a result as the analytical
methods studied in this project (refer section 2.2), given its assumptions, a simulation
can model far more generally specified systems, albeit slowly in complex cases. If a
simulation was built, we could verify and validate some of our results that we
obtained using the analytical techniques.
Lastly, we could also incorporate a ‘breachbuster’ into our queuing model. A breachbuster is a member of staff in the A&E service unit who is brought into use when there is an instance of a patient’s response time larger than 240 minutes. Although it is hard to incorporate accurately, such an addition might be able to indicate when a breachbuster would be required and whether it improved overall performance or not, compared with adding an extra nurse, for example.
Chapter 7

Conclusion

This project investigated the application of different queuing models in order to model an A&E service unit. Furthermore, response time and response time distributions of such queues are studied too. Our investigation showed that an M/M/14 queue is a surprisingly reasonable approximation to model an A&E service unit. Due to its constraints of its servers being identical and working perfectly in parallel, we studied the effect of modelling the A&E service unit using a Markovian queuing network.

This queuing network models the A&E service more accurately as it enables us capture the patient flow as well as overcomes the problems with the M/M/m queue. Response times calculated using the simple network of queues give us a very similar value to that found using real patient data of an A&E service unit of a major London hospital. Furthermore, the addition of the specialist doctor gives an even better representation of the A&E service unit. Response times calculated using this model are closer to the actual response time than when modelled using the simple network.

The response time distribution graphs for the M/M/m queue show us that when the arrival rate of such a queue is varied, the general shape of the graph varies. Furthermore, the application of Laplace Transforms enabled us to calculate the response time distributions and mean response times within the paths of the Markovian queuing network. When the weighted response time of the paths was verified with the response time calculated using Little’s law, as expected they were the same. This tends to confirm that the response time distributions were calculated correctly and enabled us to verify the response time within the network.

There are still a couple of things that could be done in order to model the A&E service unit more accurately as stated in section 6.6. An obvious extension would be to use more general service times than exponential. This is intractable in queuing networks with First-Come-First-Sever queuing discipline, but where there is a good degree of sharing of a resource amongst waiting patients, (e.g. tests) processor sharing provides a good approximation and does not require any assumptions about service time distribution in the calculation of queue length distributions at equilibrium. However, this project gives us a guideline as to which queuing models could be used and corresponding response time distributions we expect to get when modelling an A&E service unit.
Bibliography


Appendix

The tables in this section, show the number of jobs (queue lengths) at each node when the external arrival rate is varied. This is done for the different bottlenecks described in the report.

(Note: $\lambda$ represents the external arrival rate and ‘Total L’ gives us the total number of jobs in the system. The label next to each table’s legend gives us the value of the parameter.)

Tests as Bottleneck in Simple Network

A.1 - (M = 2)

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2061</td>
<td>0.0800</td>
<td>0.1800</td>
<td>0.6462</td>
<td>64.6182</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4257</td>
<td>0.1600</td>
<td>0.3601</td>
<td>1.3061</td>
<td>65.3040</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.9790</td>
<td>0.3200</td>
<td>0.7215</td>
<td>2.7433</td>
<td>68.5837</td>
</tr>
<tr>
<td>0.06</td>
<td>0.3069</td>
<td>0.2402</td>
<td>0.5441</td>
<td>1.9570</td>
<td>0.4802</td>
<td>1.0897</td>
<td>4.6181</td>
<td>76.9680</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>4.8841</td>
<td>0.6408</td>
<td>1.4770</td>
<td>8.4717</td>
<td>105.8963</td>
</tr>
<tr>
<td>0.09</td>
<td>0.4740</td>
<td>0.3608</td>
<td>0.8299</td>
<td>11.6737</td>
<td>0.7215</td>
<td>1.6841</td>
<td>15.7439</td>
<td>174.9326</td>
</tr>
</tbody>
</table>
### A.2 - \( M = 4 \)

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0800</td>
<td>0.1800</td>
<td>0.6440</td>
<td>64.4040</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4081</td>
<td>0.1600</td>
<td>0.3601</td>
<td>1.2885</td>
<td>64.4230</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8186</td>
<td>0.3200</td>
<td>0.7215</td>
<td>2.5830</td>
<td>64.5753</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6983</td>
<td>0.6408</td>
<td>1.4770</td>
<td>5.2859</td>
<td>66.0739</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5333</td>
<td>0.4013</td>
<td>0.9300</td>
<td>2.2314</td>
<td>0.9024</td>
<td>1.9052</td>
<td>6.9036</td>
<td>69.0357</td>
</tr>
<tr>
<td>0.11</td>
<td>0.5950</td>
<td>0.4418</td>
<td>1.0337</td>
<td>2.5500</td>
<td>0.8837</td>
<td>2.1457</td>
<td>7.6499</td>
<td>69.5448</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>2.9253</td>
<td>0.9659</td>
<td>2.4131</td>
<td>8.5879</td>
<td>71.5660</td>
</tr>
<tr>
<td>0.13</td>
<td>0.7268</td>
<td>0.5235</td>
<td>1.2550</td>
<td>3.3856</td>
<td>1.0482</td>
<td>2.7181</td>
<td>9.6571</td>
<td>74.2855</td>
</tr>
<tr>
<td>0.14</td>
<td>0.7977</td>
<td>0.5647</td>
<td>1.3747</td>
<td>3.9810</td>
<td>1.1315</td>
<td>3.0762</td>
<td>10.9258</td>
<td>78.0415</td>
</tr>
<tr>
<td>0.15</td>
<td>0.8727</td>
<td>0.6062</td>
<td>1.5022</td>
<td>4.8013</td>
<td>1.2159</td>
<td>3.5115</td>
<td>12.5097</td>
<td>83.3981</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>6.0381</td>
<td>1.3014</td>
<td>4.0634</td>
<td>14.6424</td>
<td>91.5152</td>
</tr>
<tr>
<td>0.17</td>
<td>1.0374</td>
<td>0.6900</td>
<td>1.7878</td>
<td>8.1757</td>
<td>1.3884</td>
<td>4.8013</td>
<td>17.8806</td>
<td>105.1978</td>
</tr>
<tr>
<td>0.18</td>
<td>1.1229</td>
<td>0.7325</td>
<td>1.9504</td>
<td>12.9022</td>
<td>1.4770</td>
<td>5.8599</td>
<td>24.0450</td>
<td>133.5831</td>
</tr>
<tr>
<td>0.19</td>
<td>1.2268</td>
<td>0.7755</td>
<td>2.1305</td>
<td>33.0122</td>
<td>1.5676</td>
<td>7.5412</td>
<td>46.2537</td>
<td>243.4407</td>
</tr>
<tr>
<td>0.195</td>
<td>1.2790</td>
<td>0.7971</td>
<td>2.2283</td>
<td>182.5950</td>
<td>1.6137</td>
<td>8.8308</td>
<td>197.3439</td>
<td>1012.0199</td>
</tr>
</tbody>
</table>

### A.3 - \( M = 8 \)

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0800</td>
<td>0.1800</td>
<td>0.6440</td>
<td>64.4037</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4080</td>
<td>0.1600</td>
<td>0.3601</td>
<td>1.2884</td>
<td>64.4181</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8160</td>
<td>0.3200</td>
<td>0.7215</td>
<td>2.5804</td>
<td>64.5096</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6321</td>
<td>0.6408</td>
<td>1.4770</td>
<td>5.2197</td>
<td>65.2458</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>2.4498</td>
<td>0.9656</td>
<td>2.4131</td>
<td>8.1121</td>
<td>67.6011</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>3.2782</td>
<td>1.3014</td>
<td>4.0634</td>
<td>11.8825</td>
<td>74.2659</td>
</tr>
<tr>
<td>0.2</td>
<td>1.3333</td>
<td>0.8189</td>
<td>2.3321</td>
<td>4.1477</td>
<td>1.6605</td>
<td>10.6898</td>
<td>20.9824</td>
<td>104.9119</td>
</tr>
</tbody>
</table>

### A.4 - \( \mu = \frac{1}{60} \)

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.4081</td>
<td>0.0800</td>
<td>0.1800</td>
<td>0.8481</td>
<td>84.8134</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.8186</td>
<td>0.1600</td>
<td>0.3601</td>
<td>1.6990</td>
<td>84.9496</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>1.6983</td>
<td>0.3200</td>
<td>0.7215</td>
<td>3.4627</td>
<td>86.5679</td>
</tr>
<tr>
<td>0.06</td>
<td>0.3069</td>
<td>0.2402</td>
<td>0.5441</td>
<td>2.9253</td>
<td>0.4802</td>
<td>1.0897</td>
<td>5.5864</td>
<td>93.1060</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>6.0381</td>
<td>0.6408</td>
<td>1.4770</td>
<td>9.6257</td>
<td>120.3209</td>
</tr>
<tr>
<td>0.09</td>
<td>0.4740</td>
<td>0.3608</td>
<td>0.8299</td>
<td>12.9022</td>
<td>0.7215</td>
<td>1.6841</td>
<td>16.9724</td>
<td>188.5826</td>
</tr>
</tbody>
</table>
### A.5 - ($\mu = \frac{2}{60}$)

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0800</td>
<td>0.1800</td>
<td>0.6440</td>
<td>64.4040</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4081</td>
<td>0.1600</td>
<td>0.3601</td>
<td>1.2885</td>
<td>64.4230</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8186</td>
<td>0.3200</td>
<td>0.7215</td>
<td>2.5830</td>
<td>64.5753</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6983</td>
<td>0.6408</td>
<td>1.4770</td>
<td>5.2859</td>
<td>66.0739</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>2.9253</td>
<td>0.9656</td>
<td>2.4131</td>
<td>8.5876</td>
<td>71.5635</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>6.0381</td>
<td>1.3014</td>
<td>4.0634</td>
<td>14.6424</td>
<td>91.5152</td>
</tr>
<tr>
<td>0.19</td>
<td>1.2268</td>
<td>0.7755</td>
<td>2.1305</td>
<td>33.0122</td>
<td>1.5676</td>
<td>7.5412</td>
<td>46.2537</td>
<td>243.4407</td>
</tr>
</tbody>
</table>

### A.6 - ($\mu = \frac{3}{60}$)

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.1360</td>
<td>0.0800</td>
<td>0.1800</td>
<td>0.5760</td>
<td>57.6037</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.2720</td>
<td>0.1600</td>
<td>0.3601</td>
<td>1.1524</td>
<td>57.6188</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.5444</td>
<td>0.3200</td>
<td>0.7215</td>
<td>2.3088</td>
<td>57.7193</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.0981</td>
<td>0.6408</td>
<td>1.4770</td>
<td>4.6857</td>
<td>58.5707</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>1.6983</td>
<td>0.9656</td>
<td>2.4131</td>
<td>7.3607</td>
<td>61.3392</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>2.4385</td>
<td>1.3014</td>
<td>4.0634</td>
<td>11.0429</td>
<td>69.0180</td>
</tr>
<tr>
<td>0.19</td>
<td>1.2268</td>
<td>0.7755</td>
<td>2.1305</td>
<td>3.2205</td>
<td>1.5676</td>
<td>7.5412</td>
<td>16.4620</td>
<td>86.6421</td>
</tr>
</tbody>
</table>

### A.7 - ($\mu = \frac{4}{60}$)

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.1020</td>
<td>0.0800</td>
<td>0.1800</td>
<td>0.5420</td>
<td>54.2037</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.2040</td>
<td>0.1600</td>
<td>0.3601</td>
<td>1.0844</td>
<td>54.2183</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.4081</td>
<td>0.3200</td>
<td>0.7215</td>
<td>2.1725</td>
<td>54.3120</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>0.8186</td>
<td>0.6408</td>
<td>1.4770</td>
<td>4.4062</td>
<td>55.0777</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>1.2414</td>
<td>0.9656</td>
<td>2.4131</td>
<td>6.9038</td>
<td>57.5315</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>1.6983</td>
<td>1.3014</td>
<td>4.0634</td>
<td>10.3027</td>
<td>64.3917</td>
</tr>
<tr>
<td>0.19</td>
<td>1.2268</td>
<td>0.7755</td>
<td>2.1305</td>
<td>2.0875</td>
<td>1.5676</td>
<td>7.5412</td>
<td>15.3290</td>
<td>80.6790</td>
</tr>
</tbody>
</table>
## Majors Doctor as Bottleneck in Simple Network

### A.8 - (M = 4)

<table>
<thead>
<tr>
<th>λ</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>1.3333</td>
<td>0.8189</td>
<td>2.3321</td>
<td>4.1477</td>
<td>1.6605</td>
<td>10.6898</td>
<td>20.9824</td>
<td>104.9119</td>
</tr>
<tr>
<td>0.21</td>
<td>1.4495</td>
<td>0.8629</td>
<td>2.5610</td>
<td>4.3790</td>
<td>1.7560</td>
<td>18.9142</td>
<td>29.9226</td>
<td>142.4885</td>
</tr>
<tr>
<td>0.22</td>
<td>1.5771</td>
<td>0.9075</td>
<td>2.8246</td>
<td>4.6191</td>
<td>1.8546</td>
<td>100.7730</td>
<td>112.5558</td>
<td>511.6175</td>
</tr>
</tbody>
</table>

### A.9 - (M = 5)

<table>
<thead>
<tr>
<th>λ</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>1.3333</td>
<td>0.8189</td>
<td>2.3321</td>
<td>4.1477</td>
<td>1.6605</td>
<td>4.6553</td>
<td>14.9479</td>
<td>74.7394</td>
</tr>
<tr>
<td>0.21</td>
<td>1.4495</td>
<td>0.8629</td>
<td>2.5610</td>
<td>4.3790</td>
<td>1.7560</td>
<td>5.2438</td>
<td>16.2521</td>
<td>77.3912</td>
</tr>
<tr>
<td>0.23</td>
<td>1.7180</td>
<td>0.9527</td>
<td>3.1334</td>
<td>4.8706</td>
<td>1.9567</td>
<td>7.0734</td>
<td>19.7048</td>
<td>85.6731</td>
</tr>
<tr>
<td>0.25</td>
<td>2.0513</td>
<td>1.0455</td>
<td>3.9533</td>
<td>5.4207</td>
<td>2.1739</td>
<td>11.3624</td>
<td>26.0070</td>
<td>104.0280</td>
</tr>
<tr>
<td>0.27</td>
<td>2.4799</td>
<td>1.1417</td>
<td>5.2588</td>
<td>6.0666</td>
<td>2.4131</td>
<td>37.1700</td>
<td>54.5301</td>
<td>201.9634</td>
</tr>
</tbody>
</table>

### A.10 - (M = 6)

<table>
<thead>
<tr>
<th>λ</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>1.3333</td>
<td>0.8189</td>
<td>2.3321</td>
<td>4.1477</td>
<td>1.6605</td>
<td>3.8949</td>
<td>14.1874</td>
<td>70.9372</td>
</tr>
<tr>
<td>0.23</td>
<td>1.7180</td>
<td>0.9527</td>
<td>3.1334</td>
<td>4.8706</td>
<td>1.9567</td>
<td>4.8525</td>
<td>17.4840</td>
<td>76.0172</td>
</tr>
<tr>
<td>0.27</td>
<td>2.4799</td>
<td>1.1417</td>
<td>5.2588</td>
<td>6.0666</td>
<td>2.4131</td>
<td>7.1545</td>
<td>24.5147</td>
<td>90.7950</td>
</tr>
<tr>
<td>0.28</td>
<td>2.7451</td>
<td>1.1913</td>
<td>6.2656</td>
<td>6.4444</td>
<td>2.5433</td>
<td>8.2005</td>
<td>27.3902</td>
<td>97.8220</td>
</tr>
<tr>
<td>0.3</td>
<td>3.4286</td>
<td>1.2941</td>
<td>10.0535</td>
<td>7.3755</td>
<td>2.8306</td>
<td>12.0611</td>
<td>37.0434</td>
<td>123.4778</td>
</tr>
<tr>
<td>0.32</td>
<td>4.4444</td>
<td>1.4023</td>
<td>25.0889</td>
<td>8.7138</td>
<td>3.1651</td>
<td>27.1625</td>
<td>69.9771</td>
<td>218.6784</td>
</tr>
<tr>
<td>0.33</td>
<td>5.1663</td>
<td>1.4588</td>
<td>100.1060</td>
<td>9.6628</td>
<td>3.3558</td>
<td>102.2100</td>
<td>221.9597</td>
<td>672.6052</td>
</tr>
</tbody>
</table>

### A.11 - (M = 7)

<table>
<thead>
<tr>
<th>λ</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>1.3333</td>
<td>0.8189</td>
<td>1.9052</td>
<td>4.1477</td>
<td>1.6605</td>
<td>3.6913</td>
<td>13.5569</td>
<td>67.7847</td>
</tr>
<tr>
<td>0.27</td>
<td>2.4799</td>
<td>1.1417</td>
<td>2.8892</td>
<td>6.0666</td>
<td>2.4131</td>
<td>5.5224</td>
<td>20.5130</td>
<td>75.9739</td>
</tr>
<tr>
<td>0.33</td>
<td>5.1663</td>
<td>1.4588</td>
<td>4.4028</td>
<td>9.6628</td>
<td>3.3558</td>
<td>9.2677</td>
<td>33.3132</td>
<td>100.9491</td>
</tr>
<tr>
<td>0.34</td>
<td>6.1261</td>
<td>1.5170</td>
<td>4.8013</td>
<td>10.9441</td>
<td>3.5661</td>
<td>10.6769</td>
<td>37.6314</td>
<td>110.6807</td>
</tr>
<tr>
<td>0.36</td>
<td>9.4737</td>
<td>1.6393</td>
<td>5.8599</td>
<td>15.7552</td>
<td>4.0634</td>
<td>16.2961</td>
<td>53.0875</td>
<td>147.4653</td>
</tr>
<tr>
<td>0.38</td>
<td>19.4872</td>
<td>1.7706</td>
<td>7.5412</td>
<td>35.9424</td>
<td>4.7068</td>
<td>46.6883</td>
<td>116.1365</td>
<td>305.6223</td>
</tr>
</tbody>
</table>
## Specialist Doctor as Bottleneck in Extended Network

### A.12 - (M = 1)

<table>
<thead>
<tr>
<th>λ</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0760</td>
<td>0.0810</td>
<td>0.1173</td>
<td>0.6584</td>
<td>65.8356</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4081</td>
<td>0.1520</td>
<td>0.1620</td>
<td>0.2658</td>
<td>1.3482</td>
<td>67.4115</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8186</td>
<td>0.3040</td>
<td>0.3240</td>
<td>0.7241</td>
<td>2.8937</td>
<td>72.3429</td>
</tr>
<tr>
<td>0.06</td>
<td>0.3069</td>
<td>0.2402</td>
<td>0.5441</td>
<td>1.2414</td>
<td>0.4562</td>
<td>0.4862</td>
<td>1.7027</td>
<td>4.9777</td>
<td>82.9611</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6983</td>
<td>0.6087</td>
<td>0.6489</td>
<td>5.2500</td>
<td>9.6756</td>
<td>120.9451</td>
</tr>
<tr>
<td>0.09</td>
<td>0.4740</td>
<td>0.3608</td>
<td>0.8299</td>
<td>1.9515</td>
<td>0.6851</td>
<td>0.7305</td>
<td>17.1818</td>
<td>22.2137</td>
<td>246.8189</td>
</tr>
</tbody>
</table>

### A.13 - (M = 2)

<table>
<thead>
<tr>
<th>λ</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0760</td>
<td>0.0810</td>
<td>0.1053</td>
<td>0.6463</td>
<td>64.6328</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4081</td>
<td>0.1520</td>
<td>0.1620</td>
<td>0.2123</td>
<td>1.2947</td>
<td>64.7374</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8186</td>
<td>0.3040</td>
<td>0.3240</td>
<td>0.4394</td>
<td>2.6090</td>
<td>65.2239</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6983</td>
<td>0.6087</td>
<td>0.6489</td>
<td>1.0199</td>
<td>5.4455</td>
<td>68.0690</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>2.9253</td>
<td>0.9164</td>
<td>0.9780</td>
<td>2.0892</td>
<td>9.1925</td>
<td>76.6038</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>6.0381</td>
<td>1.2329</td>
<td>1.3187</td>
<td>5.7065</td>
<td>17.5357</td>
<td>109.5984</td>
</tr>
</tbody>
</table>

### A.14 - (M = 3)

<table>
<thead>
<tr>
<th>λ</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0760</td>
<td>0.0810</td>
<td>0.1050</td>
<td>0.6460</td>
<td>64.6045</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4081</td>
<td>0.1520</td>
<td>0.1620</td>
<td>0.2101</td>
<td>1.2925</td>
<td>64.6254</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8186</td>
<td>0.3040</td>
<td>0.3240</td>
<td>0.4215</td>
<td>2.5911</td>
<td>64.7779</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6983</td>
<td>0.6087</td>
<td>0.6489</td>
<td>0.8629</td>
<td>5.2885</td>
<td>66.1064</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>2.9253</td>
<td>0.9164</td>
<td>0.9780</td>
<td>1.3747</td>
<td>8.4780</td>
<td>70.6497</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>6.0381</td>
<td>1.2329</td>
<td>1.3187</td>
<td>2.0683</td>
<td>13.8975</td>
<td>86.8594</td>
</tr>
</tbody>
</table>
### A.15 - (M = 4)

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0760</td>
<td>0.0810</td>
<td>0.1050</td>
<td>0.6460</td>
<td>64.6038</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4081</td>
<td>0.1520</td>
<td>0.1620</td>
<td>0.2100</td>
<td>1.2924</td>
<td>64.6206</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8186</td>
<td>0.3040</td>
<td>0.3240</td>
<td>0.4201</td>
<td>2.5897</td>
<td>64.7423</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6983</td>
<td>0.6087</td>
<td>0.6489</td>
<td>0.8430</td>
<td>5.2686</td>
<td>65.8578</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>2.9253</td>
<td>0.9164</td>
<td>0.9780</td>
<td>1.2799</td>
<td>8.3832</td>
<td>69.8597</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>6.0381</td>
<td>1.2329</td>
<td>1.3187</td>
<td>1.7560</td>
<td>13.5852</td>
<td>84.9075</td>
</tr>
</tbody>
</table>

### A.16 - ($\mu = \frac{1}{60}$)

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0760</td>
<td>0.0810</td>
<td>0.2123</td>
<td>0.7534</td>
<td>75.3379</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4081</td>
<td>0.1520</td>
<td>0.1620</td>
<td>0.4394</td>
<td>1.5218</td>
<td>76.0892</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8186</td>
<td>0.3040</td>
<td>0.3240</td>
<td>1.0199</td>
<td>3.1895</td>
<td>79.7372</td>
</tr>
<tr>
<td>0.06</td>
<td>0.3069</td>
<td>0.2402</td>
<td>0.5441</td>
<td>1.2414</td>
<td>0.4562</td>
<td>0.4862</td>
<td>2.0892</td>
<td>5.3642</td>
<td>89.4030</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6983</td>
<td>0.6087</td>
<td>0.6489</td>
<td>5.7065</td>
<td>10.1321</td>
<td>126.6516</td>
</tr>
<tr>
<td>0.09</td>
<td>0.4740</td>
<td>0.3608</td>
<td>0.8299</td>
<td>1.9515</td>
<td>0.6851</td>
<td>0.7305</td>
<td>17.6677</td>
<td>22.6996</td>
<td>252.2178</td>
</tr>
</tbody>
</table>

### A.17 - ($\mu = \frac{2}{60}$)

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0760</td>
<td>0.0810</td>
<td>0.1053</td>
<td>0.6463</td>
<td>64.6328</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4081</td>
<td>0.1520</td>
<td>0.1620</td>
<td>0.2123</td>
<td>1.2947</td>
<td>64.7374</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8186</td>
<td>0.3040</td>
<td>0.3240</td>
<td>0.4394</td>
<td>2.6090</td>
<td>65.2239</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6983</td>
<td>0.6087</td>
<td>0.6489</td>
<td>1.0199</td>
<td>5.4455</td>
<td>68.0690</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>2.9253</td>
<td>0.9164</td>
<td>0.9780</td>
<td>2.0892</td>
<td>9.1925</td>
<td>76.6038</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>6.0381</td>
<td>1.2329</td>
<td>1.3187</td>
<td>5.7065</td>
<td>17.5357</td>
<td>109.5984</td>
</tr>
</tbody>
</table>
### APPENDIX

**A.18 - \( \mu = \frac{3}{60} \)**

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0760</td>
<td>0.0810</td>
<td>0.0701</td>
<td>0.6111</td>
<td>61.1124</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4081</td>
<td>0.1520</td>
<td>0.1620</td>
<td>0.1407</td>
<td>1.2231</td>
<td>61.1548</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8186</td>
<td>0.3040</td>
<td>0.3240</td>
<td>0.2856</td>
<td>2.4552</td>
<td>61.3794</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6983</td>
<td>0.6087</td>
<td>0.6489</td>
<td>0.6076</td>
<td>5.0332</td>
<td>62.9156</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>2.9253</td>
<td>0.9164</td>
<td>0.9780</td>
<td>1.0199</td>
<td>8.1232</td>
<td>67.6930</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>6.0381</td>
<td>1.2329</td>
<td>1.3187</td>
<td>1.6317</td>
<td>13.4609</td>
<td>84.1308</td>
</tr>
</tbody>
</table>

**A.19 - \( \mu = \frac{4}{60} \)**

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0500</td>
<td>0.0400</td>
<td>0.0900</td>
<td>0.2040</td>
<td>0.0760</td>
<td>0.0810</td>
<td>0.0525</td>
<td>0.5936</td>
<td>59.3574</td>
</tr>
<tr>
<td>0.02</td>
<td>0.1003</td>
<td>0.0800</td>
<td>0.1801</td>
<td>0.4081</td>
<td>0.1520</td>
<td>0.1620</td>
<td>0.1053</td>
<td>1.1877</td>
<td>59.3849</td>
</tr>
<tr>
<td>0.04</td>
<td>0.2020</td>
<td>0.1600</td>
<td>0.3608</td>
<td>0.8186</td>
<td>0.3040</td>
<td>0.3240</td>
<td>0.2123</td>
<td>2.3819</td>
<td>59.5480</td>
</tr>
<tr>
<td>0.08</td>
<td>0.4167</td>
<td>0.3205</td>
<td>0.7325</td>
<td>1.6983</td>
<td>0.6087</td>
<td>0.6489</td>
<td>0.4394</td>
<td>4.8650</td>
<td>60.8123</td>
</tr>
<tr>
<td>0.12</td>
<td>0.6593</td>
<td>0.4826</td>
<td>1.1417</td>
<td>2.9253</td>
<td>0.9164</td>
<td>0.9780</td>
<td>0.6994</td>
<td>7.8026</td>
<td>65.0221</td>
</tr>
<tr>
<td>0.16</td>
<td>0.9524</td>
<td>0.6479</td>
<td>1.6393</td>
<td>6.0381</td>
<td>1.2329</td>
<td>1.3187</td>
<td>1.0199</td>
<td>12.8491</td>
<td>80.3071</td>
</tr>
</tbody>
</table>

### Nurse Major as Bottleneck in Extended Network

**A.20 - (M = 3)**

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>3.4286</td>
<td>1.2941</td>
<td>10.0535</td>
<td>7.3755</td>
<td>2.6114</td>
<td>6.4952</td>
<td>3.6888</td>
<td>34.9470</td>
<td>116.4901</td>
</tr>
<tr>
<td>0.31</td>
<td>3.8811</td>
<td>1.3475</td>
<td>14.3572</td>
<td>7.9745</td>
<td>2.7475</td>
<td>7.4078</td>
<td>4.2148</td>
<td>41.9303</td>
<td>135.2591</td>
</tr>
<tr>
<td>0.32</td>
<td>4.4444</td>
<td>1.4023</td>
<td>25.0889</td>
<td>8.7138</td>
<td>2.8932</td>
<td>8.6652</td>
<td>4.8841</td>
<td>56.0919</td>
<td>175.2873</td>
</tr>
<tr>
<td>0.33</td>
<td>5.1663</td>
<td>1.4588</td>
<td>100.1060</td>
<td>9.6628</td>
<td>3.0502</td>
<td>10.5244</td>
<td>5.7661</td>
<td>135.7347</td>
<td>411.3172</td>
</tr>
</tbody>
</table>
### A.21 - (M = 4)

<table>
<thead>
<tr>
<th>λ</th>
<th>Receptionist</th>
<th>Minor Nurse</th>
<th>Major Nurse</th>
<th>Tests</th>
<th>Minor Doctor</th>
<th>Major Doctor</th>
<th>Specialist Doctor</th>
<th>Total L</th>
<th>Response Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>3.4286</td>
<td>1.2941</td>
<td>3.5115</td>
<td>7.3755</td>
<td>2.6114</td>
<td>4.6671</td>
<td>3.6888</td>
<td>26.5769</td>
<td>88.5897</td>
</tr>
<tr>
<td>0.32</td>
<td>4.4444</td>
<td>1.4023</td>
<td>4.0634</td>
<td>8.7138</td>
<td>2.8932</td>
<td>5.2688</td>
<td>4.8841</td>
<td>31.6701</td>
<td>98.9689</td>
</tr>
<tr>
<td>0.36</td>
<td>9.4737</td>
<td>1.6393</td>
<td>5.8599</td>
<td>15.7552</td>
<td>3.6109</td>
<td>7.1545</td>
<td>11.6737</td>
<td>55.1671</td>
<td>153.2421</td>
</tr>
</tbody>
</table>